CLIP in medical imaging: A comprehensive survey

Zihao Zhao[a,1], Yuxiao Liu[a,b,1], Han Wu[a,1], Yonghao Li[a], Sheng Wang[a,c], Lin Teng[a], Disheng Liu[a], Zhiming Cui[a,*], Qian Wang[a,*], Dinggang Shen[a,d,e,*]

[a]School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
[b]Lingang Laboratory, Shanghai, China
[c]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[d]Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[e]Shanghai Clinical Research and Trial Center, Shanghai, China

ARTICLE INFO

Article history:

Keywords:
Contrastive language-image pre-training
Medical image analysis
Image-text alignment
Vision language model

ABSTRACT

Contrastive Language-Image Pre-training (CLIP), a simple yet effective pre-training paradigm, successfully introduces text supervision to vision models. It has shown promising results across various tasks, attributable to its generalizability and interpretability. The use of CLIP has recently gained increasing interest in the medical imaging domain, serving as a pre-training paradigm for image-text alignment, or a critical component in diverse clinical tasks. With the aim of facilitating a deeper understanding of this promising direction, this survey offers an in-depth exploration of the CLIP within the domain of medical imaging, regarding both refined CLIP pre-training and CLIP-driven applications. In this study, we (1) start with a brief introduction to the fundamentals of CLIP methodology. (2) Then, we investigate the adaptation of CLIP pre-training in the medical imaging domain, focusing on how to optimize CLIP given characteristics of medical images and reports. (3) Furthermore, we explore practical utilization of CLIP pre-trained models in various tasks, including classification, dense prediction, and cross-modal tasks. (4) Finally, we discuss existing limitations of CLIP in the context of medical imaging, and propose forward-looking directions to address the demands of medical imaging domain. We expect that this comprehensive survey will provide researchers in the field of medical image analysis with a holistic understanding of the CLIP paradigm and its potential implications. The project page of this survey can be found on Github.

1. Introduction

Despite the substantial progress in vision intelligence over the last decade (He et al., 2016; Tarvainen and Valpola, 2017; Dosovitskiy et al., 2020; Liu et al., 2021, 2022b), vision models were often trained on vision-modality annotations and tasks only, which would significantly limit the scope of encoded knowledge. In contrast, the form of text supervision is naturally rich in semantics, and the corresponding language models, especially today's large language models (Touvron et al., 2023; Xiong et al., 2023; Zhang et al., 2023a), typically contain a huge amount of knowledge. Hence, it is intuitive to integrate text supervision into vision tasks.

Drawing inspiration from contrastive pre-training, Radford et al. (2021) proposed Contrastive Language Image Pre-training (CLIP). Unlike conventional contrastive pre-training, which focuses solely on vision information, CLIP leverages both vision and language. Specifically, it considers text caption as a linguistic view of the image under consideration. It then pulls the pair of image and text representations close in the latent space. In this manner, the image-text pair is aligned through CLIP's vision and text encoders. CLIP has learned extensive knowledge from text supervision and proven useful in a wide variety of downstream areas, including image generation (Vinker et al., 2022; Ramesh et al., 2022; Yu et al., 2022; Rombach et al., 2022), segmentation (Li et al., 2022a; Rao et al., 2022; Luo et al., 2023), detection (Bangalath et al., 2022; Lin and Gong, 2023), and classification (Zhou et al., 2022c,d; Wang et al., 2023a).

Recently, CLIP has also gained increasing attention in the field of medical imaging. The interpretation of medical image data typically demands specialized clinical knowledge, which

*Corresponding author: Zhiming Cui, Qian Wang, Dinggang Shen
e-mail: cuizhm@shanghaitech.edu.cn (Zhiming Cui), qianwang@shanghaitech.edu.cn (Qian Wang), dgshen@shanghaitech.edu.cn (Dinggang Shen)
[1]These authors contributed equally to this paper

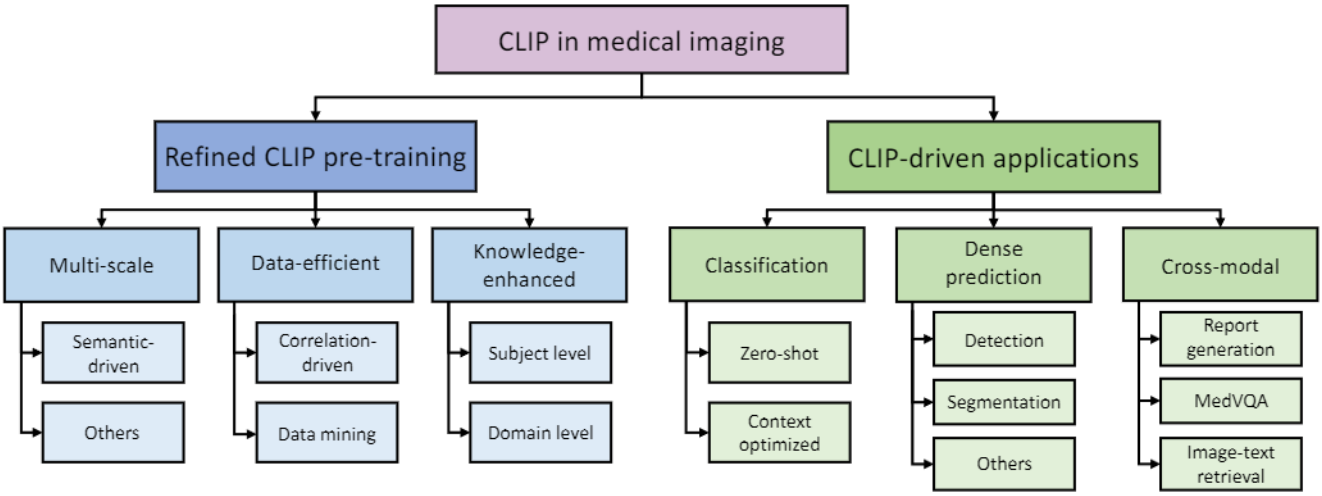


Fig. 1. Taxonomy of studies focusing on CLIP in the field of medical imaging.

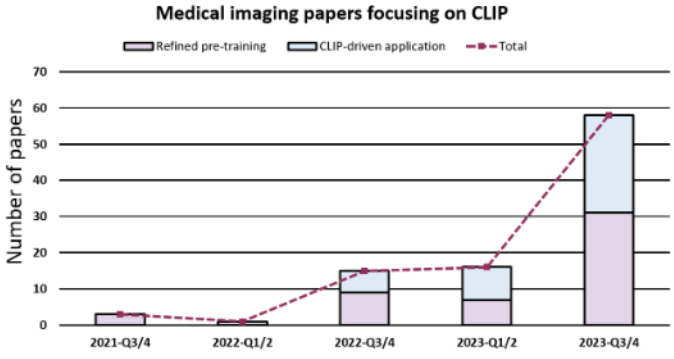


Fig. 2. The number of medical imaging papers focusing on CLIP has increased rapidly in recent years.

is not an easy goal for vision-only models. Previous studies have attempted to address the issue via fine-grained annotations such as bounding boxes (Luo et al., 2022; Ouyang et al., 2020; Tanida et al., 2023; Müller et al., 2023) and segmentation masks (Mehta et al., 2018; Zhou et al., 2019). However, collecting fine-grained annotations is non-trivial and thus hard to scale up. On the contrary, encoding clinical knowledge into deep learning models via CLIP seems to be a viable means.

In the field of medical imaging, the studies related to CLIP can be categorized into two lines, i.e., (1) refined CLIP pre-training and (2) CLIP-driven applications, as illustrated in Fig. 1. Studies focusing on pre-training methodology seek to adapt the pre-training paradigm of CLIP from web-crawled image-caption pairs to medical images and their corresponding reports. Meanwhile, researchers adopt pre-trained CLIP models as key components to implement various clinical tasks, e.g., thoracic disease diagnosis, multi-organ segmentation (Tiu et al., 2022; Pellegrini et al., 2023; Liu et al., 2023h).

Motivation and contribution. The medical imaging community has witnessed an exponential growth of CLIP-centered studies (See Fig. 2), as the literature is experiencing a large influx of contributions. Thus, a survey of the existing literature is beneficial for the community.

- This paper is a comprehensive review of CLIP in medical imaging, aiming to provide timely summary and insight for potential studies in this rapidly evolving area.
- We provide thorough coverage of existing studies and a multi-level taxonomy to serve different needs of readers.
- Furthermore, we discuss the issues and open questions in this field. We pinpoint new trends and propose future directions for further exploration.

Paper organization. The rest of the paper is organized as follows. Section 2 provides preliminary knowledge of CLIP and its variants. In Section 3, we present a systematic analysis of how to adapt CLIP to the medical imaging field, from the perspectives of key challenges and corresponding solutions. Section 4 covers various clinical applications of pre-trained CLIP and compares CLIP-driven methods with early solutions. Section 5 further discusses existing limitations, as well as potential research directions. We finally conclude this paper in Section 6.

2. Background

CLIP-related research has advanced rapidly in recent years. In this section, we provide a brief overview of CLIP, as well as its generalizability and multiple variants. Additionally, we summarize datasets of medical image-text pairs that are publicly available and usable to CLIP.

2.1. Contrastive Language-Image Pre-training

CLIP (**C**ontrastive **L**anguage-**I**mage **P**re-training) is a pre-training method developed by OpenAI, designed to bridge the gap between images and texts. It jointly optimizes a vision encoder and a text encoder, ensuring that image-text pairs are closely aligned in a shared latent space. Unlike other methods that rely on extensive manual supervision or complex structures, CLIP follows the principle of Occam's Razor (Blumer et al., 1987), favoring simplicity for effective results.

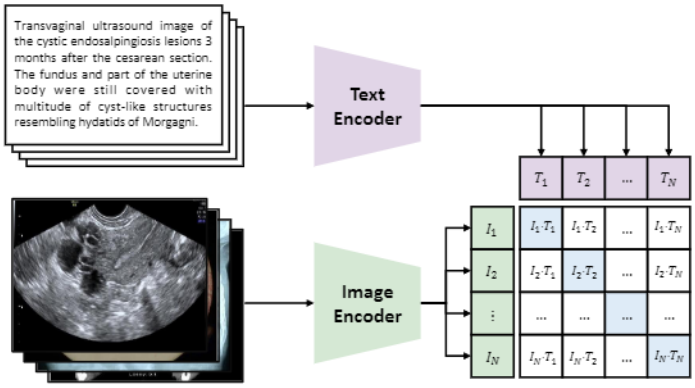


Fig. 3. Illustration of CLIP in medical imaging, with an example from the PMC-OA dataset.

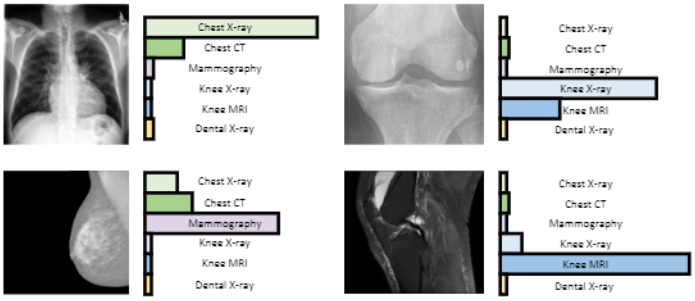


Fig. 4. Illustration of CLIP's generalizability via domain identification.

Architecture. In terms of its architecture, CLIP seamlessly integrates a vision model with a language model. The visual component can be based on either ResNet (He et al., 2016) or Vision Transformer (ViT) (Dosovitskiy et al., 2020), while the language encoder is rooted in a transformer-based model like BERT (Kenton and Toutanova, 2019). As illustrated in Fig. 3, it receives a batch of images and their corresponding text descriptions as input in each iteration. Following the encoding process, the embeddings are normalized and mapped to a joint image-text latent space. That is, the input images and texts are encoded into $I \in \mathbb{R}^{N \times D}$ and $T \in \mathbb{R}^{N \times D}$, respectively, where N denotes batch size and D represents embedding dimensionality.

Contrastive pre-training. In CLIP, contrastive pre-training plays a crucial role in aligning image-text pairs. Diverging from conventional models that are sculpted for a singular and predefined task, CLIP's learning trajectory revolves around contrastive pre-training between paired image-text information. In particular, N^2 image-text pairs can be constructed given a batch size of N, among which there are N matched image-text pairs (positive pairs, as highlighted in blue in Fig. 3) and $(N^2 - N)$ unmatched image-text pairs (negative pairs). The pre-training objective for the image encoder is hence denoted as

$$\mathcal{L}_{\text{img}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\Phi(I_i, T_i)/\tau)}{\sum_{j=1}^{N} \exp(\Phi(I_i, T_j)/\tau)}, \tag{1}$$

where $\Phi(\cdot,\cdot)$ indicates cosine similarity, τ is a learnable temperature parameter, I_i and T_i represent the i_{th} image embedding and text embedding, respectively. The objective for the text encoder is defined symmetrically:

$$\mathcal{L}_{\text{txt}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\Phi(T_i, I_i)/\tau)}{\sum_{j=1}^{N} \exp(\Phi(T_i, I_j)/\tau)}. \tag{2}$$

The total optimization objective of CLIP is hence calculated via the average of (1) and (2):

$$\mathcal{L}_{\text{total}} = \frac{\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}}}{2}. \tag{3}$$

Zero-shot capability and generalizability. Since CLIP is pre-trained to predict whether an image matches a textual description, it naturally lends itself to zero-shot recognition. This process is accomplished by comparing image embeddings with text embeddings, which correspond to textual descriptions specifying certain classes of interest. Let I_1 represent the image features extracted by the image encoder for a given image x, and let $\{W_i\}_{i=1}^{K}$ be the set of class embeddings generated by the text encoder. Here, K denotes the number of classes, and each W_i is derived from a text prompt resembling "a photo of a [CLASS]", where the class token is substituted with the specific class name. The probability of prediction is then calculated as follows:

$$p(y = i|I_1) = \frac{\exp(\Phi(I_1, W_i)/\tau)}{\sum_{j=1}^{K} \exp(\Phi(I_1, W_j)/\tau)}, \tag{4}$$

where τ is a temperature parameter learned during pre-training, and $\Phi(\cdot,\cdot)$ represents the cosine similarity. In contrast with traditional classifier learning methods where closed-set visual concepts are learned from scratch, CLIP pre-training allows for the exploration of open-set visual concepts through the text encoder. This leads to a broader semantic space and, consequently, makes the learned representations more transferable to downstream tasks.

The generalizability of the CLIP pre-trained model becomes evident when applied to specialized areas such as medical imaging. Although originally trained on internet images and their textual captions, CLIP has demonstrated the capability to recognize and categorize medical images. Fig. 4 illustrates the generalizability of CLIP via domain identification, where the class token in the text prompt is substituted with the specific class name, such as "Chest X-ray", "Mammography", "Knee X-ray", or "Dental X-ray". Its zero-shot inference capability allows it to identify the domain of a given medical image without explicit prior training on such datasets. While further studies and validations are necessary, the preliminary findings suggest that the zero-shot capability of CLIP could reduce the dependency on extensive labeled medical datasets and pave the way for more efficient and generalizable AI-driven diagnostic tools.

2.2. Variants of CLIP

After providing a concise overview of CLIP, we hereby introduce several variants of CLIP with practical applications in the area of medical imaging, which takes a step further by not only recognizing items in images but also understanding their specific details and descriptions.

Following the philosophy of CLIP, GLIP (Li et al., 2022b) reformulates detection as a grounding task by aligning each

Table 1. Summary of publicly available medical image-text datasets.

Dataset	Domain	Image	Text	Source	Language	Pre-trained CLIP
ROCO (Pelka et al., 2018)	Multiple	87K	87K	Research papers	English	PubMedCLIP
MedICaT (Subramanian et al., 2020)	Multiple	217K	217K	Research papers	English	/
PMC-OA (Lin et al., 2023b)	Multiple	1.6M	1.6M	Research papers	English	PMC-CLIP
ChiMed-VL (Liu et al., 2023g)	Multiple	580K	580K	Research papers	English & Chinese	/
FFA-IR (Li et al., 2021)	Fundus	1M	10K	Medical reports	English & Chinese	/
PadChest (Bustos et al., 2020)	Chest X-ray	160K	109K	Medical reports	Spanish	/
MIMIC-CXR (Johnson et al., 2019)	Chest X-ray	377K	227K	Medical reports	English	BioViL& BioViL-T
OpenPath (Huang et al., 2023a)	Histology	208K	208K	Social media	English	PLIP
Quilt-1M (Ikezogwo et al., 2023)	Histology	1M	1M	Research papers & Social media	English	QuiltNet

region or bounding box with corresponding text phrases. It simultaneously trains both an image encoder and a language encoder to accurately predict the associations between regions and words. A fusion module is further proposed to enhance the alignment between image and text information, improving the model's ability to learn a language-aware visual representation. Pre-trained specifically at the object level, GLIP has demonstrated remarkable performance, even comparable to fully supervised methods in zero-shot object detection and phrase grounding tasks.

Meanwhile, CLIPSeg (Lüddecke and Ecker, 2022) and CRIS (Wang et al., 2022b) extend CLIP to the area of segmentation. CLIPSeg has fixed the pre-trained CLIP image encoder and text encoder while introducing a trainable decoder for the segmentation task. The encoded image and text prompt are fused, and then input into the trainable decoder to generate the predicted segmentation mask. Conceptually aligned, a similar paradigm is proposed in CRIS. These representative variants have been favored by generalizability, further showcasing the adaptability of CLIP.

The potential of these variants lies in their attention to content details. They could combine visual and textual information to provide a more nuanced understanding of medical images. For medical imaging, it is critical to identify subtle features, e.g., related to tumors or bone fractures. Such techniques could offer significant benefits, potentially being able to spatially locate clinical findings given the provided prompt, like "malignant mass" or "calcification."

2.3. Medical image-text dataset

Large-scale datasets are essential for the alignment of image and text data in medical imaging research. Therefore, we summarize publicly available medical datasets in Table 1. These datasets encompass various medical domains and data sources. For each of them, we also signify whether there is a publicly available CLIP model pre-trained on this dataset. ROCO (Pelka et al., 2018), MedICaT (Subramanian et al., 2020), PMC-OA (Lin et al., 2023b), and ChiMed-VL (Liu et al., 2023g) are four large-scale datasets sourced from research papers. They have collected and filtered biomedical figure-caption pairs from open-access research papers via PubMed Central[2]. Since research papers could cover a wide range of topics, the resulting datasets are composed of diverse medical images. FFA-IR (Li et al., 2021), PadChest (Bustos et al., 2020), and MIMIC-CXR (Johnson et al., 2019) are collected from daily medical reports while they focus on different organs. In clinical practice, a diagnostic report is often derived from multiple images. Therefore, The number of image samples and text samples shows a significant disparity, especially in the case of the FFA-IR dataset. OpenPath (Huang et al., 2023a) has crawled histology images and paired captions from Twitter, a social media platform. Their proposed histology foundation model, PLIP, has shown impressive performance, illuminating the significance of social media-derived data. Following OpenPath, Quilt-1M (Ikezogwo et al., 2023) has extracted image-text pairs from both research papers and social media platforms like YouTube. As of December 2023, Quilt-1M stands as the largest image-text dataset for histology images. Note that not all of these datasets are in English, such as PadChest is provided in Spanish, while FFA-IR and ChiMed-VL have their respective Chinese versions.

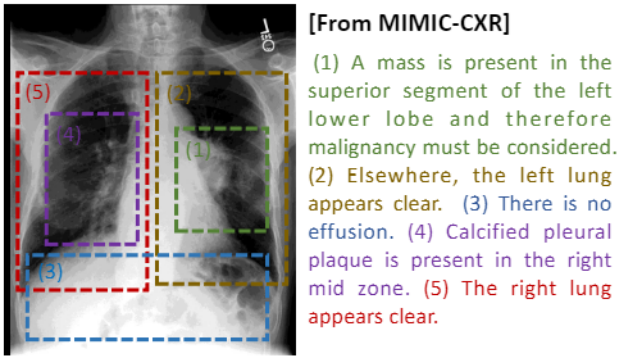


Fig. 5. Demonstration of multi-scale features of medical image-text pairs. The medical report is composed of several sentences, with each sentence focusing on region-level features instead of global-level features. Sentences are generally independent of each other, and hold different levels of significance.

3. CLIP in medical image-text pre-training

Existing CLIP models are mostly trained to encode general knowledge (Radford et al., 2021; Cherti et al., 2023; Sun et al., 2023), without emphasis on medical image and medical knowledge. Hence, several efforts have been made to overcome the challenges in the medical imaging field, e.g., to adapt the paradigm of CLIP to a specific scene (e.g., Chest X-ray, Brain

[2]https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

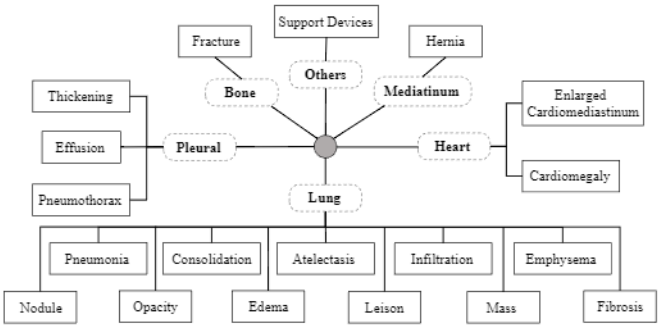


Fig. 6. An illustration of hierarchical dependencies among clinical findings in chest X-rays (sourced from Huang et al. (2023b)). Solid boxes indicate clinical findings while dotted boxes represent organs or tissues.

MRI, etc.). The goal is to yield an expert foundation model with robust expertise in its area (Zhang and Metaxas, 2023).

In this section, we describe specific challenges of medical image-text pre-training, and provide a taxonomy of existing studies given their solutions. Representative methods included in this section are shown in Table 2. Their pre-trained imaging domains, specific taxonomy, evaluation tasks, and noteworthy issues are demonstrated. Evaluation tasks here mean that the quality of pre-trained vision encoder and text encoder are evaluated by directly observing their performance on specialized tasks without much modification, which is different from CLIP-driven applications mentioned in Section 4.

3.1. Challenges of CLIP pre-training

CLIP was initially proposed on natural image datasets, which may lead to suboptimal performance on medical imaging due to three key challenges.

- **Multi-scale features.** One main difference between medical images and natural images is the significance of multi-scale features. Besides global-level vision features, local-level vision features also matter in the interpretation of medical images. For example, lesion regions often only occupy small proportions in medical images, yet their crucial visual cues can significantly influence diagnostic results. Besides image information, the corresponding text information is also characterized by multi-scale text features. Medical reports tend to be more complex than captions for natural images. Natural image captions are typically concise and provide an overview of the global features of the image. In contrast, as depicted in Fig. 5, medical reports consist of multiple sentences, with each sentence describing image findings in a specific region. For instance, the first sentence (highlighted in green) in Fig. 5 describes the presence of a mass, which is essential for accurate diagnosis. Generally speaking, besides global-level image-text contrast, both local-level image features and local-level text features should be taken into consideration during pre-training, posing challenges to the baseline CLIP pre-training, where image and text information are aligned solely at the global level.

- **Data scarcity.** Unlike natural image-text datasets, which can easily reach billion-scale (Schuhmann et al., 2022; Zhai et al., 2022; Zhu et al., 2023), medical datasets with paired images and reports (Johnson et al., 2019; Ikezogwo et al., 2023) hold a relatively limited scale. As the scale of datasets can have a significant impact on CLIP-style pre-training according to scaling laws (Cherti et al., 2023), limited medical data can hinder its performance in medical imaging domain.

- **High demands for specialized knowledge.** The hierarchical dependencies among various clinical concepts can be intricate and highly specialized. As depicted in Fig. 6, the graph is constructed based on the expert viewpoint of chest X-rays, considering correlations, characteristics, and occurrence locations of clinical findings (Huang et al., 2023b). Lack of in-depth understanding of medical concepts may lead to degraded performance when facing data from shifted distributions, or even shortcut solutions (Geirhos et al., 2020). Hence, in order to improve the reliability and robustness, explicitly incorporating knowledge during the process of pre-training may provide a viable solution.

These challenges highlight the impracticality of directly applying CLIP pre-training on medical image-text datasets, motivating related work to improve CLIP-style pre-training in the medical imaging domain.

3.2. Multi-scale contrast

Although some early-stage studies (Zhang et al., 2022b; Zhou et al., 2022a) attempt to extend CLIP pre-training to the medical imaging domain, they still follow the global-level contrast proposed in Radford et al. (2021) and hence show suboptimal performance on tasks such as semantic segmentation and object detection. To address the issue, several studies have tried to perform multi-scale contrast in pre-training.

Huang et al. (2021) made a pioneering contribution in this domain by introducing the concept of semantic-driven multi-scale contrast. The proposed GLoRIA follows a similar paradigm to CLIP to implement global-level contrast, yet it distinguishes itself by implementing local-level contrast between each word representation and its semantically similar visual counterparts. As illustrated in Fig. 7(a), GLoRIA defines each word and each image sub-region as local text and image features, respectively. To perform local-level contrast, it calculates the semantic similarity between word-wise text features and sub-region-wise image features. After obtaining the semantic similarity, GLoRIA leverages it to compute weighted summation over all local image features and hence gets an attention-weighted local image representation for each word representation, where "attention" mentioned in their paper denotes semantic similarity. The word representation and the weighted image representation are semantically similar and hence are pulled closer in the latent space via localized contrastive loss. Since local-level contrast is implemented at the word level, GLoRIA would accumulate all localized contrastive loss as the total local-level contrast objective for the medical report. Fig. 7(b)

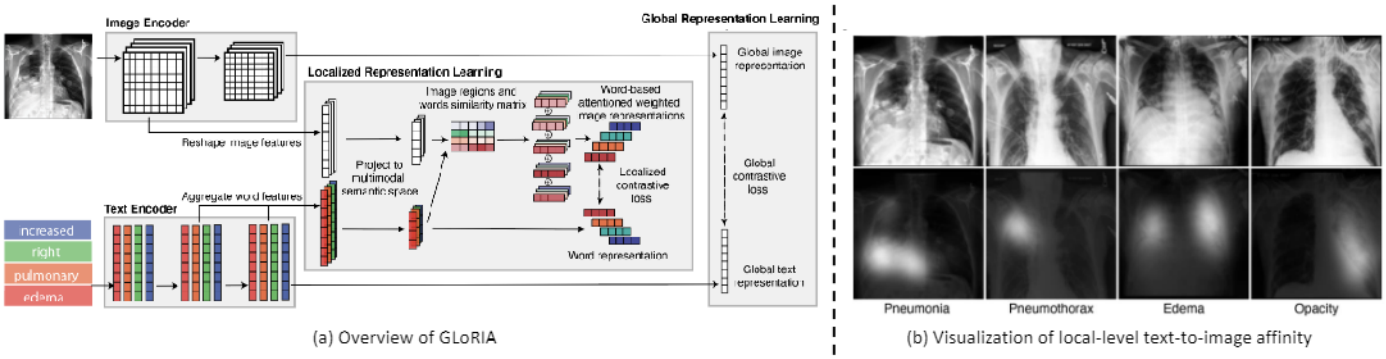


Fig. 7. Illustration of semantic-driven contrast proposed by GLoRIA (Huang et al., 2021). (a) Overview of GLoRIA, which performs multi-scale image-text alignment based on cross-modal semantic affinity. (b) Visualization of semantic affinity learned by GLoRIA.

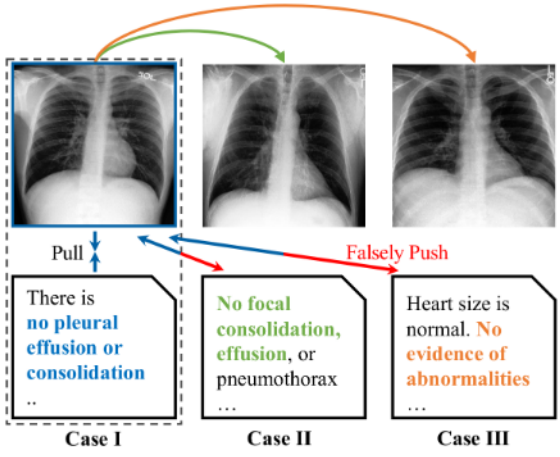


Fig. 8. Illustration of false-negative pairs (Liu et al., 2023a). In CLIP pre-training, a positive pair is defined as an image and its corresponding report, whereas all other reports are considered negatives. This can result in false negatives, where semantically similar reports from different subjects are mistakenly considered as negative pairs.

demonstrates the effectiveness of semantic affinity learned by GLoRIA. The text-to-image semantic affinity, visualized as a heatmap, is able to correctly identify related image sub-regions for a given word. For instance, the semantic affinity based on the word "Pneumonia" correctly localizes regions of the right lower lobe containing heterogenous consolidative opacities indicative of pneumonia. Additionally, attention weights associated with "Pneumothorax" accurately emphasize lucency in the right lung apex, indicative of pneumothorax. Analogous results can be observed for "Edema" and "Opacity", highlighting the efficiency of local-level alignment.

The semantic-driven multi-scale contrast proposed by Huang et al. (2021) is intuitive, but it still has some apparent weaknesses. (1) The local-level contrast is designed asymmetrically. Its contrastive objective is optimized between text and attention-weighted sub-region features. This only guarantees the alignment from text to image, dismissing the alignment from image to text. (2) While computing text-to-image local features through weighted summation of local image features is intuitive, it may struggle to capture implicit semantic correlations between image and text features. (3) The total local-level objective simply accumulates all localized contrastive loss, implying each local text feature to be treated equally. However, different sentences within medical reports hold varying levels of importance for diagnosis as illustrated in Fig. 5.

Motivated by the above-mentioned limitations, Müller et al. (2022b) proposed an improved semantic-driven contrastive method, LoVT, covering both text-to-image local alignment and image-to-text local alignment. It still takes image sub-region features as local image features but divides medical reports into sentences instead of words. To better capture the implicit semantic features, both text-to-image local features and image-to-text local features are learned by transformer layers instead of weighted summation. Moreover, the weight for each word's localized contrastive loss is assigned adaptively via transformer's attention map (Dosovitskiy et al., 2020). Cheng et al. (2023) extended the LoVT by incorporating a conditional reconstruction task for image and text representations. This extension facilitates cross-modality feature interaction and learns more fine-grained scale alignment. Additionally, they propose a prototype memory bank for sentence-level embeddings, expecting to learn high-level text features in the joint image-text space. In a parallel study (Zhang et al., 2023b), a similar methodology is employed. However, it focuses on reconstructing raw text reports instead of using a prototype memory bank.

Besides this line of research, there are many other studies focusing on multi-scale contrast. Liao et al. (2021) optimized the estimated mutual information between local image features and sentence-level text representations to implement local feature alignment. Seibold et al. (2022) assumed that each sentence could convey distinct information for diagnosis, and proposed to perform image-sentence alignment. Further, Palepu and Beam (2023) sought to penalize the entropy of the text-token to image-patch similarity scores.

3.3. Data-efficient contrast

Large-scale medical imaging datasets with paired reports are hard to obtain due to ethical concerns, which adversely influences the effectiveness of CLIP pre-training due to its data-thirsty nature. To tackle this challenge, various studies have endeavored to implement contrastive image-text pre-training in a more efficient manner, broadly falling into two categories.

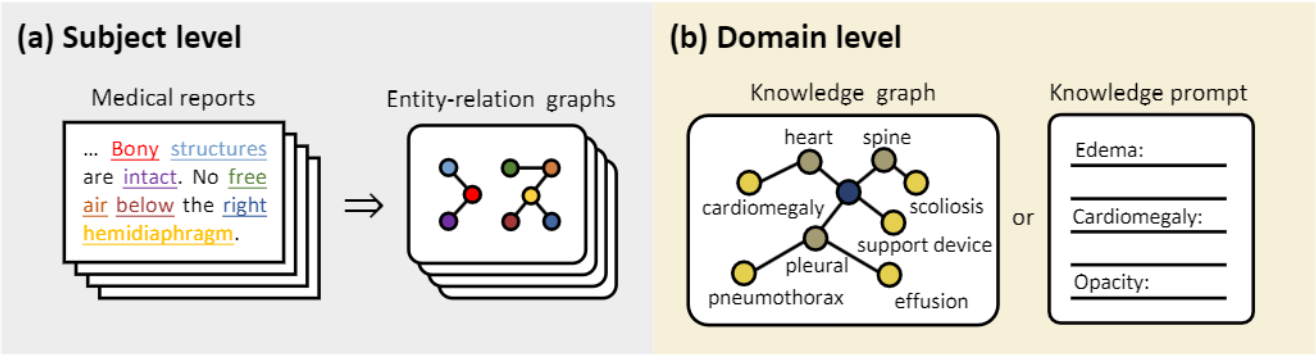


Fig. 9. Illustration of knowledge enhancement at different levels (with the chest X-ray as an example). (a) At the subject level, external knowledge aids in converting medical reports into entity-relation graphs, which elucidate the causal relationships among medical entities in the report. (b) At the domain level, the domain knowledge is presented as a knowledge graph or descriptive knowledge prompt, directly providing human prior guidance for pre-training.

Correlation-driven contrast. Several studies managed to boost the efficiency of contrastive pre-training based on semantic correlation. One notable distinction between medical reports and image captions lies in the fact that medical reports were written with a clear diagnostic purpose. Since a small proportion of diseases/findings typically cover most cases (Bustos et al., 2020), the semantic overlap between medical reports can be significant, especially for normal cases as shown in Fig. 8. As a result, simply treating unpaired images and reports as negative pairs can lead to issues of false-negative and degrade the efficacy of pre-training. Motivated by this observation, Wang et al. (2022d) followed the practice of NegBio (Peng et al., 2018) to construct a correlation matrix for all medical reports within the training set. The pre-computed correlation is then employed as a soft optimization target, instead of the original one-hot optimization target, for image-text alignment, enabling the effective utilization of unpaired false-negative reports. Under conditions where medical reports are not available, Wang (2023) constructed multi-hot optimization targets for contrastive pre-training by measuring the correlation between different samples' ground-truth labels. SDA-CLIP (Li et al., 2023) extends this method to surgical video data. It selects the Kullback-Leibler (KL) divergence to evaluate the correlation, and the predicted correlation distribution should approach the ground-truth annotation distribution. In a recent study, Liu et al. (2023a) further categorized all image-text pairs into positive, negative, and neutral pairs based on inter-report correlation. This improved categorization of sample pairs allows for a more precise mining of false-negative pairs. MGCA (Wang et al., 2022a) focuses on the disease-level inter-sample correlation, a level of abstraction higher than that of image-level semantics. It has designed a novel cross-modal disease-level alignment framework to serve samples with the same diseases. The semantic correlation can also be extended to the imaging modality level. For multi-modal brain MRI and corresponding modality-wise reports, UniBrain (Lei et al., 2023b) aligns the modality-wise image-text features and then concatenates these features together to realize subject-wise image-text alignment.

Data mining. Simultaneously, many other studies tried to boost training efficiency by mining supplementary information. Within a diagnostic report, the *Findings* section provides a detailed description of clinical observations, while the *Impression* typically encapsulates these findings and offers an overall assessment (Wallis and McCoubrie, 2011). While previous studies primarily focused on extracting the *Findings* section from raw diagnostic reports, often overlooking the *Impression*, Boecking et al. (2022) incorporated the latter section to enrich the information available for image-text alignment. In addition, since the dependency between sentences is weak (see Fig. 5), they also proposed to randomly shuffle sentences within each section. CXR-CLIP (You et al., 2023) has explored and utilized uncertainty annotations (Irvin et al., 2019). It generates prompts based on uncertainty annotations to provide supplementary information for image-text alignment.

3.4. Explicit knowledge enhancement

While studies in Section 3.2 and Section 3.3 in nature still focus on **internal** information of the dataset, some researches have investigated the integration of **external** medical knowledge to enhance the pre-training process.

Employing the unified medical language system (UMLS) as the external knowledge base (Bodenreider, 2004) for medical concepts, existing studies typically incorporate knowledge enhancement at the **subject level** and the **domain level** as shown in Fig. 9. At the subject level, ScispaCy (Neumann et al., 2019), a name entity recognition tool, is adopted to extract medical entities from each report and link them to corresponding medical concepts in the UMLS for entity disambiguation. Then, an entity-relation graph is built based on relations defined in UMLS or RadGraph (Jain et al., 2021), where the former establishes relations for general medical concepts and the latter is tailored to the needs of chest X-ray. These relations, including causal, positional, and modifying relationships, can provide an illustration of the image's visual structures and the process of human reasoning, enhancing the alignment within each image-text pair. For domain-level enhancement, knowledge is typically represented as a domain-specific knowledge graph or a descriptive knowledge prompt for the targeted medical imaging domain (e.g., chest X-ray, brain MRI), covering related organs, tissues, or clinical findings. The domain-specific knowledge graph can either be defined as a trainable symbolic graph, or a

Table 2. Overview of representative studies focusing on improving CLIP pre-training framework. CLS: classification; ZSC: zero-shot classification; SEG: segmentation; DET: detection; ITR: image-text retrieval; VQA: visual question answering; PG: phrase grounding; RG: report generation; ITC: image-text classification.

Method	Domain	Taxonomy	Evaluation tasks	Highlights
GLoRIA (Huang et al., 2021)	Chest X-ray	Multi-scale	CLS, SEG, ZSC, ITR	GLoRIA jointly learned global and local image-text representations by contrasting attention-weighted image regions with words in the paired reports.
LocalMI (Liao et al., 2021)	Chest X-ray	Multi-scale	CLS	This study proposed to estimate and optimize mutual information between local image representation and sentence-level text representation.
LoVT (Müller et al., 2022b)	Chest X-ray	Multi-scale	DET, SEG	LoVT leveraged self-attention to align sentence-level text representation and patch-level image representation. It adaptively weighed local representations via transformer's attention map.
Seibold et al. (2022)	Chest X-ray	Multi-scale	CLS	It proposed to align the image with each sentence in the report simultaneously.
PRIOR (Cheng et al., 2023)	Chest X-ray	Multi-scale	CLS, SEG, DET, ZSC, ITR	To realize more fine-grained alignment, a cross-modality conditional reconstruction module was proposed for masked image modeling and sentence prototype generation.
TIER (Palepu and Beam, 2023)	Chest X-ray	Multi-scale	CLS	A regularization strategy focusing on local image-text similarity was proposed to improve localization performance.
Müller et al. (2022a)	Chest X-ray	Multi-scale	DET, SEG	It analyzed from the view of distribution prior and argued that global-level and local-level alignment act complementarily. A local uniformity loss is hence proposed to replace local alignment.
MRM (Zhou et al., 2022b)	Chest X-ray	Multi-scale	CLS, SEG	Masked image reconstruction and report completion acted as two complementary objectives during pre-training.
Pan et al. (2022)	Placenta	Multi-scale	CLS	To address the issue of feature suppression during contrastive alignment, this study proposed a NegLogCosh similarity to replace cosine similarity.
Pan et al. (2023)	Placenta	Multi-scale	CLS	A Distributional Feature Recomposition (DFR) module was proposed to estimate the significance of each local text feature in a distribution-aware manner.
TCSA (Lei et al., 2023a)	Chest X-ray	Multi-scale Data-efficient	CLS, ZSC, ITR	Besides global-local image-text contrast, TCSA learns view-specific latent space for multi-view samples and then projects them to a common latent space, which enables the effective utilization of multi-view information.
MedCLIP (Wang et al., 2022d)	Chest X-ray	Data-efficient	CLS, ZSC, ITR	Employ inter-report semantical correlation as the soft optimization target for the alignment between image and text.
SAT (Liu et al., 2023a)	Chest X-ray	Multi-scale Data-efficient	CLS, SEG, ZSC, ITR	SAT categorized image-text pairs into positive/neutral/negative pairs given inter-report similarity and expected to more precisely alleviate false-negative problems. This technique can serve as a plug-and-play module for fine-grained contrast methods.
UMCL (Wang, 2023)	Chest X-ray	Data-efficient	CLS, ZSC, ITR	Construct positive/negative pairs through the multi-hot annotation to eliminate false negative pairs.
CXR-CLIP (You et al., 2023)	Chest X-ray	Data-efficient	CLS, ZSC, ITR	CXR-CLIP generates prompts based on ground-truth labels to provide supplementary information for alignment.
Jang et al. (2022)	Chest X-ray	Data-efficient	ZSC	This study introduced a sentence-level text augmentation technique and proposed a novel similarity function to compel the model to concentrate on perfectly negative samples instead of false-negative pairs.
IMITATE (Liu et al., 2023c)	Chest X-ray	Data-efficient	CLS, SEG, DET, ZSC	Feature maps derived from different stages of the vision model were aligned with *Findings* and *Impression* sections of medical reports simultaneously.
MGCA (Wang et al., 2022a)	Chest X-ray	Data-efficient	CLS, DET, SEG	MGCA sought to leverage disease-level semantic information to cluster samples with high semantic correlation.
BioViL (Boecking et al., 2022)	Chest X-ray	Data-efficient	CLS, SEG, ZSC, PG	Text data augmentation was adopted to boost training efficiency. Optimized using global alignment alone, BioViL still shows impressive phrase grounding performance compared with methods adopting multi-scale contrast.
BioViL-T (Bannur et al., 2023)	Chest X-ray	Data-efficient	CLS, PG, RG	It leveraged the temporal connectivity commonly present in diagnostic reports, which were usually discarded in previous studies.

set of top-K most common entity triplets (Wu et al., 2023a). The descriptive knowledge prompt usually provides detailed observations or explanations for encompassed entities. To incorporate external knowledge into the pre-training phase, an auxiliary **knowledge encoder** is usually involved. It serves to convert knowledge information into knowledge embeddings that can be interpreted by neural networks. The choice of the knowledge encoder may involve the selection of a graph neural network or a pre-trained BERT model.

ARL (Chen et al., 2022) is the representative study of subject-level knowledge enhancement. It adopts TransE (Bordes et al., 2013) algorithm to train a graph attention network (Veličković et al., 2018) as the knowledge encoder. All medical reports in the training set are pre-processed to construct subject-specific entity-relation graphs according to relations defined in the UMLS (Bodenreider, 2004). In contrast, KoBo (Zhang et al., 2023e) and FLAIR (Wang et al., 2023b) prioritize domain-level enhancement. KoBo extracts a knowledge graph containing radiological medical concepts from UMLS and utilizes CompGCN (Zhang et al., 2022c) as the knowledge encoder. It has proposed a knowledge semantic enhancement module and a knowledge semantic guidance module to mitigate negative sample noise and adjust semantic shifting, respectively. Similarly, FLAIR leverages Clini-

Table 2. (continued)

Method	Domain	Taxonomy	Evaluation tasks	Highlights
SDA-CLIP (Li et al., 2023)	Surgical video	Data-efficient	CLS	It leveraged the Kullback-Leibler divergence to evaluate the correlation.
UniBrain (Lei et al., 2023b)	Brain MRI	Knowledge Data-efficient	CLS, ZSC	• Knowledge encoder: MedKEBERT (Gu et al., 2021) • Multiple pre-defined knowledge prompts serve as the query for the diagnosis of brain disease
FLAIR (Wang et al., 2023b)	Retina	Knowledge	CLS, ZSC	• Knowledge encoder: ClinicalBERT (Alsentzer et al., 2019) • It incorporated domain knowledge explicitly through descriptive textual prompts during both pre-training and zero-shot inference.
ARL (Chen et al., 2022)	Chest X-ray	Knowledge	VQA, ITR, ITC	• Knowledge encoder: Graph Attention Network (Veličković et al., 2018). • ARL extracted the entity-relation graph of each medical report as subject-level knowledge. Both image and text embeddings were aligned with knowledge embeddings.
MedKLIP (Wu et al., 2023a)	Chest X-ray	Knowledge	CLS, SEG, ZSC, RG	• Knowledge encoder: ClinicalBERT (Alsentzer et al., 2019). • MedKLIP pre-processed raw reports and extracts medical-related triplets as the subject-level knowledge. Top-K most commonly appearing triplets were leveraged as the domain-level knowledge.
KAD (Zhang et al., 2023e)	Chest X-ray	Knowledge	ZSC	• Knowledge encoder: PubMedBERT (Gu et al., 2021). • A set of disease descriptions was pre-defined as knowledge queries, allowing for the diagnosis of unseen diseases.
KoBo (Zhang et al., 2023e)	Chest X-ray	Knowledge Data-efficient	CLS, ZSC, SEG	• Knowledge encoder: CompGCN with LTE (Zhang et al., 2022c) • Medical knowledge was incorporated to distinguish noisy negative samples.
MOTOR (Lin et al., 2023a)	Chest X-ray	Knowledge	CLS, RG, ITR, VQA	• Knowledge encoder: SciBERT (Beltagy et al., 2019) • The image embedding was first fused with knowledge embedding and then aligned with text embeddings.
PathOmics (Ding et al., 2023)	Pathology image	Others	SP	PathOmics explored genotype-phenotype interactions in complex cancer data. It aligned omics tabular data and image patches in a joint latent space via mean squared error instead of commonly adopted cosine similarity.
CMTA (Zhou and Chen, 2023)	Pathology image	Others	SP	CMTA focused on the intrinsic cross-modal correlations between genomic profile and pathology image.
SCA-Net (Chen et al., 2023b)	Surgical video	Others	RG	SCA-Net mutually aligned image and text representations with prototype representations of the other modality.
M-FLAG (Liu et al., 2023b)	Chest X-ray	Others	CLS, DET, SEG	M-FLAG kept the text encoder fixed during pre-training to alleviate the collapsed solution problem. It explicitly optimized the latent geometry towards orthogonal using the proposed optimization objective.

calBERT (Alsentzer et al., 2019) to interpret human-refined knowledge descriptions for the domain of fundus imaging, including detailed descriptions of visual features and inter-concept relationships.

In addition to the studies mentioned above, there are also studies adopting both subject-level and domain-level knowledge enhancements. MedKLIP (Wu et al., 2023a) has pre-processed and extracted entity triplets from each raw text report, constituting the subject-level knowledge. Then, the top-K most frequently occurring entities, K=75 in practical implementation, are identified to form an entity query set, functioning as the domain-level knowledge graph. The ClinicalBERT (Alsentzer et al., 2019) is adopted for subject-level and domain-level knowledge enhancements, simultaneously. This paradigm was also adopted in KAD (Zhang et al., 2023e), revealing its robustness. MOTOR (Lin et al., 2023a) adopted a symbolic graph regarding key clinical findings as the domain-level knowledge and also extracted entity-relation graphs for subject-level enhancement. The pre-trained SciBERT (Beltagy et al., 2019) was adopted as the knowledge encoder of MOTOR. The improved performance indicates the rationality of explicit knowledge enhancement in CLIP-style pre-training.

3.5. Others

While existing methods can typically be categorized according to the taxonomy previously outlined, exceptions exist. Such studies encourage the community to explore the proper combination of the three ways in future work. Moreover, there are also multiple interesting studies that may provide insight for potential research. M-FLAG (Liu et al., 2023b) focuses on the issue of collapse solution. That is, image and text features are encoded into the same constant feature embedding to minimize their distance in the latent space. It keeps the pre-trained text encoder frozen during pre-training, and adopts an orthogonality loss to encourage the orthogonality of visual representations. CMTA (Zhou and Chen, 2023) and PathOmics (Ding et al., 2023) investigate the alignment between omics tabular data and pathology images, which could potentially inspire the exploration of aligning other forms of data with images beyond diagnostic reports.

3.6. Summary

In this section, we provide a comprehensive overview of adapting CLIP pre-training for medical imaging, which is taxnomized into three categories. Despite their categorization, these methods are inherently the same – they all attempt to explore and utilize the inherent consistency relationships within medical images. The multi-scale contrast approach focuses on consistency between local image features and text features, enabling a more detailed interpretation. Simultaneously, the data-efficient contrast approach emphasizes the efficient use of data, maintaining consistency across different samples by leveraging inter-sample correlations. Lastly, the knowledge-enhanced

Table 3. Overview of representative classification applications.

Method	Organ	Imaging modality	Test dataset	Pre-trained CLIP
CheXzero (Tiu et al., 2022)	Chest	X-ray	CheXpert (Irvin et al., 2019) PadChest (Bustos et al., 2020)	Fine-tuned CLIP (Radford et al., 2021)
Seibold et al. (2022)	Chest	X-ray	CheXpert (Irvin et al., 2019) PadChest (Bustos et al., 2020) ChestX-ray14 Wang et al. (2017)	Fine-tuned CLIP (Radford et al., 2021)
Xplainer (Pellegrini et al., 2023)	Chest	X-ray	CheXpert (Irvin et al., 2019) ChestX-ray14 Wang et al. (2017)	BioViL (Boecking et al., 2022)
Kumar et al. (2022)	Chest	X-ray	MIMIC-CXR (Johnson et al., 2019)	Fine-tuned BioViL (Boecking et al., 2022)
Pham et al. (2023)	Chest	X-ray	EGD-CXR (Karargyris et al., 2021)	BiomedCLIP (Zhang et al., 2023c)
Mishra et al. (2023)	Chest	X-ray	VinDr-CXR (Nguyen et al., 2022)	Fine-tuned CLIP (Radford et al., 2021)
ETP (Liu et al., 2023d)	Heart	ECG	PTB-XL (Wagner et al., 2020) CPSC2018 (Liu et al., 2018)	An ECG CLIP trained from scratch
CITE (Zhang et al., 2023g)	Stomach	Histology	PatchGastric (Tsuneki and Kanavati, 2022)	CLIP (Radford et al., 2021)
CLIP-Lung (Lei et al., 2023c)	Lung	CT	LIDC-IDRI (Armato III et al., 2011)	CLIP (Radford et al., 2021)
Kim et al. (2023b)	Skin	Dermatology	HAM10000 (Tschandl et al., 2018)	CLIP (Radford et al., 2021)
DCPL (Cao et al., 2023)	Brain, Colorectal	MRI, Histology	BTMRI (Nickparvar, 2021) CHMNIST (Kather et al., 2016)	CLIP (Radford et al., 2021)
CoOPLVT (Baliah et al., 2023)	Eye	Fundus	EyePACS (Emma et al., 2015) APTOS (Karthik, 2019) Messidor (Decencière et al., 2014)	CLIP (Radford et al., 2021)
Byra et al. (2023)	Chest, Breast	X-ray, Ultrasound	Pneumonia (Kermany et al., 2018) UDIAT (Yap et al., 2017)	CLIP (Radford et al., 2021)

methods integrate expert-level medical knowledge, going beyond basic image-text matching to ensure that associations and interpretations are in line with the nuanced and complex knowledge of medical experts. Each of these methods contributes uniquely to the medical imaging domain, showcasing the adaptability and potential of CLIP pre-training to not only enhance traditional image-text relationships but also to introduce a new depth and precision in medical image analysis.

4. CLIP-driven applications

Leveraging large-scale text supervision, the pre-trained CLIP model effectively aligns visual features with human language, a capability that extends to medical images (referring to Fig. 4). This capability is particularly significant in clinical settings where interpretability is of importance. Meanwhile, the rich human knowledge embedded in CLIP can also act as an external supervision for annotation-demanding tasks, such as tumor segmentation. These strengths of CLIP explain the growing adoption of CLIP in various clinical applications.

4.1. Classification

The pre-training of CLIP involves image-text alignment, making it a natural fit for medical image classification. This task typically requires a global assessment of the image (e.g., determining whether it is benign or malignant or identifying specific diseases). In Table 3, we present existing studies that employ CLIP for image classification, which can generally be categorized into two approaches: zero-shot classification and context optimization. **Zero-shot classification** focuses on leveraging the diagnostic potential of the pre-trained, domain-specific CLIP model through effective prompt engineering. **Context optimization**, on the other hand, aims to adapt the non-domain-specific CLIP model to the medical domain in a manner that is both parameter-efficient and data-efficient.

4.1.1. Zero-shot classification

This research avenue attempts to fully leverage the potential of pre-trained CLIP models in clinical settings. Hence, the diagnostic performance would largely depend on the built-in knowledge, which influences the choice of domain-specific CLIP. In existing studies, a domain-specific CLIP is typically obtained by either finetuning the original CLIP on the target medical imaging domain (Tiu et al., 2022; Seibold et al., 2022; Mishra et al., 2023; Kumar et al., 2022) or by adopting an open-source specialized CLIP model (Boecking et al., 2022; Pham et al., 2023). For unique cases like electrocardiogram (ECG) (Liu et al., 2023d), where the data exists in the form of one-dimensional multi-channel signals, researchers often train ECG-specific CLIP models from scratch.

Besides the choice of pre-trained CLIP, the other key point of zero-shot classification lies in prompt engineering. While the standard CLIP zero-shot prompt described in Section 2.1 is effective for tasks like BI-RADS grading, it falls short in disease diagnosis. The shortfall is attributed to the softmax operations in the probability calculations (see Eq. 4), which treat each class as mutually exclusive. This does not align with the reality of disease diagnosis, as patients may simultaneously suffer from multiple diseases. To address this issue, CheXzero (Tiu et al., 2022) defined the positive and negative prompts (such as 'Pneumothorax' versus 'no Pneumothorax') to implement zero-shot disease diagnosis for different potential diseases in a compatible

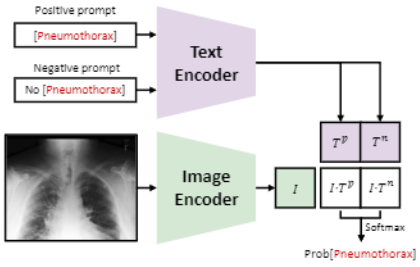


Fig. 10. Illustration of the positive/negative prompt engineering for zero-shot disease diagnosis. The diagnosis of Pneumothorax is demonstrated here, while other potential diseases can also be diagnosed in this way.

manner (see Fig. 10).

However, the diagnosis provided by CheXzero lacks explainability. To alleviate this issue, several methods have tried to incorporate detailed descriptions of the image in the prompt design, including texture, shape, and location. (Byra et al., 2023; Kim et al., 2023b; Yan et al., 2023; Liu et al., 2023f; Pellegrini et al., 2023). Among them, Pellegrini et al. (2023) introduced Xplainer, a representative method for explainable zero-shot diagnosis. Specifically, instead of directly predicting a diagnosis, they prompted the model to classify the existence of descriptive observations, for which a radiologist would look for on an X-ray scan, and used the joint probabilities of all observations to estimate the overall probability. These descriptive prompts are originally generated by ChatGPT by querying observations that may indicate specific diseases on a Chest X-ray. To improve the reliability of these prompts, radiologists are asked to further refine them based on their experience. During the diagnosis of a specific disease (e.g., Pneumothorax), all related observations would be fed into the CLIP text encoder in the form of positive/negative prompt, as shown in Fig. 11(a). According to the probability of each observation (from Prob[#1] to Prob[#N]), a joint probability can be estimated as the final result. Fig. 11(b) gives a qualitative demonstration of Xplainer's explainable diagnosis. It can be observed that Xplainer can correctly detect true positive and true negative cases. Although it fails to always make correct decisions in false positive and false negative cases, both of them show contradictory findings (e.g., bronchogram tends to co-occur with consolidation), which means they can be easily detected and corrected by radiologists. The illustration of underlying reasons in Xplainer would no doubt improve the explainability of zero-shot diagnosis.

4.1.2. Context optimization

While the concept of zero-shot disease diagnosis appears impressive and promising, its broader application in the medical imaging community is constrained by the limited availability of domain-specific CLIP models. For example, most existing open-source biomedical CLIP models are predominantly focused on Chest X-rays. Consequently, the other line of studies has turned to non-domain-specific pre-trained CLIP models, aiming to efficiently adapt these models to the context of medical imaging domains with optimal use of trainable parameters.

Although parameter-efficient tuning studies (Zhou et al., 2022d,c) have been proposed to adapt CLIP to out-of-distribution natural image datasets, none of them considers the medical imaging domain. The lack of domain awareness may lead to an inadequate perception of medical images and results in suboptimal performance. To tackle this issue, several studies focusing on context optimization (Cao et al., 2023; Zheng et al.; Baliah et al., 2023; Lei et al., 2023c) have been proposed. CLIP-Lung (Lei et al., 2023c) has proposed the channel-wise conditional prompt (CCP) for lung nodule malignancy prediction as Fig. 12 shows. Different from CoCoOp (Zhou et al., 2022c), it constructs learnable prompts based on channel-level information of feature maps. This adaptively learnable prompt successfully leads to more explainable attention maps.

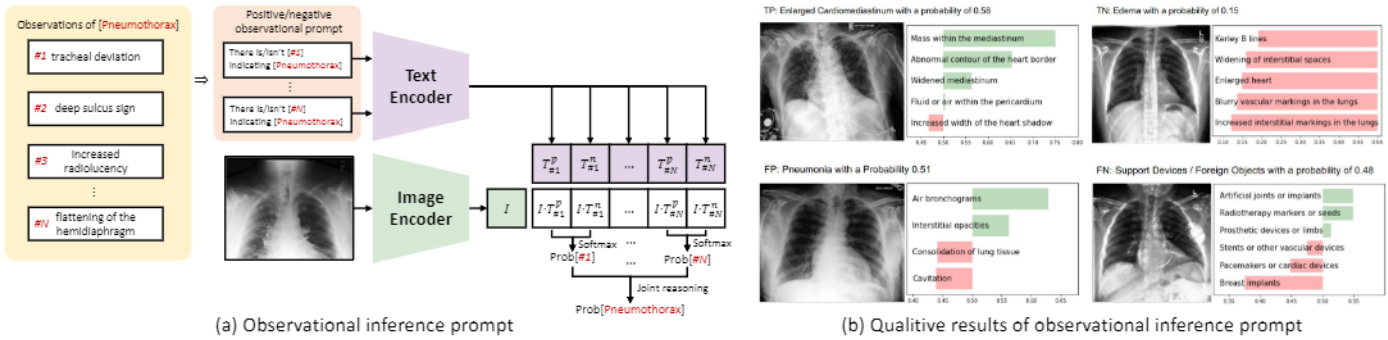
4.2. Dense prediction

Dense prediction (Zuo et al., 2022; Rao et al., 2022; Wang et al., 2021) involves generating outputs (like labels or coordinates) for every pixel or subset of pixels in an image. This task contrasts with classification, where a single label is assigned to the entire image without providing detailed spatial information. Owing to the robust feature extraction and image-text alignment capability, CLIP as well as its variants has been applied to a wide variety of dense prediction tasks. Methods involved in this line of study typically function as an auxiliary tool, providing clinicians with valuable information (e.g., potential lesion region) to support their decision-making.

4.2.1. Detection

Detection is an essential task for clinical practice such as surgical planning, pathological diagnosis, and post-operative assessment. Previous approaches (Baumgartner et al., 2021; Ickler et al., 2023; Wittmann et al., 2022; Yüksel et al., 2021) in medical image detection typically focus on leveraging image-based features extracted through various convolutional neural networks or with transformer-based architectures. These methods, while effective to a certain extent, often struggle with the nuanced and complex nature of medical images, especially in cases where visual cues are subtle or ambiguous. The pipeline for detection tasks in medical image has been significantly influenced by the advancements and integration of visual-language models, such as directly using CLIP (Müller et al., 2022c) or its extension GLIP (Li et al., 2022b). Guo et al. (2023) proposed a prompt-ensemble technique for the amalgamation of diverse textual descriptions, which fully leveraged GLIP's proficiency in interpreting complex medical scenarios. Moreover, VLPMNuD (Wu et al., 2023c) introduces GLIP for zero-shot nuclei detection in H&E stained images. It has proposed an automated prompt design method and adopted a self-training framework to polish the predicted boxes iteratively.

While object detection focuses on identifying and localizing specific and pre-defined items, such as tumors or fractures, anomaly detection is also a crucial application in medical imaging, which aims to identify deviations from the norm. AnomalyCLIP (Zhou et al., 2023) showcases CLIP's capabilities in zero-shot anomaly detection across medical domains. AnomalyCLIP employs object-agnostic text prompts to capture the essence of normality and abnormality across various images.



(a) Observational inference prompt (b) Qualitive results of observational inference prompt

Fig. 11. Overview of Xplainer. (a) Illustration of the observational inference prompt proposed by Xplainer. The diagnosis of Pneumothorax is demonstrated here, while other potential diseases can also be implemented in this way. (b) The qualitative results of Xplainer.

Table 4. Overview of representative dense prediction applications.

Method	Task	Organ	Imaging modality	Pre-trained CLIP
Guo et al. (2023)	Detection	Skin lesion, Polyp, Cell	Endoscopy, Cytology	GLIP (Li et al., 2022b)
VLPMNuD (Wu et al., 2023c)	Detection	Nuclei	Cytology images	GLIP (Li et al., 2022b)
AnomalyCLIP (Zhou et al., 2023)	Anomaly Detection	Skin, Colon, Thyroid, Chest, etc.	CT, MRI, Ultrasound, X-ray, Colonoscopy	CLIP (Radford et al., 2021)
Anand et al. (2023)	Segmentation	Spleen, Liver, Kidney, Shoulder, etc.	CT, MRI, Ultrasound	Fine-tuned CLIP (Radford et al., 2021)
MedVLSM (Poudel et al., 2023)	Segmentation	Colon, Skin, Chest, etc.	X-ray, Ultrasound, Endoscopy,	Fine-tuned CLIPSeg (Lüddecke and Ecker, 2022) Fine-tuned CRIS (Wang et al., 2022b)
SyntheticBoost (Adhikari et al., 2023)	Segmentation	Heart	Ultrasound	Fine-tuned CLIPSeg (Lüddecke and Ecker, 2022) Fine-tuned CRIS (Wang et al., 2022b)
Liu et al. (2023i)	Segmentation	Multi-organ	CT, MRI	CLIP (Radford et al., 2021)
Zhang et al. (2023h)	Segmentation	Multi-organ	CT	CLIP (Radford et al., 2021)
TPRO (Zhang et al., 2023d)	Segmentation	Breast, Lung	Histology	MedCLIP (Madhawa, 2021)
TCEIP (Yang et al., 2023)	Keypoints Localization	Tooth	Tooth crown images	CLIP (Radford et al., 2021)

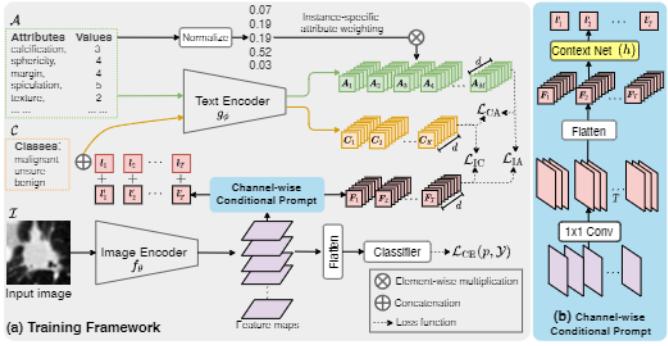
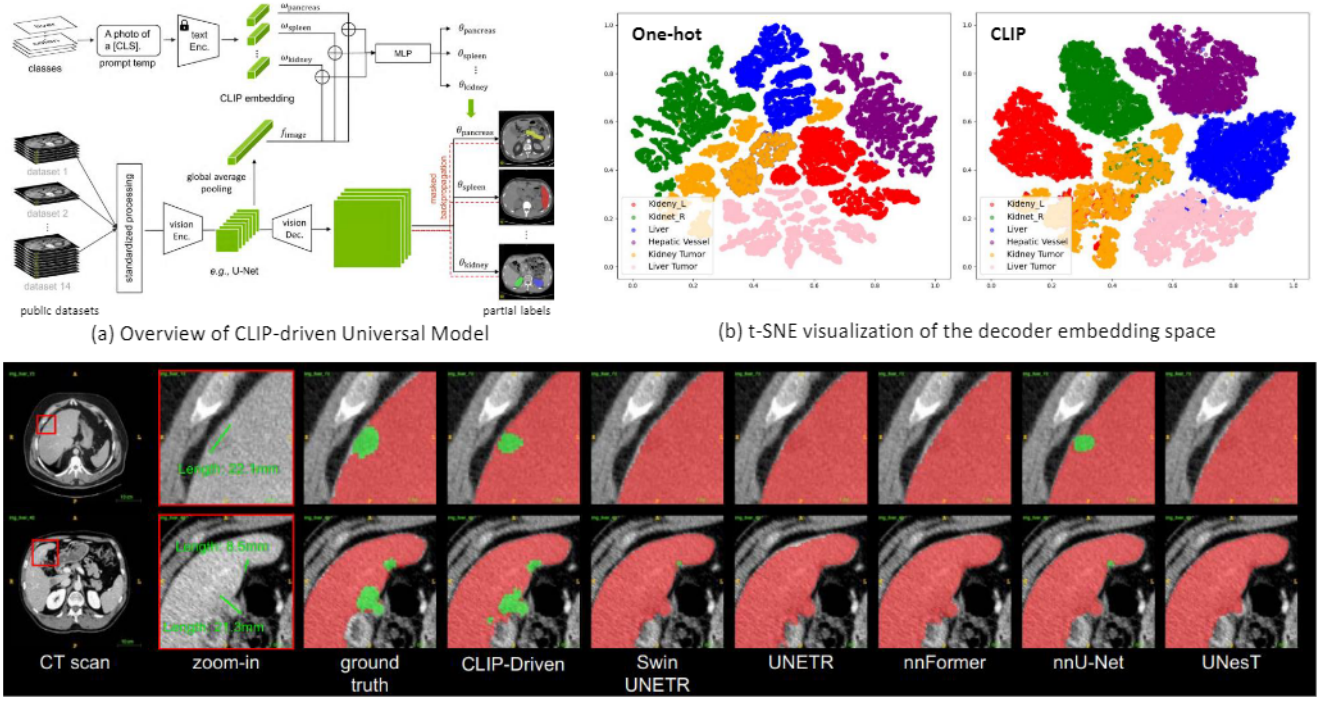


Fig. 12. Context optimization for lung nodule classification (from Lei et al. (2023c)).

This has forced CLIP to pay more attention to abnormal regions rather than to the main objects shown in the image, thereby facilitating a more generalized recognition of anomalies compared with previous methods (Zhou et al., 2022c; Sun et al., 2022; Chen et al., 2023a).

4.2.2. 2D medical image segmentation

CLIP is originally pre-trained on the 2D image domain via text supervision. Therefore, it can be easily integrated into 2D medical image segmentation with finetuning. Following this idea, Müller et al. (2022c); Anand et al. (2023) applied the CLIP pre-trained image encoder across various medical imaging modalities, including X-rays, Ultrasound, and CT/MR (by taking 3D data as 2D slices). Their works demonstrate that CLIP's image encoder, originally trained on natural images, can also deliver impressive performance in medical image segmentation tasks. Further, Poudel et al. (2023); Adhikari et al. (2023) employed both pre-trained CLIP image and text encoders to construct a vision-language segmentation model, and finetuned



(a) Overview of CLIP-driven Universal Model

(b) t-SNE visualization of the decoder embedding space

(c) Performance on liver tumor segmentation

Fig. 13. (a) Overview of CLIP-driven segmentation model for universal segmentation(Liu et al., 2023i). (b) t-SNE visualization of the decoder embedding space between one-shot task encoding and CLIP label encoding. (c) Performance on liver tumor segmentation (green for tumor and red for organ).

it to serve 2D medical image segmentation tasks.

4.2.3. 3D medical image segmentation

A growing number of publicly available datasets (Simpson et al., 2019; Heller et al., 2019; Liu et al., 2020; Bilic et al., 2023) have allowed researchers to train 3D segmentation models for anatomical structures and lesions from volume imaging data. However, most of these datasets typically only focus on certain organs or anatomical structures while all task-irrelevant organs and tumors are treated as the background, leading to the issue of partial label (Yan et al., 2020; Lyu et al., 2021). Consequently, it remains a constraint on how to break the barrier of individual datasets and fully leverage existing data cohorts to expand the capabilities of segmentation models.

Previously, DoDNet (Zhang et al., 2021) was the first universal segmentation model to introduce a dynamic segmentation head (Tian et al., 2020) tailored to specific tasks, with the task represented as a one-hot embedding. However, such label orthogonality encoding ignores the natural semantic relationship between organs. This limitation is exacerbated as the number of distinct segmentation tasks increases. Hence, label orthogonality encoding fails to generalize effectively when the diversity of tasks grows more complex. To tackle the aforementioned challenges and limitations, Liu et al. (2023i) proposed a CLIP-Driven Universal Segmentation Model (see Fig. 13(a)) by introducing the text embedding learned from CLIP to replace the one-hot encoding used in DoDNet (Zhang et al., 2021). Specifically, they have utilized the pre-trained CLIP text encoder to encode task prompts like 'liver', 'liver tumor', 'left kidney',' right kidney', 'hepatic vessel', 'kidney tumor' etc. Such embeddings are then concatenated with the pooled image features to generate the dynamic segmentation head, which is utilized to refine the segmentation results after the vision decoder. The advantage of CLIP text encoder over DoDNet's one-hot label encoding is shown in Fig. 13(b). Correlations between organs and tumors are better established with the fixed CLIP label embedding, i.e., the relationship between liver and liver tumor, liver tumor, and kidney tumor. This method provides superior performance not only in organ segmentation but also in more challenging tumor segmentation tasks, surpassing other image-only segmentation models as shown in Fig. 13(c). The CLIP-Driven Uniserval Model was compared with five vision-only SOTA segmentation methods. By reviewing the segmentation of tumors, the CLIP-Driven segmentation model succeeded in detecting small tumors, even in cases showing multiple tiny tumors, which were ignored by most image-only methods.

Following Liu et al. (2023i), Zhang et al. (2023h) extended this framework into continual learning by leveraging additional heads and text prompts to tackle new tasks. As previous studies (Ozdemir et al., 2018; Ozdemir and Goksel, 2019; Liu et al., 2022a) have focused on developing novel loss functions as additional constraints, or memory modules to preserve patterns from the original data, Zhang et al. (2023h) leveraged text prompts to represent the correlation between current task and previously learned tasks. The semantic correlation contained in text prompts enables the proposed model to filter and reserve task-specific information with superior performance.

4.2.4. Others

In weakly supervised segmentation, class activation map (CAM) (Zhou et al., 2016) is commonly used for attention localization and pseudo-label generation. Yet CAM only focuses on the most distinctive parts of the object and often leads to low-quality labels due to boundary neglection. Despite recent attempts to broaden CAM's coverage (Lee et al., 2021; Han et al., 2022; Li et al., 2022c; Zhang et al., 2022a), this fundamental problem persists. Notably, Zhang et al. (2023d) proposed to integrate language prior knowledge into weakly supervised learning to provide reliable assistance in finding object structures. Specifically, they introduce a text-prompting-based weakly supervised segmentation method (TPRO) by employing a pre-trained MedCLIP (Madhawa, 2021) to convert semantic labels into class-level embedding. An additional BioBERT (Lee et al., 2020) is also adopted to extract detailed information about the label's corresponding text descriptions. These additional text supervisions are then fused with the image features, effectively improving the quality of pseudo labels and thus providing superior performance compared to other CAM-based methods.

Considering keypoint localization, various methods (O'Neil et al., 2018; Payer et al., 2020; Wu et al., 2023b) have been developed for challenges in the medical imaging domain. While these methods have demonstrated solid performance in many cases, they still struggle to handle complex localization environments. TCEIP (Yang et al., 2023) integrates the instructional text embedding of the target region into a regression network to guide the prediction of implant position. By leveraging CLIP, TCEIP is able to interpret and process instructions such as 'left', 'middle', and 'right' alongside visual data, ensuring more precise and context-aware results. Its performance has surpassed the capabilities of previous image-only detection methods, especially in the challenging cases with multiple missing teeth or sparse teeth.

4.3. Cross-modal tasks

In addition to the previously mentioned pure vision tasks, CLIP has also propelled the development of cross-modal tasks. Here cross-modal task refers to the interaction between image and text modalities. Representative studies are shown in Table 5.

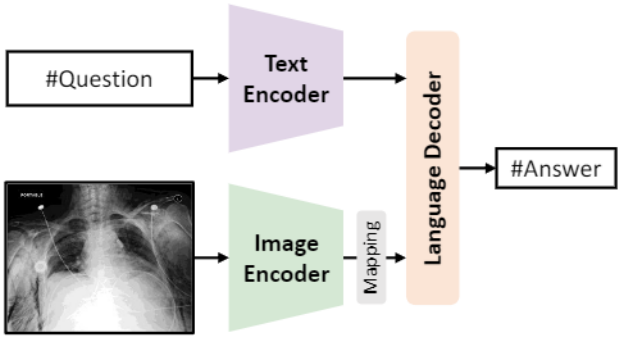


Fig. 14. Illustration of CLIP-driven methods for open-ended MedVQA.

4.3.1. Report generation

Given the time-intensive process of manually transcribing reports in clinical settings, there has been an increasing inclination toward automating the generation of medical reports (Liu et al., 2019; Yu et al., 2023). Since the effective generation of medical reports necessitates the recognition of crucial findings, attributes, and inter-finding semantic relations, CLIP is inherently suitable for this task due to its semantic awareness.

Wang et al. (2022c) adopted CLIP's vision encoder to extract semantic-aware image representations from chest X-ray, which then interacted with learnable concept embeddings and hence benefitted the performance of report generation. Keicher et al. (2022) fully leveraged the strengths of CLIP by reformulating the report generation task as a multi-label classification task, with labels indicating the presence or absence of specific findings. They compile all possible combinations of clinical findings and corresponding locations in the training set to form a prompt set. They utilize CLIP's zero-shot capability to calculate the likelihood of each prompt appearing in the image.

4.3.2. Medical visual question answering

Medical visual question answering (MedVQA) is a task that demands an in-depth understanding of both text-based questions and visual content of medical images. It has drawn attention from the community as it would lead to more efficient and accurate diagnoses and treatment decisions. Since CLIP has long been favored by its ability to align visual and text content, recent efforts have been made to apply CLIP in MedVQA.

Previous efforts have incorporated CLIP into closed-ended MedVQA tasks. These studies (Eslami et al., 2023; Liu et al., 2023e) usually only integrate CLIP's image encoder into the original MedVQA framework, aiming to enhance image representation with semantic understanding. However, they tend to overlook the comprehensive utilization of image-text alignment. Moreover, closed-ended MedVQA typically offers all potential answer options for each question, which essentially transforms the task into a classification problem. Consequently, the practical utility of these methods appears constrained due to these limitations.

Conversely, open-ended MedVQA has shown promise due to the development of CLIP. It does not pre-define options for each question, expanding its applicability to various scenarios and necessitating a heightened capacity for image-text comprehension. Hence, Zhang et al. (2023f) leveraged CLIP's image encoder and text encoder for question and image understanding, respectively, with a subsequent language decoder for answer generation. We illustrate the CLIP-driven open-ended MedVQA in Fig. 14. To mitigate the domain gap between CLIP's pre-training dataset and the current MedVQA dataset, a mapping layer is commonly employed. The question embedding and the transformed image embedding are concatenated and directly input into a language decoder, which may take the form of a multi-layer transformer or a language model, facilitating the generation of answers.

4.3.3. Image-text retrieval

Retrieval augmentation (Komeili et al., 2022), which involves supplementing data by retrieving relevant information,

Table 5. Overview of representative cross-modality applications.

Method	Tasks	Organ	Modality	Pre-trained CLIP
FlexR (Keicher et al., 2022)	Report generation	Chest	X-ray	Fine-tuned CLIP (Radford et al., 2021)
MCGN (Wang et al., 2022c)	Report generation	Chest	X-ray	CLIP(Radford et al., 2021)
X-TRA (van Sonsbeek and Worring, 2023)	Image-text Retrieval	Chest	X-ray	Fine-tuned CLIP (Radford et al., 2021), Fine-tuned PubMedCLIP (Eslami et al., 2021)
MONET (Kim et al., 2023a)	Image-text Retrieval	Skin	Dermatology	CLIP (Radford et al., 2021)
VQA-adapter (Liu et al., 2023e)	MedVQA	Multiple	Multiple	CLIP (Radford et al., 2021)
Eslami et al. (2021)	MedVQA	Multiple	Multiple	PubMedCLIP (Radford et al., 2021)
Sonsbeek et al. (van Sonsbeek et al., 2023)	MedVQA	Multiple	Multiple	CLIP (Radford et al., 2021)

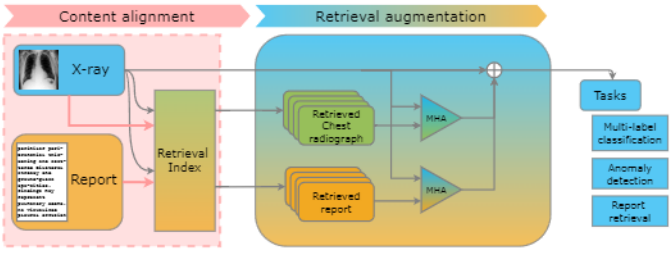


Fig. 15. Architecture overview of X-TRA (from van Sonsbeek and Worring (2023)).

allows utilization of up-to-time information from a trusted knowledge source, essentially providing a non-parametric memory expansion (Ramos et al., 2023). This approach has gained attention for its versatility, especially in the area of retrieval-augmented large language model (Zhao et al., 2023; Asai et al., 2023). However, existing retrieval methods often focus on global image features (Ionescu et al., 2023), which can lead to sub-optimal results in medical imaging. Unlike global features that may resemble across patients, subtle image details have effects on disease diagnosis and are of significance.

To address the domain shift between medical and natural images, van Sonsbeek and Worring (2023) proposed a CLIP-based multi-modal retrieval framework. This method comprises two main parts, illustrated in Fig. 15. The first part involves finetuning the original CLIP model to construct the retrieval model. Given the visual similarity of medical images and the significance of small, localized markers as disease indicators, they propose a content classifier to implement supervised content-based alignment. The second part utilizes the output of the retriever in cross-modal retrieval augmentation, enhancing downstream tasks with multi-head attention (MHA). When evaluating the performance of their retrieval method in comparison to previous approaches for disease classification and report retrieval, van Sonsbeek and Worring (2023) demonstrated a substantial performance improvement, outperforming all existing retrieval methods by a significant margin. The observed performance difference underscores the potential of CLIP in constructing a robust retrieval method.

4.4. Summary

In this section, we demonstrate some representative CLIP-driven applications to show the performance improvements under CLIP assistance. While these studies focus on various tasks, they generally indicate that the strength of a pre-trained CLIP model lies in its ability to **interpret and convey human knowledge**. As best illustrated in Pellegrini et al. (2023); Zhang et al. (2023d); Yang et al. (2023), where descriptive text prompts are fed to the CLIP, the experimental results showcase CLIP's adeptness in comprehending the semantics embedded within prompts and effectively conveying semantics to other components within the framework. It implies that CLIP-driven applications can be adaptable to different groups of patients by simply modifying the specific content of the input prompt, which is beneficial for the diagnosis or prognosis of diseases having regional or age-related differences. For instance, diseases like sepsis often exhibit distinct progression patterns among different racial groups (Khoshnevisan and Chi, 2021; Tintinalli et al., 2016), while the survival rate of Community-Acquired Pneumonia correlates with the patient's age (Stupka et al., 2009; Ravioli et al., 2022). By adjusting the content within the descriptive prompt, the developed CLIP-driven application can seamlessly transition between various groups without necessitating re-training or fine-tuning.

5. Discussions and future directions

The aforementioned sections have delved into research studies that either leverage a refined CLIP pre-training paradigm or showcase CLIP-driven clinical applications within the medical imaging community. Despite significant strides, there still exist several challenges and open questions. In this section, we summarize key challenges and offer discussions on potential future directions.

Inconsistency between pre-training and application. Some readers might notice that the two sections – refined CLIP pre-training and CLIP-driven application – are currently uncoordinated. Ideally, refined CLIP pre-training is responsible for offering the domain-specific CLIP for CLIP-driven applications. Unfortunately, the CLIP-driven applications covered in this survey still primarily rely on OpenAI's pre-trained CLIP (trained on natural image-text datasets). This would significantly limit their performance in clinical practice. In Fig. 16,

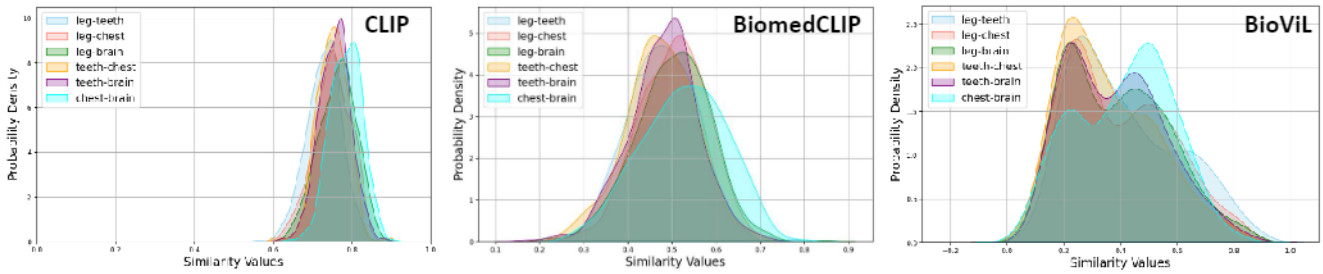


Fig. 16. Comparison between non-domain-specific and domain-specific pre-training. CLIP (OpenAI) tends to present high inter-disease similarity, which is significantly alleviated in BiomedCLIP and BioViL, revealing the irrationality of adopting CLIP (OpenAI) in medical applications.

we select the top 20 to 30 most frequently occurring diseases for each organ, computing inter-organ similarity distributions of textual disease embeddings. Despite the inherent semantic differences among the selected diseases, the resulting similarity distribution reveals a challenge for CLIP in effectively discriminating them, as the distribution curves overlap significantly in the left panel of Fig. 16. Notably, this issue is markedly mitigated in BiomedCLIP, underscoring the importance of domain-specific CLIP pre-training. Simultaneously, BioViL, a model tailored for Chest X-ray analysis, demonstrates the best performance. This observation underscores the efficacy of specialized pre-trained CLIP model, emphasizing their ability to outperform generalized counterparts, especially in contexts where fine-grained discrimination among diseases is crucial. Hence, we argue that future work focusing on CLIP-driven applications should adopt a pre-trained CLIP specific to their target organ. Even for research studies focusing on context optimization (see in Section 4.1.2), which aim to efficiently fine-tune a non-domain-specific CLIP to specific medical imaging scenarios, we still recommend the use of BiomedCLIP (Zhang et al., 2023c) rather than OpenAI's general-purpose CLIP.

Incomprehensive evaluation of refined pre-training. As previously illustrated in Section 3.5, studies focusing on refined CLIP pre-training commonly assess the quality of pre-training through various evaluation tasks. These evaluation tasks include those primarily aimed at evaluating vision encoders, such as CLS/ZSC/SEG/DET, and tasks that simultaneously assess image and text encoders, such as ITR/VQA/PG. However, the issue lies in the fact that existing studies tend to favor vision-biased evaluation tasks, somewhat overlooking the evaluation of text encoders, which is evidently demonstrated in Table 2. The essence of CLIP lies in the alignment between images and text. Only when both the vision encoder and the text encoder demonstrate high quality, they can effectively function as foundational components in domain-specific CLIP-driven applications. BioViL (Boecking et al., 2022) and BioViL-T (Bannur et al., 2023) deserve recognition as they implement relatively comprehensive evaluations for their pre-trained vision and text encoder, and BioViL has been adopted in some CLIP-driven applications due to its robust performance(see Table 3). For future work, we encourage researchers to conduct more comprehensive evaluations. These evaluations could encompass their performance across tasks such as report generation (IU-Xray (Pavlopoulos et al., 2019)), phrase grounding (MS-CXR (Boecking et al., 2022)), and VQA (EHRXQA (Bae et al., 2023)).

Limited scope of refined CLIP pre-training. Presently, domain-specific CLIP models are tailored specifically for Chest X-rays within medical imaging, leaving other prevalent image types like mammography, knee MRI, and histology without adequate research. This limitation is primarily attributed to the scarcity of publicly available datasets. Previously, MIMIC-CXR stood as the predominant large-scale dataset for image-text alignment in medical imaging. However, with the recent release of FFA-IR (in 2021) and two additional histology datasets (in 2023), there is a pressing need for further advancements in CLIP pre-training that prioritize these two domains rather than solely focusing on chest X-rays. These two domains also have their specific challenges, which make them different from Chest X-ray. The FFA-IR dataset is featured by multi-view diagnosis. A fundus fluorescein angiography (FFA) examination may include tens or even more images to comprehensively assess the status of the eye's vascular system, which is much larger than that of Chest X-ray (only 1 or 2 views). At the same time, histology images are characterized by giga-pixel resolution and are usually processed at the patch level, which encourages the investigation of patch-level alignment and slide-level alignment. We expect that future work could develop more sophisticated CLIP-style pre-training methods to address these issues on domains beyond Chest X-ray.

Exploring the potential of metadata. The potential of metadata has been largely underexplored. This type of data typically includes a range of patient attributes, many of which may exhibit a strong correlation with visual morphology. For example, a common attribute such as a patient's age can yield significant insights into brain tissue segmentation. Fig. 17 illustrates the varied morphology and tissue contrast observed throughout the human lifespan, highlighting the potential importance of age information in brain-related tasks. The integration of metadata into prompts can significantly enhance the comprehension and interpretation capabilities of deep learning-based models. Unlike previous methods that encoded metadata directly using a multi-layer perceptron (Cetin et al., 2023; Zhao et al., 2021), CLIP can offer a more semantically rich approach to text embeddings due to large-scale pre-training. This suggests a promising avenue for future research and exploration in the field.

Incorporation of high-order correlation. Existing CLIP-

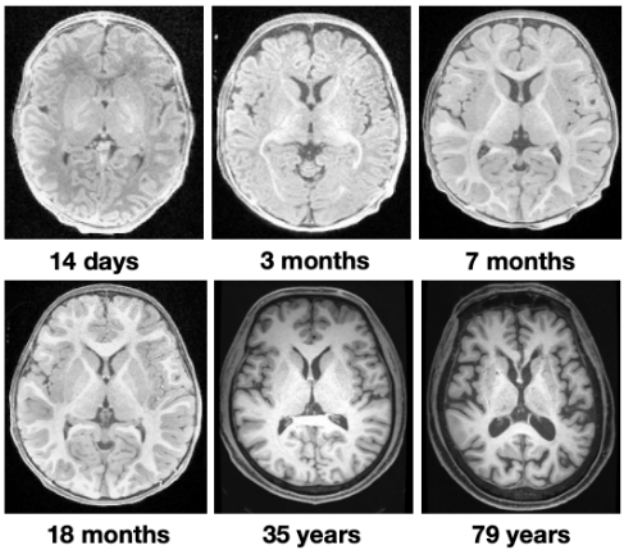


Fig. 17. Brain morphology and intensity contrast vary as a function of age. Each figure represents T1-weighted MR image with one-time point.

style pre-training methods in the medical imaging domain still predominantly adhere to orthogonal alignment between images and texts, lacking explicit consideration for inter-sample correlation. This conventional practice involves orthogonal alignment of each image with its corresponding ground-truth report. As elucidated in Section 3.3, this approach may result in performance degradation due to substantial semantic overlap among medical samples. Although attempts have been made to mitigate this issue through inter-report semantic similarity, their success has primarily relied on handcrafted rules and low-order inter-report correlation. Consequently, the integration of high-order correlation emerges as a promising solution.

The effectiveness of high-order correlation has been well-established in tasks involving multiple information sources or those requiring interpretations of complex relationships, including brain network analysis (Ohiorhenuan et al., 2010; Chen et al., 2016; Zhang et al., 2017; Owen et al., 2021; Liu et al., 2023j), multi-label classification (Zhang et al., 2014; Nazmi et al., 2020; Si et al., 2023), and multi-view clustering (Li et al., 2022d). Likewise, medical image-text pre-training involves two kinds of information (i.e., image and text), and their semantic correlations need to be further explored. Hence, we expect that future studies will devote increased attention to comprehending the intricate semantic correlations between medical image-text samples, addressing the challenge of orthogonal image-text alignment via the methodology of high-order correlation.

Beyond image-text alignment. The philosophy of CLIP revolves around achieving alignment between different modalities, specifically images and text. Alignment, in this context, refers to the model's ability to understand and establish meaningful connections between visual and textual content. By comprehending the intrinsic connections between visual and textual information, CLIP can perform exceptionally well in various cross-modal applications, reflecting a broader trend. Extending the alignment philosophy of CLIP to other multimodal medical imaging can be a promising direction. Medical imaging often involves various modalities like X-rays, MRIs, CT scans, each providing unique insights into different aspects of a patient's condition. Analogous to CLIP's methodology, aligning these diverse imaging modalities within a unified embedding space could potentially revolutionize medical data analysis, presenting a progressive direction for medical research and diagnostics.

6. Conclusion

In conclusion, we present the first comprehensive review of the CLIP in medical imaging. Starting by introducing the foundational concepts that underpin CLIP's success, we then delve into an extensive literature review from two aspects: refined CLIP pre-training methods and diverse CLIP-driven applications. For refined CLIP pre-training, our survey offers a structured taxonomy based on the unique challenges that CLIP pre-training encountered in the medical imaging domain, aiming to chart a clear pathway for researchers to advance this field progressively. In exploring CLIP-driven applications, we compare these approaches against solely vision-driven methods, emphasizing the added value that pre-trained CLIP models could bring. Notably, through thoughtful design, they could serve as valuable supplementary supervision signals, significantly enhancing the performance across various tasks. Beyond simply reviewing existing studies in these two sections, we also discuss common issues, laying the groundwork for future directions. By illuminating the potential and challenges of employing CLIP in medical imaging, we aim to push the field forward, encouraging innovation and paving the way for human-aligned medical AI.

References

Adhikari, R., Dhakal, M., Thapaliya, S., Poudel, K., Bhandari, P., Khanal, B., 2023. Synthetic boost: Leveraging synthetic data for enhanced vision-language segmentation in echocardiography, in: International Workshop on Advances in Simplifying Medical Ultrasound, Springer. pp. 89–99.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M., 2019. Publicly available clinical, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics.

Anand, D., Singhal, V., Shanbhag, D.D., KS, S., Patil, U., Bhushan, C., Manickam, K., Gui, D., Mullick, R., Gopal, A., et al., 2023. One-shot localization and segmentation of medical images with foundation models. arXiv preprint arXiv:2310.18642 .

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38, 915–931.

Asai, A., Min, S., Zhong, Z., Chen, D., 2023. Retrieval-based language models and applications, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts), pp. 41–46.

Bae, S., Kyung, D., Ryu, J., Cho, E., Lee, G., Kweon, S., Oh, J., Ji, L., Chang, E.I.C., Kim, T., Choi, E., 2023. EHRXQA: A multi-modal question answering dataset for electronic health records with chest x-ray images, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track. URL: https://openreview.net/forum?id=Pk2x7FPuZ4.

Baliah, S., Maani, F.A., Sanjeev, S., Khan, M.H., 2023. Exploring the transfer learning capabilities of clip in domain generalization for diabetic retinopathy, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 444–453.

Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F., 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. Advances in Neural Information Processing Systems 35, 33781–33794.

Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al., 2023. Learning to exploit temporal structure for biomedical vision-language processing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15016–15027.

Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H., 2021. nndetection: a self-configuring method for medical object detection, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer. pp. 530–539.

Beltagy, I., Lo, K., Cohan, A., 2019. Scibert: A pretrained language model for scientific text, in: EMNLP, Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/D19-1371.

Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al., 2023. The liver tumor segmentation benchmark (lits). Medical Image Analysis 84, 102680.

Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K., 1987. Occam's razor. Information processing letters 24, 377–380.

Bodenreider, O., 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research 32, D267–D270.

Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al., 2022. Making the most of text semantics to improve biomedical vision–language processing, in: European conference on computer vision, Springer. pp. 1–21.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 26.

Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M., 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis 66, 101797.

Byra, M., Rachmadi, M.F., Skibbe, H., 2023. Few-shot medical image classification with simple shape and texture text descriptors using vision-language models. arXiv preprint arXiv:2308.04005 .

Cao, Q., Xu, Z., Chen, Y., Ma, C., Yang, X., 2023. Domain-controlled prompt learning. arXiv preprint arXiv:2310.07730 .

Cetin, I., Stephens, M., Camara, O., Ballester, M.A.G., 2023. Attri-vae: Attribute-based interpretable representations of medical images with variational autoencoders. Computerized Medical Imaging and Graphics 104, 102158.

Chen, X., Han, Y., Zhang, J., 2023a. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. arXiv preprint arXiv:2305.17382 .

Chen, X., Zhang, H., Gao, Y., Wee, C.Y., Li, G., Shen, D., Initiative, A.D.N., 2016. High-order resting-state functional connectivity network for mci classification. Human brain mapping 37, 3282–3296.

Chen, Z., Guo, Q., Yeung, L.K., Chan, D.T., Lei, Z., Liu, H., Wang, J., 2023b. Surgical video captioning with mutual-modal concept alignment, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 24–34.

Chen, Z., Li, G., Wan, X., 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 5152–5161.

Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., Tang, X., 2023. Prior: Prototype representation joint learning from medical images and reports, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21361–21371.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J., 2023. Reproducible scaling laws for contrastive language-image learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2818–2829.

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al., 2014. Feedback on a publicly distributed image database: the messidor database. Image Analysis & Stereology 33, 231–234.

Ding, K., Zhou, M., Metaxas, D.N., Zhang, S., 2023. Pathology-and-genomics multimodal transformer for survival outcome prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 622–631.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Emma, D., Jared, Jorge, Will, C., 2015. Diabetic retinopathy detection. URL: https://kaggle.com/competitions/diabetic-retinopathy-detection.

Eslami, S., Meinel, C., De Melo, G., 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain?, in: Findings of the Association for Computational Linguistics: EACL 2023, pp. 1151–1163.

Eslami, S., de Melo, G., Meinel, C., 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906 .

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) 3, 1–23.

Guo, M., Yi, H., Qin, Z., Wang, H., Men, A., Lao, Q., 2023. Multiple prompt fusion for zero-shot lesion detection using vision-language models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 283–292.

Han, C., Lin, J., Mai, J., Wang, Y., Zhang, Q., Zhao, B., Chen, X., Pan, X., Shi, Z., Xu, Z., et al., 2022. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Medical Image Analysis 80, 102487.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 .

Huang, S.C., Shen, L., Lungren, M.P., Yeung, S., 2021. 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, pp. 3942–3951.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023a. A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine 29, 2307–2316.

Huang, Z., Zhang, X., Zhang, S., 2023b. Kiut: Knowledge-injected u-transformer for radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19809–19818.

Ickler, M.K., Baumgartner, M., Roy, S., Wald, T., Maier-Hein, K.H., 2023. Taming detection transformers for medical object detection, in: BVM Workshop, Springer. pp. 183–188.

Ikezogwo, W.O., Seyfioglu, M.S., Ghezloo, F., Geva, D.S.C., Mohammed, F.S., Anand, P.K., Krishna, R., Shapiro, L., 2023. Quilt-1m: One million image-text pairs for histopathology. arXiv preprint arXiv:2306.11207 .

Ionescu, B., Müller, H., Drăgulinescu, A.M., Yim, W.W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M., Rückert, J., García Seco de Herrera, A., et al., 2023. Overview of the imageclef 2023: Multimedia retrieval in medical, social media and internet applications, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer. pp. 370–396.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI conference on artificial intelligence, pp. 590–597.

Jain, S., Agrawal, A., Saporta, A., Truong, S., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C., Rajpurkar, P., 2021. Radgraph: Extracting clinical entities and relations from radiology reports, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). URL: https://openreview.net/forum?id=pMWtc5NKd7V.

Jang, J., Kyung, D., Kim, S.H., Lee, H., Bae, K., Choi, E., 2022. Signif-

icantly improving zero-shot x-ray pathology classification via fine-tuning pre-trained image-text encoders. arXiv preprint arXiv:2212.07050 .

Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S., 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data 6, 317.

Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J.T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E.A., et al., 2021. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. Scientific data 8, 1–18.

Karthik, Maggie, S.D., 2019. Aptos 2019 blindness detection. URL: `https://kaggle.com/competitions/aptos2019-blindness-detection`.

Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. Scientific reports 6, 27988.

Keicher, M., Zaripova, K., Czempiel, T., Mach, K., Khakzar, A., Navab, N., 2022. Flexr: Few-shot classification with language embeddings for structured reporting of chest x-rays. arXiv preprint arXiv:2203.15723 .

Kenton, J.D.M.W.C., Toutanova, L.K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, pp. 4171–4186.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. cell 172, 1122–1131.

Khoshnevisan, F., Chi, M., 2021. Unifying domain adaptation and domain generalization for robust prediction across minority racial groups, in: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21, Springer. pp. 521–537.

Kim, C., Gadgil, S.U., DeGrave, A.J., Cai, Z.R., Daneshjou, R., Lee, S.I., 2023a. Fostering transparent medical image ai via an image-text foundation model grounded in medical literature. medRxiv .

Kim, I., Kim, J., Choi, J., Kim, H.J., 2023b. Concept bottleneck with visual concept filtering for explainable medical image classification. arXiv preprint arXiv:2308.11920 .

Komeili, M., Shuster, K., Weston, J., 2022. Internet-augmented dialogue generation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8460–8478.

Kumar, B., Palepu, A., Tuwani, R., Beam, A., 2022. Towards reliable zero shot classification in self-supervised models with conformal prediction. arXiv preprint arXiv:2210.15805 .

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240.

Lee, S., Lee, M., Lee, J., Shim, H., 2021. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5495–5505.

Lei, H., Huang, H., Yang, B., Cui, G., Wang, R., Wu, D., Li, Y., 2023a. Tcsa: A text-guided cross-view medical semantic alignment framework for adaptive multi-view visual representation learning, in: International Symposium on Bioinformatics Research and Applications, Springer. pp. 136–149.

Lei, J., Dai, L., Jiang, H., Wu, C., Zhang, X., Zhang, Y., Yao, J., Xie, W., Zhang, Y., Li, Y., et al., 2023b. Unibrain: Universal brain mri diagnosis with hierarchical knowledge-enhanced pre-training. arXiv preprint arXiv:2309.06828 .

Lei, Y., Li, Z., Shen, Y., Zhang, J., Shan, H., 2023c. Clip-lung: Textual knowledge-guided lung nodule malignancy prediction, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland, Cham. pp. 403–412.

Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R., 2022a. Language-driven semantic segmentation, in: International Conference on Learning Representations. URL: `https://openreview.net/forum?id=RriDjddCLN`.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al., 2022b. Grounded language-image pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10965–10975.

Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al., 2021. Ffa-ir: Towards an explainable and reliable medical report generation benchmark, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Li, Y., Jia, S., Song, G., Wang, P., Jia, F., 2023. Sda-clip: surgical visual domain adaptation using video and text labels. Quantitative Imaging in Medicine and Surgery 13, 6989.

Li, Y., Yu, Y., Zou, Y., Xiang, T., Li, X., 2022c. Online easy example mining for weakly-supervised gland segmentation from histology images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 578–587.

Li, Z., Tang, C., Zheng, X., Liu, X., Zhang, W., Zhu, E., 2022d. High-order correlation preserved incomplete multi-view subspace clustering. IEEE Transactions on Image Processing 31, 2067–2080.

Liao, R., Moyer, D., Cha, M., Quigley, K., Berkowitz, S., Horng, S., Golland, P., Wells, W.M., 2021. Multimodal representation learning via maximization of local mutual information, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 273–283.

Lin, B., Chen, Z., Li, M., Lin, H., Xu, H., Zhu, Y., Liu, J., Cai, W., Yang, L., Zhao, S., et al., 2023a. Towards medical artificial general intelligence via knowledge-enhanced multimodal pretraining. arXiv preprint arXiv:2304.14204 .

Lin, J., Gong, S., 2023. Gridclip: One-stage object detection by grid-level clip representation learning. arXiv preprint arXiv:2303.09252 .

Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W., 2023b. Pmc-clip: Contrastive language-image pre-training using biomedical documents. arXiv preprint arXiv:2303.07240 .

Liu, B., Lu, D., Wei, D., Wu, X., Wang, Y., Zhang, Y., Zheng, Y., 2023a. Improving medical vision-language contrastive pretraining with semantics-aware triage. IEEE Transactions on Medical Imaging .

Liu, C., Cheng, S., Chen, C., Qiao, M., Zhang, W., Shah, A., Bai, W., Arcucci, R., 2023b. 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, Springer. pp. 637–647.

Liu, C., Cheng, S., Shi, M., Shah, A., Bai, W., Arcucci, R., 2023c. Imitate: Clinical prior guided hierarchical vision-language pre-training. arXiv preprint arXiv:2310.07355 .

Liu, C., Wan, Z., Cheng, S., Zhang, M., Arcucci, R., 2023d. Etp: Learning transferable ecg representations via ecg-text pre-training. arXiv preprint arXiv:2309.07145 .

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., et al., 2018. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics 8, 1368–1373.

Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M., 2019. Clinically accurate chest x-ray report generation, in: Machine Learning for Healthcare Conference, PMLR. pp. 249–269.

Liu, J., Hu, T., Zhang, Y., Feng, Y., Hao, J., Lv, J., Liu, Z., 2023e. Parameter-efficient transfer learning for medical visual question answering. IEEE Transactions on Emerging Topics in Computational Intelligence .

Liu, J., Hu, T., Zhang, Y., Gai, X., Feng, Y., Liu, Z., 2023f. A chatgpt aided explainable framework for zero-shot medical image diagnosis. arXiv preprint arXiv:2307.01981 .

Liu, J., Wang, Z., Ye, Q., Chong, D., Zhou, P., Hua, Y., 2023g. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. arXiv preprint arXiv:2310.17956 .

Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023h. Clip-driven universal model for organ segmentation and tumor detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164.

Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023i. Clip-driven universal model for organ segmentation and tumor detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164.

Liu, P., Wang, X., Fan, M., Pan, H., Yin, M., Zhu, X., Du, D., Zhao, X., Xiao, L., Ding, L., et al., 2022a. Learning incrementally to segment multiple organs in a ct image, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 714–724.

Liu, Q., Dou, Q., Yu, L., Heng, P.A., 2020. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE transactions on medical imaging 39, 2713–2724.

Liu, Y., Liu, M., Zhang, Y., Shen, D., 2023j. Learning hierarchical-order func-

tional connectivity networks for mild cognitive impairment diagnosis, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022b. A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976–11986.

Lüddecke, T., Ecker, A., 2022. Image segmentation using text and image prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7086–7096.

Luo, H., Bao, J., Wu, Y., He, X., Li, T., 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation, in: International Conference on Machine Learning, PMLR. pp. 23033–23044.

Luo, L., Chen, H., Xiao, Y., Zhou, Y., Wang, X., Vardhanabhuti, V., Wu, M., Han, C., Liu, Z., Fang, X.H.B., et al., 2022. Rethinking annotation granularity for overcoming shortcuts in deep learning–based radiograph diagnosis: A multicenter study. Radiology: Artificial Intelligence 4, e210299.

Lyu, F., Yang, B., Ma, A.J., Yuen, P.C., 2021. A segmentation-assisted model for universal lesion detection with partial labels, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer. pp. 117–127.

Madhawa, K., 2021. Medclip: Fine-tuning a clip model on the roco medical dataset. URL: `https://github.com/Kaushalya/medclip`.

Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L., 2018. Y-net: joint segmentation and classification for diagnosis of breast biopsy images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, Springer. pp. 893–901.

Mishra, A., Mittal, R., Jestin, C., Tingos, K., Rajpurkar, P., 2023. Improving zero-shot detection of low prevalence chest pathologies using domain pre-trained language models. arXiv preprint arXiv:2306.08000 .

Müller, P., Kaissis, G., Rueckert, D., 2022a. The role of local alignment and uniformity in image-text contrastive learning on medical images. arXiv preprint arXiv:2211.07254 .

Müller, P., Kaissis, G., Zou, C., Rueckert, D., 2022b. Joint learning of localized representations from medical images and reports, in: European Conference on Computer Vision, Springer. pp. 685–701.

Müller, P., Kaissis, G., Zou, C., Rueckert, D., 2022c. Radiological reports improve pre-training for localized imaging tasks on chest x-rays, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 647–657.

Müller, P., Meissen, F., Brandt, J., Kaissis, G., Rueckert, D., 2023. Anatomy-driven pathology detection on chest x-rays, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 57–66.

Nazmi, S., Yan, X., Homaifar, A., Doucette, E., 2020. Evolving multi-label classification rules by exploiting high-order label correlations. Neurocomputing 417, 176–186.

Neumann, M., King, D., Beltagy, I., Ammar, W., 2019. Scispacy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669 .

Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al., 2022. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data 9, 429.

Nickparvar, M., 2021. Brain tumor mri dataset. URL: `https://www.kaggle.com/dsv/2645886`, doi:`10.34740/KAGGLE/DSV/2645886`.

Ohiorhenuan, I.E., Mechler, F., Purpura, K.P., Schmid, A.M., Hu, Q., Victor, J.D., 2010. Sparse coding and high-order correlations in fine-scale cortical networks. Nature 466, 617–621.

O'Neil, A.Q., Kascenas, A., Henry, J., Wyeth, D., Shepherd, M., Beveridge, E., Clunie, L., Sansom, C., Seduikyte Keith Muir, E., Poole, I., 2018. Attaining human-level performance with atlas location autocontext for anatomical landmark detection in 3d ct data, in: Proceedings of the European conference on computer vision (ECCV) Workshops, pp. 0–0.

Ouyang, X., Karanam, S., Wu, Z., Chen, T., Huo, J., Zhou, X.S., Wang, Q., Cheng, J.Z., 2020. Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis. IEEE transactions on medical imaging .

Owen, L.L., Chang, T.H., Manning, J.R., 2021. High-level cognition during story listening is reflected in high-order dynamic correlations in neural activity patterns. Nature Communications 12, 5728.

Ozdemir, F., Fuernstahl, P., Goksel, O., 2018. Learn the new, keep the old: Extending pretrained models with new anatomy and images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11, Springer. pp. 361–369.

Ozdemir, F., Goksel, O., 2019. Extending pretrained segmentation networks with additional anatomical structures. International journal of computer assisted radiology and surgery 14, 1187–1195.

Palepu, A., Beam, A., 2023. Tier: Text-image entropy regularization for medical clip-style models. Proceedings of Machine Learning Research LEAVE UNSET 1, 21.

Pan, Y., Cai, T., Mehta, M., Gernand, A.D., Goldstein, J.A., Mithal, L., Mwinyelle, D., Gallagher, K., Wang, J.Z., 2023. Enhancing automatic placenta analysis through distributional feature recomposition in vision-language contrastive learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 116–126.

Pan, Y., Gernand, A.D., Goldstein, J.A., Mithal, L., Mwinyelle, D., Wang, J.Z., 2022. Vision-language contrastive learning approach to robust automatic placenta analysis using photographic images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 707–716.

Pavlopoulos, J., Kougia, V., Androutsopoulos, I., 2019. A survey on biomedical image captioning, in: Proceedings of the second workshop on shortcomings in vision and language, pp. 26–36.

Payer, C., Stern, D., Bischof, H., Urschler, M., 2020. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net., in: VISIGRAPP (5: VISAPP), pp. 124–133.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in context (roco): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer. pp. 180–189.

Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N., 2023. Xplainer: From x-ray observations to explainable zero-shot diagnosis. arXiv preprint arXiv:2303.13391 .

Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z., 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings 2018, 188.

Pham, T.T., Brecheisen, J., Nguyen, A., Nguyen, H., Le, N., 2023. Decoding radiologists intense focus for accurate cxr diagnoses: A controllable and interpretable ai system. arXiv preprint arXiv:2309.13550 .

Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., Khanal, B., 2023. Exploring transfer learning in medical image segmentation using vision-language models. arXiv preprint arXiv:2308.07706 .

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR. pp. 8748–8763.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 1, 3.

Ramos, R., Martins, B., Elliott, D., Kementchedjhieva, Y., 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2840–2849.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J., 2022. Denseclip: Language-guided dense prediction with context-aware prompting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18082–18091.

Ravioli, S., Germann, C., Gygli, R., Exadaktylos, A.K., Lindner, G., 2022. Age-and sex-related differences in community-acquired pneumonia at presentation to the emergency department: a retrospective cohort study. European journal of emergency medicine 29, 366–372.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems 35, 25278–25294.

Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J., 2022. Breaking with fixed set pathology recognition through report-guided contrastive training, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 690–700.

Si, C., Jia, Y., Wang, R., Zhang, M.L., Feng, Y., Chongxiao, Q., 2023. Multi-label classification with high-rank and high-order label correlations. IEEE Transactions on Knowledge and Data Engineering .

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 .

van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M., 2023. Open-ended medical visual question answering through prefix tuning of language models. arXiv preprint arXiv:2303.05977 .

van Sonsbeek, T., Worring, M., 2023. X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 471–482.

Stupka, J.E., Mortensen, E.M., Anzueto, A., Restrepo, M.I., 2009. Community-acquired pneumonia in elderly patients. Aging health 5, 763–774.

Subramanian, S., Wang, L.L., Mehta, S., Bogin, B., van Zuylen, M., Parasa, S., Singh, S., Gardner, M., Hajishirzi, H., 2020. Medicat: A dataset of medical images, captions, and textual references. arXiv preprint arXiv:2010.06000 .

Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y., 2023. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 .

Sun, X., Hu, P., Saenko, K., 2022. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. Advances in Neural Information Processing Systems 35, 30569–30582.

Tanida, T., Müller, P., Kaissis, G., Rueckert, D., 2023. Interactive and explainable region-guided radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7433–7442.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems 30.

Tian, Z., Shen, C., Chen, H., 2020. Conditional convolutions for instance segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer. pp. 282–298.

Tintinalli, J.E., Stapczynski, J.S., Ma, O., Yealy, D., Meckler, G., Cline, D., 2016. Tintinalli's Emergency Medicine: A Comprehensive Study Guide, 8e. McGraw Hill Education.

Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P., 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering 6, 1399–1406.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 1–9.

Tsuneki, M., Kanavati, F., 2022. Inference of captions from histopathological patches, in: International Conference on Medical Imaging with Deep Learning, PMLR. pp. 1235–1250.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2018. Graph attention networks, in: International Conference on Learning Representations.

Vinker, Y., Pajouheshgar, E., Bo, J.Y., Bachmann, R.C., Bermano, A.H., Cohen-Or, D., Zamir, A., Shamir, A., 2022. Clipasso: Semantically-aware object sketching. ACM Transactions on Graphics (TOG) 41, 1–11.

Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T., 2020. Ptb-xl, a large publicly available electrocardiography dataset. Scientific data 7, 154.

Wallis, A., McCoubrie, P., 2011. The radiology report—are we getting the message across? Clinical radiology 66, 1015–1022.

Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L., 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems 35, 33536–33549.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097–2106.

Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2021. Dense contrastive learning for self-supervised visual pre-training, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR).

Wang, Y., 2023. Unified medical image-text-label contrastive learning with continuous prompt. arXiv preprint arXiv:2307.05920 .

Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., Tan, T., 2023a. Improving zero-shot generalization for clip with synthesized prompts, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3032–3042.

Wang, Z., Liu, C., Zhang, S., Dou, Q., 2023b. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T., 2022b. Cris: Clip-driven referring image segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11686–11695.

Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L., 2022c. A medical semantic-assisted transformer for radiographic report generation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 655–664.

Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022d. Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 .

Wittmann, B., Navarro, F., Shit, S., Menze, B., 2022. Focused decoding enables 3d anatomical detection by transformers. arXiv preprint arXiv:2207.10774 .

Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023a. Medklip: Medical knowledge enhanced language-image pre-training. medRxiv , 2023–01.

Wu, H., Zhang, J., Fang, Y., Liu, Z., Wang, N., Cui, Z., Shen, D., 2023b. Multi-view vertebra localization and identification from ct images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 136–145.

Wu, Y., Zhou, Y., Saiyin, J., Wei, B., Lai, M., Shou, J., Fan, Y., Xu, Y., 2023c. Zero-shot nuclei detection via visual-language pre-trained models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 693–703.

Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Wang, Q., Shen, D., 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097 .

Yan, A., Wang, Y., Zhong, Y., He, Z., Karypis, P., Wang, Z., Dong, C., Gentili, A., Hsu, C.N., Shang, J., et al., 2023. Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 .

Yan, K., Cai, J., Zheng, Y., Harrison, A.P., Jin, D., Tang, Y., Tang, Y., Huang, L., Xiao, J., Lu, L., 2020. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. IEEE Transactions on Medical Imaging 40, 2759–2770.

Yang, X., Xie, J., Li, X., Li, X., Li, X., Shen, L., Deng, Y., 2023. Tceip: Text condition embedded regression network for dental implant position prediction. arXiv preprint arXiv:2306.14406 .

Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R., 2017. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics 22, 1218–1226.

You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B., 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 101–111.

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al., 2023. Evaluating progress in automatic chest x-ray radiology report generation. Patterns 4.

Yu, Y., Zhan, F., Wu, R., Zhang, J., Lu, S., Cui, M., Xie, X., Hua, X.S., Miao, C., 2022. Towards counterfactual image manipulation via clip, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 3637–3645.

Yüksel, A.E., Gültekin, S., Simsar, E., Özdemir, Ş.D., Gündoğar, M., Tokgöz, S.B., Hamamcı, İ.E., 2021. Dental enumeration and multiple treatment detection on panoramic x-rays using deep learning. Scientific reports 11, 12342.

Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L., 2022. Scaling vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12104–12113.

Zhang, B., Wang, Y., Chen, F., 2014. Multilabel image classification via high-order label correlation driven active learning. IEEE Transactions on Image Processing 23, 1430–1441.

Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., et al., 2023a. Huatuogpt, towards taming language model to be a doctor. arXiv preprint arXiv:2305.15075 .

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2021. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1195–1204.

Zhang, K., Yang, Y., Yu, J., Jiang, H., Fan, J., Huang, Q., Han, W., 2023b. Multi-task paired masking with alignment modeling for medical vision-language pre-training. IEEE Transactions on Multimedia .

Zhang, S., Metaxas, D., 2023. On the challenges and perspectives of foundation models for medical image analysis. arXiv preprint arXiv:2306.05705 .

Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al., 2023c. Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 .

Zhang, S., Zhang, J., Xia, Y., 2022a. Transws: Transformer-based weakly supervised histology image segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 367–376.

Zhang, S., Zhang, J., Xie, Y., Xia, Y., 2023d. Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 109–118.

Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y., 2023e. Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications 14, 4542.

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W., 2023f. Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 .

Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D., 2023g. Text-guided foundation model adaptation for pathological image classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 272–282.

Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., 2022b. Contrastive learning of medical visual representations from paired images and text, in: Machine Learning for Healthcare Conference, PMLR. pp. 2–25.

Zhang, Y., Li, X., Chen, H., Yuille, A.L., Liu, Y., Zhou, Z., 2023h. Continual learning for abdominal multi-organ and tumor segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 35–45.

Zhang, Y., Zhang, H., Chen, X., Lee, S.W., Shen, D., 2017. Hybrid high-order functional connectivity networks using resting-state functional mri for mild cognitive impairment diagnosis. Scientific reports 7, 6530.

Zhang, Z., Wang, J., Ye, J., Wu, F., 2022c. Rethinking graph convolutional networks in knowledge graph completion, in: Proceedings of the ACM Web Conference 2022, pp. 798–807.

Zhao, G., Feng, Q., Chen, C., Zhou, Z., Yu, Y., 2021. Diagnose like a radiologist: Hybrid neuro-probabilistic reasoning for attribute-based medical image diagnosis. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 7400–7416.

Zhao, Z., Wang, S., Gu, J., Zhu, Y., Mei, L., Zhuang, Z., Cui, Z., Wang, Q., Shen, D., 2023. Chatcad+: Towards a universal and reliable interactive cad using llms. arXiv preprint arXiv:2305.15964 .

Zheng, F., Cao, J., Yu, W., Chen, Z., Xiao, N., Lu, Y., . Exploring low-resource medical image classification with weakly supervised prompt learning. Available at SSRN 4578827 .

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

Zhou, F., Chen, H., 2023. Cross-modal translation and alignment for survival analysis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21485–21494.

Zhou, H.Y., Chen, X., Zhang, Y., Luo, R., Wang, L., Yu, Y., 2022a. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nature Machine Intelligence 4, 32–40.

Zhou, H.Y., Lian, C., Wang, L., Yu, Y., 2022b. Advancing radiograph representation learning with masked record modeling, in: The Eleventh International Conference on Learning Representations.

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022c. Conditional prompt learning for vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825.

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022d. Learning to prompt for vision-language models. International Journal of Computer Vision 130, 2337–2348.

Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J., 2023. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. arXiv preprint arXiv:2310.18961 .

Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., Shao, L., 2019. Collaborative learning of semi-supervised segmentation and classification for medical images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2079–2088.

Zhu, W., Hessel, J., Awadalla, A., Gadre, S.Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W.Y., Choi, Y., 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. arXiv preprint arXiv:2304.06939 .

Zuo, S., Xiao, Y., Chang, X., Wang, X., 2022. Vision transformers for dense prediction: A survey. Knowledge-Based Systems 253, 109552.