Knowledge-enhanced Visual-Language Pretraining for Computational Pathology

Xiao Zhou[1], Xiaoman Zhang[1,2], Chaoyi Wu[1,2], Ya Zhang[1,2], Weidi Xie[1,2], Yanfeng Wang[1,2]

[1] Shanghai Artificial Intelligence Laboratory
[2] Shanghai Jiao Tong University
https://github.com/MAGIC-AI4Med/KEP
zhouxiao@pjlab.org.cn, {xm99sjtu, wtzxxxwcy02, ya_zhang, weidi, wangyanfeng} @sjtu.edu.cn

Abstract. In this paper, we consider the problem of visual representation learning for computational pathology, by exploiting large-scale image-text pairs gathered from public resources, along with the domain-specific knowledge in pathology. Specifically, we make the following contributions: (i) We curate a pathology knowledge tree that consists of 50,470 informative attributes for 4,718 diseases requiring pathology diagnosis from 32 human tissues. To our knowledge, this is the first comprehensive structured pathology knowledge base; (ii) We develop a knowledge-enhanced visual-language pretraining approach, where we first project pathology-specific knowledge into latent embedding space via a language model, and use it to guide the visual representation learning; (iii) We conduct thorough experiments to validate the effectiveness of our proposed components, demonstrating significant performance improvement on various downstream tasks, including cross-modal retrieval, zero-shot classification on pathology patches, and zero-shot tumor subtyping on whole slide images (WSIs).

Keywords: Pathology knowledge · Visual-language pretraining

1 Introduction

Pathology diagnosis is currently the golden standard for examining various diseases in clinical applications, especially in the diagnosis of neoplasm [43]. In the last decade, the prosperity of deep learning on computer vision has led to the rapid development of computational pathology, for example, approaches with supervised learning [6,22,34,45,46], and weakly supervised learning [5,11,32,38,59]. Despite being promising, these approaches have been fundamentally limited by the scale of costly labels. Alternatively, self-supervised pretraining on numerous unlabeled pathological images [8,9,26,48] has attracted unprecedented attention, yet it still requires supervised fine-tuning for downstream deployments.

In the recent literature [25, 41], studies on the multimodal foundation model have demonstrated conspicuous improvement in downstream zero-shot tasks, by

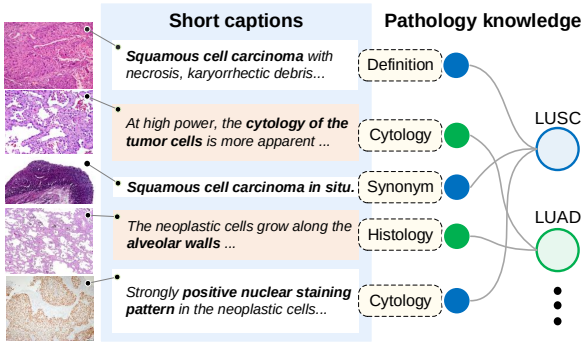

Fig. 1: Knowledge-enhanced pathology image-text alignment. The short caption of a pathology image crawled from public websites is typically unstructured and with varying granularities, which introduces noticeable ambiguities for image-text alignment. While the implicit structures and correlations between different image-caption pairs could be constructed by explicit disease attributes (dashed boxes), which can be well-aligned by a pathology knowledge tree. LUSC and LUAD suggest lung squamous cell carcinoma, and lung adenocarcinoma, respectively.

simply training to align visual and language embedding space on massive image-text pairs crawled from public websites. In contrast to computer vision [19,25,41], representation learning in pathology requires tremendous expertise and domain knowledge. An ideal training corpus would thus consist of well-structured medical reports from the hospital, however, it is often extremely difficult to acquire, due to privacy concerns. Alternatively, recent works [23, 24, 36, 37] propose to gather large-scale image-caption pairs from public resources (PubMed [42] papers, Twitter, Youtube videos). Compared to discrete image labels, image captions from medical reports and academic articles can potentially provide more valuable medical information without manual annotation.

For existing works that adopt simple contrastive learning on pathology image and caption pairs, they suffer from the following challenges: *First*, the short captions from these web-crawled image-text pairs are typically noisy, unstructured, and lack domain knowledge (Fig. 1), which can be sub-optimal for constructing high-quality pathology-specific visual representations; *Second*, the varying granularities of the free text from captions will inevitably introduce ambiguities when aligning with images, thus causing the model to be highly sensitive to the used text prompts during the inference stage [36]. Thus, effectively harnessing the potential of such web-crawled data remains to be a challenge.

In this paper, to tackle the above challenges, we anticipate that introducing pathology knowledge is of great significance to make up for the deficiency of short image captions. To this end, we make the following contributions: (i) We curate a pathology knowledge tree, **PathKT**, by collecting 50,470 informative pathological attributes of 4,718 diseases in 32 tissues from publicly available educational

resources and OncoTree[3] [31]. (ii) We propose a novel knowledge encoder pretraining approach, that projects the attributes of each disease from PathKT into latent embedding space, where the attributes of the same disease, including disease synonyms, definitions, histology, and cytology features, share similar representations. (iii) We develop a knowledge-enhanced pretraining (**KEP**) approach to align pathology visual-language representations, which freezes the knowledge encoder and continuously injects domain-specific knowledge into the image-text embedding space. To demonstrate the effectiveness of our proposed approach, we conduct thorough experiments on three downstream tasks, including retrieval on three pathological image-caption datasets, zero-shot patch classification on eight patch-level pathology image datasets, and zero-shot WSI tumor subtyping on three datasets from The Cancer Genome Atlas (TCGA) [4]. Quantitative experiments suggest that knowledge guidance can significantly enhance the performance across different tasks.

2 Related work

Vision-Language Pretraining (VLP). Current vision-language pretraining approaches are typically categorized into two groups. The first group, referred to as two-stream approaches [41, 51], involves using two separate encoders to extract features for visual and textual data, respectively. The second group, known as single-stream methods [10, 29], utilizes a cross-modal fusion encoder to enhance interactions between vision and text features. In the field of medical VLP, existing methods mainly adopt the two-stream approach [7, 12, 20, 35, 39, 56, 58]. Despite the valuable contributions of these methods, the reliance on data-driven representation learning restricts the utilization of systematic and structured medical knowledge, leading to less than professional diagnosis.

VLP in Computational Pathology. In the recent literature, Huang et al. proposed a pathology VLP model named PLIP [23] and released OpenPath that contains 200K image-caption pairs from the social media of Twitter and other public resources. To extend the pathology data scale, Ikezogwo et al. curate Quilt1m [24] with over one million histopathology image-text pairs by capturing keyframes and speech from YouTube videos. In addition to evaluating the zero-shot classification on patch-level pathology images, Lu et al. propose MI-Zero [37] and CONCH [36] to extend the transfer ability of pathology VLP models on gigapixel whole slide images (WSIs). In contrast to these existing work, that finetunes a CLIP [41] or CoCa [55] on pathology image-caption pairs, we propose to first train a pathology-specific knowledge encoder, and use it to guide visual-language representation learning.

Medical Knowledge-enhanced Learning. In the medical community, leveraging external medical knowledge to enhance deep learning models has become an increasingly important topic [49, 50, 57]. Generally, existing approaches can

[3] `https://oncotree.info/`
[4] `https://portal.gdc.cancer.gov/`

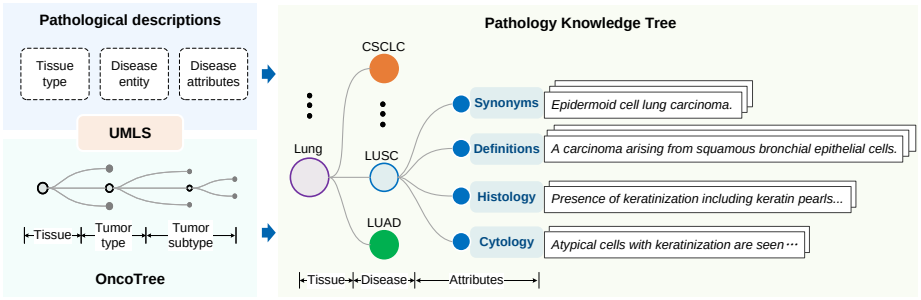

Fig. 2: The construction of pathology knowledge tree. OncoTree is adopted as the base architecture to construct the PathKT. The tissue types, disease entities, and attributes are first extracted from web-crawled pathological descriptions, where cancers are then matched to OncoTree based on their tissue types and tumor types/subtypes using UMLS CUIs. Moreover, non-tumor diseases are added to the knowledge tree according to their tissue types. Finally, the pathology knowledge tree integrates 4718 diseases from 32 tissues, with each disease containing various synonyms, definitions, and histological and cytological features. CSCLC in this figure suggests combined small cell lung cancer.

be categorized based on the ways of using medical knowledge: model-based approaches [14, 21] adopts the prior knowledge of radiology or diagnosis summarized by doctors to design algorithms; and input-based methods [13, 33, 53, 54] directly exploit knowledge as the external input to guide representation training. However, most of these works are focused on the analysis of chest X-rays.

3 Methods

Our primary goal is to leverage structured pathology knowledge to enhance visual-language representation learning. To start with, we construct a **Path**ology **K**nowledge **T**ree, termed as **PathKT**, that consists of 50,470 informative attributes of 4718 diseases from 32 human tissues (Sec. 3.1). We then train a knowledge encoder that projects the structured pathology knowledge into an embedding space (Sec. 3.2). We further employ the knowledge encoder to guide visual-language pretraining for computational pathology, termed as KEP (**K**nowledge-**E**nhanced **P**re-training, Sec. 3.3).

3.1 PathKT Construction

Here, we detail the procedure for building up a pathology knowledge tree with various online sources, which will be further used for knowledge encoding.

Knowledge Source. We collect pathology-specific knowledge from publicly available educational resources, such as text books, professional websites, and structured databases (OncoTree [31]). Specifically, we extract pathological descriptions of 884 tumor subtypes from OncoTree and 4360 diseases from text books and professional websites, including all domain knowledge required for

diagnosis in clinical scenarios, *i.e.,* disease name/ synonyms, definitions, histological features, and cytological features.

Knowledge Tree Construction. We structure these knowledge sources into a knowledge tree by expanding the OncoTree, as shown in Fig. 2. OncoTree is a tree-structure cancer classification system [31], which consists of 884 tumor subtypes from 32 tissues, with each type of tumor linked to a Concept Unique Identifier (CUI) from Unified Medical Language System (UMLS) [2]. Specifically, we first extract tissue types, disease entities, and disease attributes from pathological descriptions of 4360 diseases, in which 168 cancers are found overlapped with OncoTree by using SciSpacy package [40] to link the UMLS CUIs. The rest 4192 diseases are then added to the knowledge tree according to their tissue types. After deduplication and noise reduction (358 diseases without any informative descriptions are deleted), all diseases are organized into the corresponding tissues. The final histopathology knowledge tree contains 4718 diseases from 32 tissues. The attributes of each disease node are constructed by a varying number of synonyms, definitions, and histological and cytological descriptions. The statistics of PathKT is shown in Table S1 and Fig. S1 in Supplementary Materials. In the final PathKT, for instance, the tissue node of the lung connects all lung diseases, including combined small cell lung cancer (CSCLC), lung squamous cell carcinoma (LUSC), and lung adenocarcinoma (LUAD), etc. The LUSC node connects four kinds of attributes, shown in Fig. 2.

3.2 Pathology Knowledge Encoding

In this section, we describe details for projecting tree-structure pathological knowledge into a latent embedding space, by training a knowledge encoder. Specifically, we align different disease entities with their corresponding free-text attributes via metric learning, such that, the synonyms, definitions, and corresponding histological/cytological features are close in the embedding space. As a consequence, the model can link the pathology images to their implicit disease labels during the visual-language pertaining, since the free-form text descriptions in image-text pairs may contain disease attributes, such as pathological features and disease definitions, which have already been aligned with disease names/synonyms. Therefore, the alignment of pathology knowledge can help the model link to the disease entities for better performance during diagnosis.

Problem Setting. Given a set of disease entities with their corresponding attributes, $\mathcal{D} = \{(d_1, \mathbf{a}_1), \ldots, (d_n, \mathbf{a}_n)\}$, where d_i denotes the i-th disease entity, and $\mathbf{a}_i = \{a_i^1, \ldots, a_i^k\}$ refer to the associated k attributes, both disease and attributes are represented in the format of natural language. Note that, for different disease entities, k also varies, our goal here is to train a model that satisfies:

$$\mathrm{sim}(\Phi_{\mathrm{k}}(a_i^p), \Phi_{\mathrm{k}}(a_i^q)) \gg \mathrm{sim}(\Phi_{\mathrm{k}}(a_i^p), \Phi_{\mathrm{k}}(a_j^t)),\ i \neq j, \tag{1}$$

where $\Phi_{\mathrm{k}}(\cdot)$ denotes the knowledge encoder, $\mathrm{sim}(\cdot, \cdot)$ refers to the similarity, a_i^p, a_i^q and a_j^t refer to the randomly sampled attributes from the i, j-th disease entity.

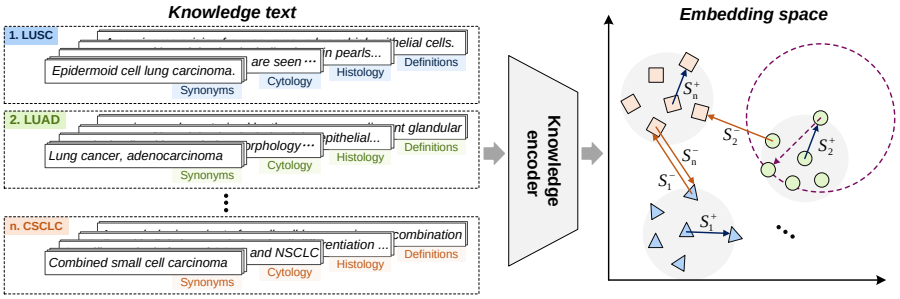

Fig. 3: Knowledge encoder pretraining based on metric learning. n disease entities and each with k attributes, including disease synonyms, definitions, cytology and pathology features, construct a mini-batch (left part of the figure), which are fed to a knowledge encoder for pretraining. In the embedding space (right part of the figure), the markers in different shapes represent the embeddings of attributes of different diseases. S_i^+ suggests the max-min positive attribute similarity within the i-th disease, while S_i^- denotes the maximal attribute similarity between the i-th disease and other diseases. The goal of metric learning is to increase S_i^+ and meanwhile decreasing S_i^-. The purple-dashed arrow and circle denote the minimal positive attribute similarity in the second class and the hypersphere it spans.

Intuitively, the knowledge encoder enables the attributes of the same disease to be pulled together, while attributes from different diseases are pushed apart.

Training. In order to achieve the objective defined in Eq. 1, metric learning is exploited to construct an embedding space where the representations of intra/inter-class instances are clustered/separated. In specific, given a mini-batch that contains n random diseases and each with k attributes, we denote the normalized embedding for the p-th attribute of i-th disease as: $\mathbf{z}_p^i = \Phi_{\mathrm{k}}(a_i^p)/\|\Phi_{\mathrm{k}}(a_i^p)\|$. We adopt the recently proposed AdaSP loss [60], which finds out a max-min positive similarity (Fig. 3) and then shapes a loss with the maximal negative similarity:

$$\mathcal{L}_{\mathrm{metic}} = \frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{(S_i^- - S_i^+)/\tau}\right), \tag{2}$$

where τ is a temperature parameter. S_i^+ and S_i^- denote the max-min positive and the maximal negative similarity, which can be computed by the soft version:

$$S_i^+ = \max_p \min_q \left\langle \mathbf{z}_p^i, \mathbf{z}_q^i \right\rangle \approx \tau \log\left(\sum_{p=1}^{k} \frac{1}{\sum_{q=1}^{k} e^{-\frac{\left\langle \mathbf{z}_p^i, \mathbf{z}_q^i \right\rangle}{\tau}}}\right) \tag{3}$$

$$S_i^- = \max_{j,p,q} \left\langle \mathbf{z}_p^i, \mathbf{z}_q^j \right\rangle \approx \tau \log\left(\sum_{p=1}^{k}\sum_{j=1,j\neq i}^{n}\sum_{q=1}^{k} e^{\frac{\left\langle \mathbf{z}_p^i, \mathbf{z}_q^j \right\rangle}{\tau}}\right) \tag{4}$$

where $\langle\cdot\rangle$ represents the cosine similarity, the details about the soft version can be found in Supplementary Materials.

Discussion. The conventional triplet loss [18] with batch hard mining strategy is widely adopted in metric learning, while it is not suitable in our case due to the following two reasons: *First,* in the pathology knowledge, each disease entity is associated with at most four types of attributes, the text descriptions of different attribute types might reveal significant divergence, which causes a large intra-class variation, marked by the purple dashed circle in Fig. 3, during training. *Second,* the fine-grained diseases might share high similarities, such as tumor subtypes: breast invasive ductal carcinoma and breast invasive lobular carcinoma, which cause low inter-class variations during training. When a mini-batch meets these two conditions, traditional triplet loss that enforces the instance similarities could typically produce the bad local minima of optimization [52], which, therefore, undermines the metric learning.

3.3 Pathology Knowledge Enhanced Pretraining

In this section, we present a simple yet effective pretraining approach, termed KEP, that leverages the established knowledge encoder to guide visual-language pretraining for computational pathology.

Visual-Language Pretraining. Given paired image and captions, denoted as $\mathcal{F} = \{(x_1, c_1), \ldots, (x_n, c_n)\}$, our goal is to construct a visual-language embedding space from paired image-text data, that satisfies:

$$\text{sim}(\Phi_\text{v}(x_i), \Phi_\text{t}(c_i)) \gg \text{sim}(\Phi_\text{v}(x_i), \Phi_\text{t}(c_j)),\ i \neq j, \tag{5}$$

where $\Phi_\text{v}(\cdot)$ and $\Phi_\text{t}(\cdot)$ denote the visual and the textual encoder, respectively. In this work, ViT-B-32/16 is adopted as the backbone for the visual encoder and initialized from the visual weights of CLIP [41] / BiomedCLIP [56]. A projection head is added on top of the visual encoder to bridge the gap caused by the non-pathology initialization, shown in Fig. 4. The text encoder is initialized by the pretrained knowledge encoder, as described in Sec. 3.2. In order to learn effective visual-language representation, we optimise an infoNCE contrastive objective:

$$\mathcal{L}_\text{vt} = -(\log \frac{e^{\mathbf{v}_i^T \mathbf{t}_i/\tau}}{\sum_{j=1}^{n} e^{\mathbf{v}_i^T \mathbf{t}_j/\tau}} + \log \frac{e^{\mathbf{t}_i^T \mathbf{v}_i/\tau}}{\sum_{j=1}^{n} e^{\mathbf{t}_i^T \mathbf{v}_j/\tau}}), \tag{6}$$

where $\mathbf{v}_i = \Phi_\text{v}(x_i)$, and $\mathbf{t}_i = \Phi_\text{t}(c_i)$, refer to the normalized embedding vectors from visual and text encoders, respectively. τ denotes a temperature parameter.

Knowledge Distillation. To keep the alignment between images and free-form captions inside the knowledge space and thus the images can be linked to their implicit disease entities (right part in Fig. 4), we adopt an additional frozen branch to continuously distill pathology knowledge to the text encoder. Specifically, we use the text-knowledge embedding pairs to construct a contrastive loss item $\mathcal{L}_\text{tk}$, which can be computed by Eq. 6 with the normalized visual embedding vectors replaced by the knowledge embedding vectors:

$$\mathcal{L}_\text{tk} = -(\log \frac{e^{\mathbf{k}_i^T \mathbf{t}_i/\tau}}{\sum_{j=1}^{n} e^{\mathbf{k}_i^T \mathbf{t}_j/\tau}} + \log \frac{e^{\mathbf{t}_i^T \mathbf{k}_i/\tau}}{\sum_{j=1}^{n} e^{\mathbf{t}_i^T \mathbf{k}_j/\tau}}), \tag{7}$$

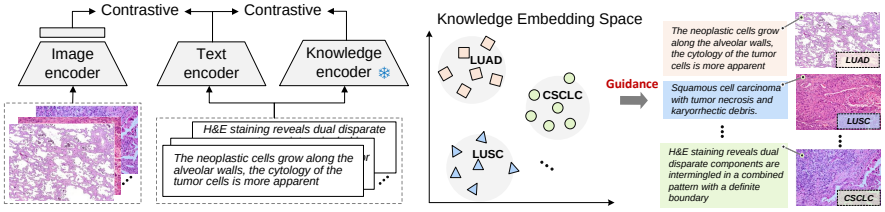

Fig. 4: Model architecture (left graph). A projection head is added on the top of the visual encoder to bridge the gap between the image and the text encoder. The knowledge encoder is frozen across the whole training stage to distill pathology knowledge to the learnable text encoder. As a result, the pathology images can be aligned with their implicit disease labels (marked by dashed boxes in the right graph) during visual-language pretraining, since the captions contain disease attributes that have been already aligned with disease names/synonyms in the knowledge embedding space.

where $\mathbf{k}_i = \Phi_{\mathrm{k}}(c_i)$, refer to the normalized embedding vectors from the knowledge encoder. The overall training loss can thus be computed as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{vt}} + \alpha \mathcal{L}_{\mathrm{tk}}, \tag{8}$$

where α denotes a weight parameter. It is worth emphasizing that the key contribution in KEP is to initialize the text encoder with the pre-obtained pathology knowledge encoder and adopt it to continuously distill pathology knowledge to the text encoder, while for the other parts, we keep the same as PLIP [23].

4 Experiments

In this section, we first introduce the datasets used for training and evaluation in this paper, followed by the evaluation metrics and implementation details.

4.1 Training Datasets

Dataset for Knowledge Encoding. We carry out pathology knowledge encoding with PathKT, where 4,718 disease nodes with a total number of 50,470 attributes are extracted, shown in Table S1 in Supplementary Materials.

Datasets for KEP Pretraining. Our dataset is composed of two parts. *First*, we collect the **OpenPath** data provided by PLIP [23] and obtain 138,874 image-text pairs after denoising and pre-processing. *Second*, we collect the **Quilt1M** [24] dataset, which gathers pathology image-text pairs from four public sources: PubMed articles, LAION [44], OpenPath [23], and Youtube videos. Since one of the downstream evaluation datasets Arch-PubMed [15] contains pathology images from PubMed articles, we remove this data source from Quilt1M to avoid data leaking. Considering that many images in Quilt1M are matched with multiple captions, we concatenate the captions related to the same image and finally obtain 576,608 image-text pairs for the Quilt1M dataset.

4.2 Downstream Tasks

We evaluate the pretrained models on three tasks, namely, retrieval, zero-shot patch classification, and zero-shot WSI tumor subtyping. The details for all evaluation datasets are exhibited in Table S2 in Supplementary Materials.

Retrieval. This task involves cross-modal retrieval and disease retrieval. Cross-modal retrieval aims to retrieve the correct caption for a given image and vice versa. Disease retrieval, on the other hand, utilizes the disease names to retrieve captions or images with the same disease label, which is proposed to demonstrate the effectiveness of the knowledge encoder. For cross-modal retrieval, we follow PLIP [23] to split the ARCH [15] dataset into Arch-PubMed and Arch-book. In addition, we also gather image-text pairs from publicly available educational resources and curate a retrieval dataset, termed as **PathPair**, consisting of 9,358 pathology image-caption pairs with known 1676 disease labels. For disease retrieval, we utilize the captions, images, and their disease labels in PathPair.

Zero-shot Patch Classification. This task involves zero-shot classification on patch-level pathology images. Specifically, at inference time, we randomly select one template from the 21 templates in CONCH [36] and one type synonym from the corresponding name list (exhibited in Supplementary Materials) to yield a text prompt for each type, e.g. a histopathological image of CLASSNAME. (template) + beast invasive carcinoma (synonym) $\rightarrow$ a histopathological image of beast invasive carcinoma. This process is repeated 100 times in every experiment. Following PLIP [23] and Quilt1M [24], we adopt these patch-level pathology image datasets: BACH [1], NCT-CRC-HE-100K [27], KatherColon [28], LC25000 [3], RenalCell [4], SICAP [47], SkinCancer [30], WSSS4LUAD [17]. Each dataset includes multiple types of H&E stained cell micrographs.

Zero-shot WSI Tumor Subtyping. This task involves zero-shot tumor subtyping on pathology whole slide images of common and rare cancers. We follow MI-Zero [37] and CONCH [36] for evaluation. Specifically, we first divide WSI into 256×256 patches and then predict the class label of each image patch in a zero-shot manner. The tumor type of the whole slide is then provided by integrating Top-K predictions on patches. For this task, we also employ the templates and tumor synonyms from CONCH [36] to randomly generate 100 text prompts (listed in Supplementary materials) for each tumor subtype. (i) **For common cancers**, we follow existing research [36,37] and collect 525 WSIs with three tumor types, including 150 breast carcinoma (BRCA), 150 non-small cell lung cancer (NSCLC) histopathology slides, and 225 renal cell carcinoma (RCC) slides, from TCGA. Specifically, BRCA consists of two subtypes: invasive ductal carcinoma (IDC) and invasive lobular carcinoma (IDC). NSCLC contains lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). RCC is divided into chromophobe renal cell carcinoma (CHRCC), clear-cell renal cell carcinoma (CCRCC), and papillary renal cell carcinoma (PRCC). Each tumor subtype has 75 WSIs. (ii) **For rare cancers**, we collected 41 WSIs of rare subtypes for breast cancer (6 for Intraductal papillary adenocarcinoma with invasion, 6 for Medullary carcinoma, NOS, 13 for Metaplastic carcinoma, NOS and

16 for Mucinous adenocarcinoma), from TCGA and merge them with common breast cancers (75 for IDC and 75 for ILC).

4.3 Evaluation Metrics.

Retrieval. To evaluate the retrieval performance, we adopt the Recall@K metric, suggesting the ratio of correctly retrieved queries in Top-K retrieved samples.

Zero-shot Patch Classification. For zero-shot classification tasks, we adopt the same metric as PLIP [23], namely, weighted F1 (wF1). We report the median, the first, and the third quartile (Q1, Q3) of wF1 across all text 100 prompts.

Zero-shot WSI Tumor Subtyping. We adopt the same metric as MI-Zero [37], namely, balanced accuracy at Top-K pooling to measure the zero-shot tumor subtyping performance on WSIs. We also report the median, the first, and the third quartile (Q1, Q3) of the balanced accuracy across all 100 text prompts.

4.4 Implementation Details

Knowledge Encoder Pretraining. We adopt the architecture of PubMedBERT [16] to encode knowledge. The embedding dimension is set to 512. The temperature parameter τ is set to 0.04 in Eq. 2. The batch size is set to 256, including 32 disease entities with 8 instances per entity.

Pathology Image-text Pretraining. To achieve a fair comparison with PLIP and Quilt1M, we utilize the same visual encoder (ViT-B-32) and initialization of the visual encoder (CLIP [41]), termed by KEP-32, and set the input image size to 224×224. Additionally, we also develop a KEP variant named KEP-16 that is initialized by the visual weights of BiomedCLIP [56] (ViT-B-16). The batch size and learning rate are set to 256 and 1e-5, respectively. The temperature in Eq. 6 and Eq. 7 is set to 0.04 across all experiments. For OpenPath, we conduct the pathology VLP for 30 epochs, While for Quilt1M, 15 epochs are adopted.

5 Results

In this section, we show the results on the downstream tasks, to evaluate the effectiveness of proposed knowledge-enhanced representation learning. Note that for a fair comparison with the publicly available model released by the original authors, in all experiments, we separately train KEP on the OpenPath [23] dataset, and Quilt1M [24] dataset, and then report results on downstream tasks.

5.1 Retrieval

In this section, we evaluate the performance of retrieval, including cross-modal retrieval on three datasets and disease retrieval on PathPair.

Cross-modal Retrieval. In Table 1, we demonstrate the results for different models pretrained on OpenPath and Quilt1M. It can be seen that KEP-32

Table 1: Performance comparison with PLIP and QuiltNet on three retrieval datasets. All models are pre-trained on the OpenPath and Quilt1M datasets, respectively. i2t and t2i denote image-to-text and text-to-image retrieval, respectively. Bold fonts suggest the best performance.

Training dataset	Task	Model	Knowledge Enhancement	Arch-PubMed		Arch-book		PathPair	
				Recall@10	Recall@50	Recall@10	Recall@50	Recall@10	Recall@50
OpenPath	i2t	PLIP	✗	0.067	0.185	0.152	0.393	0.038	0.119
		KEP-32	✓	0.098	0.283	0.164	0.404	0.071	0.205
		KEP-16	✓	**0.163**	**0.398**	**0.244**	**0.537**	**0.108**	**0.275**
	t2i	PLIP	✗	0.067	0.181	0.165	0.419	0.047	0.133
		KEP-32	✓	0.085	0.226	0.148	0.365	0.061	0.171
		KEP-16	✓	**0.138**	**0.339**	**0.238**	**0.533**	**0.093**	**0.247**
Quilt1M	i2t	QuiltNet	✗	0.139	0.326	0.188	0.407	0.065	0.166
		KEP-32	✓	0.140	0.327	0.240	0.521	0.084	0.221
		KEP-16	✓	**0.196**	**0.421**	**0.282**	**0.564**	**0.108**	**0.254**
	t2i	QuiltNet	✗	0.122	0.293	0.204	0.429	0.071	0.195
		KEP-32	✓	0.135	0.326	0.275	0.568	0.106	0.276
		KEP-16	✓	**0.176**	**0.404**	**0.340**	**0.621**	**0.136**	**0.326**

Table 2: Performance comparison of disease retrieval. l2t and i2l denote label-to-text and image-to-label, respectively. Bold fonts suggest the best performance.

Task	Metrics	OpenPath			Quilt1M		
		PLIP	KEP-32	KEP-16	QuiltNet	KEP-32	KEP-16
l2t	Recall@10	0.408	0.693	**0.699**	0.211	0.648	0.643
	Recall@50	0.536	0.832	**0.835**	0.325	0.812	0.808
i2l	Recall@10	0.114	0.162	0.223	0.135	0.226	**0.259**
	Recall@50	0.292	0.381	0.478	0.322	0.473	**0.519**

pretrained on OpenPath outperforms PLIP on all datasets with respect to the image-to-text retrieval task, especially on Arch-PubMed and PathPair with more than 3% boost on R10. As for the text-to-image task, KEP-32 also achieves better performance on the dataset of Arch-PubMed and PathPair. Furthermore, our KEP-16 model improves the retrieval performance by a large margin on all datasets across both retrieval tasks. Similar results can be concluded that KEP-32 pretrained on Quilt1M improves the performance on all datasets for both retrieval tasks, which demonstrates the effectiveness of knowledge guidance for visual-language pretraining on cross-modal retrieval.

Disease Retrieval. Table 2 shows the performance of disease retrieval on the PathPair dataset for different models. It can be seen that although the scale of Quilt1M is 5 times OpenPath, models pretrained on OpenPath often outperforms that on Quilt1M for the label-to-text task. We conjecture this may be caused by the quality of captions in Quilt1M. Our approach KEP-32 outperforms PLIP and Quilt1M by a large margin on both label-to-text and image-to-label tasks, suggesting that the knowledge encoder contributes to paying attention to the key disease information when encoding captions, and thus improves the alignment between images and their disease labels.

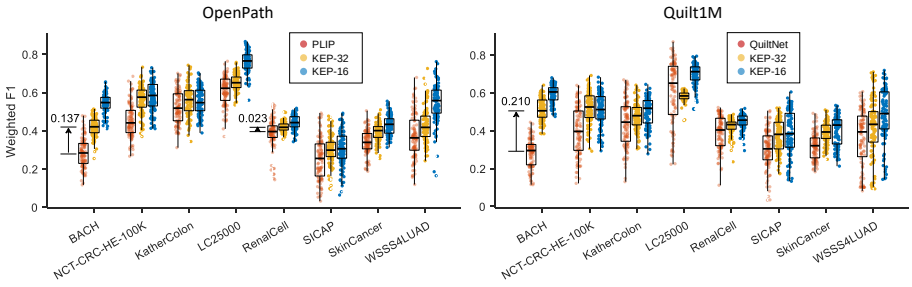

Fig. 5: The comparison of zero-shot patch classification between different models. The left and the right subfigures suggest pretraining on OpenPath and Quilt1M, respectively. The visual encoders of KEP-32 and KEP-16 are initialized by CLIP (ViT-B-32) and BiomedCLIP (Vit-B-16), respectively. The number of points for every box is 100, with each representing the performance of one text prompt. The upper, center, and lower line of each box denote the first, median, and third quartile of the distribution.

5.2 Zero-shot Patch Classification

In this section, we evaluate the performance of zero-shot classification on patch-level pathology images from 8 datasets. We report the performance distribution (Fig. 5), where each point in the figure denotes the performance of one prompt.

Comparison to PLIP. In Fig. 5 (left), we demonstrate the performance comparison with PLIP [23], by only pretraining on OpenPath image-text pairs. Compared with PLIP, KEP-32 achieves better zero-shot classification performance on all datasets, with a maximum enhancement of 0.137 on the BACH dataset. Moreover, with the medical-specific initialization of the visual encoder (BiomedCLIP), KEP-16 can further improve the performance significantly in most datasets. Note that, the boxes of KEP variants are generally shorter than that of PLIP, suggesting that our approach is less sensitive to the varying text prompts than PLIP, which can be demonstrated by the visualization of prompt embeddings, shown in Supplementary Materials.

Comparison to QuiltNet. As shown in Fig. 5 (right), we exhibit the performance comparison between KEP and QuiltNet [24] pretrained on Quilt1M dataset. Similar to results on OpenPath, KEP-32 outperforms Quilt1M in seven out of eight datasets, especially for the BACH, KatherColon, and SkinCancer datasets. KEP-16 can further improve the weighted F1 score. Most boxes of KEP variants are also shorter than that of Quilt1M. The visualization of prompt embeddings, shown in Supplementary Materials, maintains that pathology knowledge can reduce the ambiguities of image-text alignment and thus improve the zero-shot classification performance.

5.3 Zero-shot WSI Tumor Subtyping

In this section, we evaluate the transfer ability of different models for tumor subtyping on common and rare cancers.

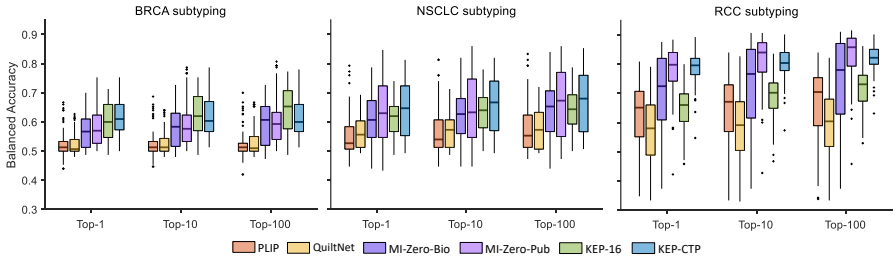

Fig. 6: The performance comparison of tumor subtyping on TCGA-BRCA (common), TCGA-NSCLC, and TCGA-RCC WSIs. The upper, center and lower line of each box denote the first, median, and third quartile of the performance distribution, respectively. The scattered points represent outliers. KEP-16 and KEP-CTP are trained on OpenPath with the visual encoder initialized by BiomedCLIP [56] and CTransPath [48], respectively. MI-Zero-Bio and MI-Zero-Pub are two variants of MI-Zero with different initialization of text encoder. Their visual encoders are initialized by CTransPath. In addition, MI-Zero adopts in-house pathology reports to pretrain their text encoders.

Table 3: Performance comparison of tumor subtyping on six BRCA subtypes, including 2 common and 4 rare cancers. Bold fonts suggest the best performance.

Model	Training dataset	Top-1		Top-10	
		Common	Rare	Common	Rare
PLIP	OpenPath	0.113	0.115	0.119	0.119
QuiltNet	Quilt1M	0.278	0.094	0.279	0.102
KEP-16	OpenPath	**0.343**	**0.217**	**0.368**	**0.245**
MI-Zero-Bio	In-house & Web data	0.390	0.214	0.403	0.237
MI-Zero-Pub	In-house & Web data	0.331	**0.282**	0.347	**0.304**
KEP-CTP	OpenPath	**0.443**	**0.282**	**0.432**	0.301

Common cancers. Fig. 6 shows the performance comparison of tumor subtyping on common cancers, including TCGA-BRCA (common), TCGA-NSCLC and TCGA-RCC WSIs. MI-Zero-Bio and MI-Zero-Pub are two variants of MI-Zero with different text encoders. The visual encoders of the two MI-Zero variants are initialized by CTransPath [48]. To achieve a fair comparison, we also train a KEP variant named KEP-CTP with the visual encoder initialized by CTransPath. Both KEP-16 and KEP-CTP are trained on the OpenPath dataset. It can be seen that our approach KEP-16 and KEP-CTP outperform PLIP and QuiltNet on all datasets. Note that, the comparisons between variants of MI-Zero and KEP-CTP are actually not fair for our method, as the MI-Zero variants have also been pretrained on massive in-house data – over 550k pathology reports from hospitals [37]. Yet, KEP-CTP still outperforms MI-Zero-Bio across all datasets and achieves comparable performance with MI-Zero-Pub on NSCLC and RCC.

Rare cancers. Table 3 exhibits the median balanced accuracy of tumor subtyping on common and rare breast cancers. It can be seen that our method KEP-16 pretrained on OpenPath outperforms PLIP and QuiltNet by a large margin on both common and rare breast cancers. Moreover, KEP-CTP significantly outper-

Table 4: Experimental results of ablation study on model architecture and initialization. Arch-PubMed and PathPair suggest the text-to-image retrieval and the disease retrieval task, respectively. PMB and KB denote PubMedBERT and our pretrained pathology knowledge encoder, respectively. BCLIP denotes BiomedCLIP. Bold fonts and underline suggest the best and the second-best performance, respectively.

Visual Init.	Text Init.	Projection head	Knowledge distill	Metric loss	KatherColon Median (Q1, Q3)	Arch-PubMed R10	Arch-PubMed R50	PathPair R1	PathPair R5
Scratch	PMB				0.322 (0.272, 0.392)	0.037	0.121	0.056	0.113
CLIP	PMB				0.407 (0.358, 0.452)	0.081	0.214	0.161	0.316
CLIP	PMB	✓			0.486 (0.448, 0.517)	0.073	0.210	0.161	0.305
BCLIP	PMB	✓			0.553 (0.503, 0.591)	<u>0.127</u>	**0.347**	0.187	0.350
CLIP	PMB	✓	✓(PMB)		0.484 (0.435, 0.520)	0.063	0.182	0.034	0.089
CLIP	KB	✓		Adasp	0.530 (0.477, 0.600)	0.086	0.232	0.270	0.482
CLIP	KB	✓	✓(KB)	Adasp	<u>0.563</u> (<u>0.505</u>, <u>0.610</u>)	0.085	0.226	<u>0.409</u>	**0.618**
CLIP	KB	✓	✓(KB)	Triplet	0.531 (0.482, 0.580)	0.087	0.227	0.265	0.493
BCLIP	KB	✓	✓(KB)	Adasp	**0.580** (**0.517**, **0.631**)	**0.138**	<u>0.339</u>	**0.424**	**0.618**

forms MI-Zero variants on common breast cancers while achieving comparable performance on the rare ones.

5.4 Ablation Study

In this section, we explore the impact of model architecture, initialization, and other hyperparameters, as shown in Table 4 and Table S3 in Supplementary Materials. Comparing the performance between different models, we can draw the following observations: (i) visual projection head can facilitate the zero-shot classification performance; (ii) our pretrained knowledge encoder can enhance the performance across all tasks; (iii) the frozen knowledge branch can further improve the performance of zero-shot classification and disease retrieval tasks while the frozen PubmedBERT can not; (iv) a better initialization of visual encoder contributes to significant performance improvement on all tasks. (v) the Adasp loss is better than the Triplet loss for the knowledge encoding.

6 Conclusion

In this paper, we address the problem of knowledge-enhanced visual-language pretraining on computational pathology. We first curate a pathology knowledge tree that integrates the informative attributes of diseases requiring pathological diagnosis. We then propose a novel knowledge encoding approach based on metric learning to model structured pathological knowledge. With the guidance of the pretrained knowledge encoder, we conduct extensive visual-language pretraining on pathology image-caption pairs. To demonstrate the effectiveness of our approach, we evaluate pretrained VLP models on three downstream tasks, including retrieval, zero-shot classification on patch-level pathology images, and zero-shot tumor subtyping on pathology WSIs. Quantitative experimental results demonstrate that pathology knowledge can significantly improve the performance across different tasks.

Acknowledgements

This work is supported by the National Key R&D Program of China (No. 2022ZD0160702) and China Postdoctoral Science Foundation (Certificate Number: 2023M741850).

References

1. Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al.: Bach: Grand challenge on breast cancer histology images. Medical Image Analysis **56**, 122–139 (2019)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research **32**(suppl_1), D267–D270 (2004)
3. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142 (2019)
4. Brummer, O., Pölönen, P., Mustjoki, S., Brück, O.: Integrative analysis of histological textures and lymphocyte infiltration in renal cell carcinoma using deep learning. bioRxiv pp. 2022–08 (2022)
5. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019)
6. Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15661–15670 (2023)
7. Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., Golland, P.: Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–539. Springer (2020)
8. Chen, C., Lu, M.Y., Williamson, D.F., Chen, T.Y., Schaumberg, A.J., Mahmood, F.: Fast and scalable search of whole-slide images via self-supervised deep learning. Nature Biomedical Engineering **6**(12), 1420–1434 (2022)
9. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022)
10. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations (2019)
11. Chen, Y.C., Lu, C.S.: Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23936–23945 (2023)
12. Chen, Z., Li, G., Wan, X.: 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 (2022)
13. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1439–1449 (2020)

14. Cui, H., Xu, Y., Li, W., Wang, L., Duh, H.: Collaborative learning of cross-channel clinical attention for radiotherapy-related esophageal fistula prediction from ct. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 212–220. Springer (2020)
15. Gamper, J., Rajpoot, N.: Multiple instance captioning: Learning representations from histopathology textbooks and articles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16549–16559 (2021)
16. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
17. Han, C., Pan, X., Yan, L., Lin, H., Li, B., Yao, S., Lv, S., Shi, Z., Mai, J., Lin, J., et al.: Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. arXiv preprint arXiv:2204.06455 (2022)
18. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
19. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: European Conference on Computer Vision. pp. 533–549. Springer (2022)
20. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: 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 (2021), official Implementation: `https://github.com/marshuang80/gloria`
21. Huang, X., Fang, Y., Lu, M., Yan, F., Yang, J., Xu, Y.: Dual-ray net: automatic diagnosis of thoracic diseases using frontal and lateral chest x-rays. Journal of Medical Imaging and Health Informatics **10**(2), 348–355 (2020)
22. Huang, Y., Zhao, W., Wang, S., Fu, Y., Jiang, Y., Yu, L.: Conslide: Asynchronous hierarchical interaction transformer with breakup-reorganize rehearsal for continual whole slide image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21349–21360 (2023)
23. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
24. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in Neural Information Processing Systems **36** (2024)
25. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
26. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3344–3354 (2023)
27. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo10 **5281** (2018)
28. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Medicine **16**(1), e1002730 (2019)

29. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
30. Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Janssen, C., Meliss, R.R., Muley, T., Sack, U., Steinbuss, G., Kriegsmann, M.: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. Frontiers in Oncology **12**, 1022967 (2022)
31. Kundra, R., Zhang, H., Sheridan, R., Sirintrapun, S.J., Wang, A., Ochoa, A., Wilson, M., Gross, B., Sun, Y., Madupuri, R., et al.: Oncotree: a cancer classification system for precision oncology. JCO clinical cancer informatics **5**, 221–230 (2021)
32. Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., Yang, L.: Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7454–7463 (2023)
33. Li, M., Cai, W., Verspoor, K., Pan, S., Liang, X., Chang, X.: Cross-modal clinical graph transformer for ophthalmic report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20656–20665 (2022)
34. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023)
35. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2023)
36. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Zhang, A., Le, L.P., et al.: Towards a visual-language foundation model for computational pathology. arXiv preprint arXiv:2307.12914 (2023)
37. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19764–19775 (2023)
38. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
39. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Joint learning of localized representations from medical images and reports. In: European Conference on Computer Vision. pp. 685–701. Springer (2022)
40. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: Fast and robust models for biomedical natural language processing. In: Demner-Fushman, D., Cohen, K.B., Ananiadou, S., Tsujii, J. (eds.) Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 319–327. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-5034, https://aclanthology.org/W19-5034
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
42. Roberts, R.J.: Pubmed central: The genbank of the published literature (2001)
43. Rorke, L.B.: Pathologic diagnosis as the gold standard (1997)

44. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
45. Shaban, M., Awan, R., Fraz, M.M., Azam, A., Tsang, Y.W., Snead, D., Rajpoot, N.M.: Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE Transactions on Medical Imaging **39**(7), 2395–2405 (2020)
46. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
47. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer Methods and Programs in Biomedicine **195**, 105637 (2020)
48. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis **81**, 102559 (2022)
49. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training. medRxiv pp. 2023–01 (2023)
50. Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S.: A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis **69**, 101985 (2021)
51. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6787–6800 (2021)
52. Xuan, H., Stylianou, A., Liu, X., Pless, R.: Hard negative examples are hard, but useful. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 126–142. Springer (2020)
53. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical Image Analysis **80**, 102510 (2022)
54. Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H.: Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3208–3216 (2021)
55. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
56. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
57. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
58. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare (2022), highest Starred Implementation: `https://github.com/edreisMD/ConVIRT-pytorch`
59. Zhou, X., Cheng, Z., Gu, M., Chang, F.: Lirnet: Local integral regression network for both strongly and weakly supervised nuclei detection. In: 2020 IEEE Interna-

tional Conference on Bioinformatics and Biomedicine (BIBM). pp. 945–951. IEEE (2020)
60. Zhou, X., Zhong, Y., Cheng, Z., Liang, F., Ma, L.: Adaptive sparse pairwise loss for object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19691–19701 (2023)