Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images

Ming Y. Lu[†,1,2,3], Bowen Chen[†,2,3], Andrew Zhang[1,2,3], Drew F.K. Williamson[2,3],
Richard J. Chen[2,3], Tong Ding[2,3], Long Phi Le[2,3], Yung-Sung Chuang[1], Faisal Mahmood[2,3]
[1]Massachusetts Institute of Technology [2]Harvard University [3]Mass General Brigham
mingylu@mit.edu, bchen18@bwh.harvard.edu, faisalmahmood@bwh.harvard.edu

Abstract

Contrastive visual language pretraining has emerged as a powerful method for either training new language-aware image encoders or augmenting existing pretrained models with zero-shot visual recognition capabilities. However, existing works typically train on large datasets of image-text pairs and have been designed to perform downstream tasks involving only small to medium sized-images, neither of which are applicable to the emerging field of computational pathology where there are limited publicly available paired image-text datasets and each image can span up to 100,000 $\times$ 100,000 pixels. In this paper we present MI-Zero, a simple and intuitive framework for unleashing the zero-shot transfer capabilities of contrastively aligned image and text models on gigapixel histopathology whole slide images, enabling multiple downstream diagnostic tasks to be carried out by pretrained encoders without requiring any additional labels. MI-Zero reformulates zero-shot transfer under the framework of multiple instance learning to overcome the computational challenge of inference on extremely large images. We used over 550k pathology reports and other available in-domain text corpora to pretrain our text encoder. By effectively leveraging strong pretrained encoders, our best model pretrained on over 33k histopathology image-caption pairs achieves an average median zero-shot accuracy of 70.2% across three different real-world cancer subtyping tasks. Our code is available at: https://github.com/mahmoodlab/MI-Zero.

1. Introduction

Weakly-supervised deep learning for computational pathology (CPATH) has rapidly become a standard approach for modelling whole slide image (WSI) data [9, 30, 47, 71, 73]. To obtain "clinical grade" machine learning performance on par with human experts for a given clinical task, many approaches adopt the following model development life cycle: 1) curate a large patient cohort ($N > 1000$ samples) with diagnostic whole-slide images and clinical labels, 2) unravel and tokenize the WSI into a sequence of patch features, 3) use labels to train a slide classifier that learns to aggregate the patch features for making a prediction, and 4) transfer the slide classifier for downstream clinical deployment [9, 43, 91].

Successful examples of task-specific model development (*e.g.* training models from scratch for each task) include prostate cancer grading and lymph node metastasis detection [5, 7–9, 50, 70]. However, this paradigm is intractable if one wishes to scale across the hundreds of tumor types across the dozens of different organ sites in the WHO classification system[1], with most tumor types under-represented in public datasets or having inadequate samples for model development [41, 92]. To partially address these limitations, self-supervised learning has been explored for learning the patch representations within the WSI with the idea that certain local features, such as tumor cells, lymphocytes, and stroma, may be conserved and transferred across tissue types [10, 16, 39, 40, 44, 64, 77]. Though morphological features at the patch-level are captured in a task-agnostic fashion, developing the slide classifier still requires supervision, which may not be possible for disease types with small sample sizes. To scale slide classification across the vast number of clinical tasks and possible findings in CPATH, an important shift needs to be made from task-specific to task-agnostic model development.

Recent works [33, 55] have demonstrated that large-scale pretraining using massive, web-sourced datasets of noisy image-text pairs can not only learn well-aligned representation spaces between image and language, but also transfer the aligned latent space to perform downstream tasks such as image classification. Specifically for CLIP [55], after pretraining a vision encoder in a task-agnostic fashion, the vision encoder can be "prompted" with text from the label

[†]These authors contributed equally to this work.

[1]tumourclassification.iarc.who.int/

space (referred to as "zero-shot transfer", as no labeled examples are used in the transfer protocol). Despite the volume of zero-shot transfer applications developed for natural images [21, 33, 45, 49, 57, 80, 82] and certain medical imaging modalities (*e.g.* radiology [29, 60, 68, 78, 90]), zero-shot transfer for pathology has not yet been studied[2]. We believe this is due to 1) the lack of large-scale, publicly available datasets of paired images and captions in the highly specialized field of pathology, and 2) fundamental computational challenges associated with WSIs, as images can span up to $100,000 \times 100,000$ pixels and do not routinely come with textual descriptions, bounding box annotations or even region of interest labels.

In this work, we overcome the above data and computational challenges and develop the first zero-shot transfer framework for the classification of histopathology whole slide images. On the data end, we curated the largest known dataset of web-sourced image-caption pairs specifically for pathology. We propose "MI-Zero", a simple and intuitive multiple instance learning-based [3, 30] method for utilizing the zero-shot transfer capability of pretrained visual-language encoders for gigapixel-sized WSIs that are routinely examined during clinical practice. We validate our approach on 3 different real-world cancer subtyping tasks, and perform multiple ablation experiments that explore image pretraining, text pretraining, pooling strategies, and sample size choices for enabling zero-shot transfer in MI-Zero.

2. Related Work

Contrastive visual representation learning. Contrastive learning has emerged as a powerful pretraining technique for learning task-agnostic representations of data. It works by constructing collections of similar samples (positive pairs) that would have embeddings maximally aligned in the model's latent space, as well as dissimilar samples (negative pairs), for which embeddings should be spread far apart [14, 24, 51, 93]. Examples from computer vision include different augmented views of the same unlabeled image [14, 25], images with the same class label [20, 37, 83], and different sensory views of the same scene [67]. In computational pathology, recent works [4, 16, 39, 64, 74, 77] have leveraged contrastive learning and unlabeled images from histopathology datasets to pretrain domain-specific image encoders that achieve strong performance on downstream visual recognition tasks compared to transfer learning.

Visual language pretraining. Contrastive learning has also been shown to be a highly effective and scalable strategy to pretrain dual-encoder image-text models that can excel at a range of downstream visual recognition tasks. In medical imaging, ConVIRT [90] considered paired chest X-ray images and reports for learning aligned visual language representation. Later, representative works such as CLIP [56] and ALIGN [33] showed that by scaling to large, diverse web-source datasets of paired images and captions, we can train models capable of exhibiting fairly robust zero-shot transfer capabilities through the use of prompts that exploit the cross-modal alignment between image and text learned by the model during pretraining. Recent methods have explored ways to improve the sample efficiency and zero-shot performance of CLIP-like models [21, 29, 45, 49, 80, 83, 86]. Other works have also explored incorporating a generative loss and masked language modeling loss into the pretraining objective either in addition to or in lieu of contrastive-based objectives [17, 46, 53, 59, 75, 79, 84]. Notably, VirTex [17] has been used to learn visual representations for histopathology images using a generative captioning loss and the ARCH dataset [22] containing 7,562[3] histopathology image-caption pairs from pathology textbooks and PubMed research articles.

Multiple instance learning. Multiple instance learning (MIL) [3] refers to a family of methods that considers learning from weakly-annotated data where each input is a bag or collection of instances such that only an unknown subset of instances are relevant to or representative of the label. In CPATH, algorithms based around the framework of MIL have been proposed for various diagnostic tasks in a weakly-supervised setting, utilizing trainable aggregation operators to learn WSI-level embeddings independent of the size of the bag [9, 30]. ABMIL [30] proposed to use attention-guided weighted averaging as a generic operator to aggregate instance-level embeddings. CLAM [48] took a first step towards data-efficient weakly-supervised learning for WSIs by embedding instances using a frozen ResNet encoder pretrained on ImageNet. Other variants and extensions of the vanilla MIL and ABMIL formulations have been studied [28, 32, 76, 81, 85, 87], including extensions with self-supervised learning [10, 43, 65, 91], multi-scale feature aggregation [10, 43, 69, 88], graphs [11, 18, 42, 52, 91], Transformer attention [10, 36, 61], learning to zoom [6, 38, 66], and multi-modal fusion [12, 13].

3. Methods

3.1. Image caption dataset.

For this work, we curated a histopathology dataset of image-caption pairs by scraping from publicly available educational resources and incorporating the existing ARCH

[2]Concurrent to our work, BiomedCLIP [89] was developed using figure-caption pairs mined from PubMed articles. It was benchmarked on both patch-level histopathology datasets and radiology datasets, but did not study zero-shot transfer for gigapixel WSIs.

[3]The number differs from [22] due to removing a few empty (" "), invalid (*e.g.* "`(continued)`") or unpaired captions.

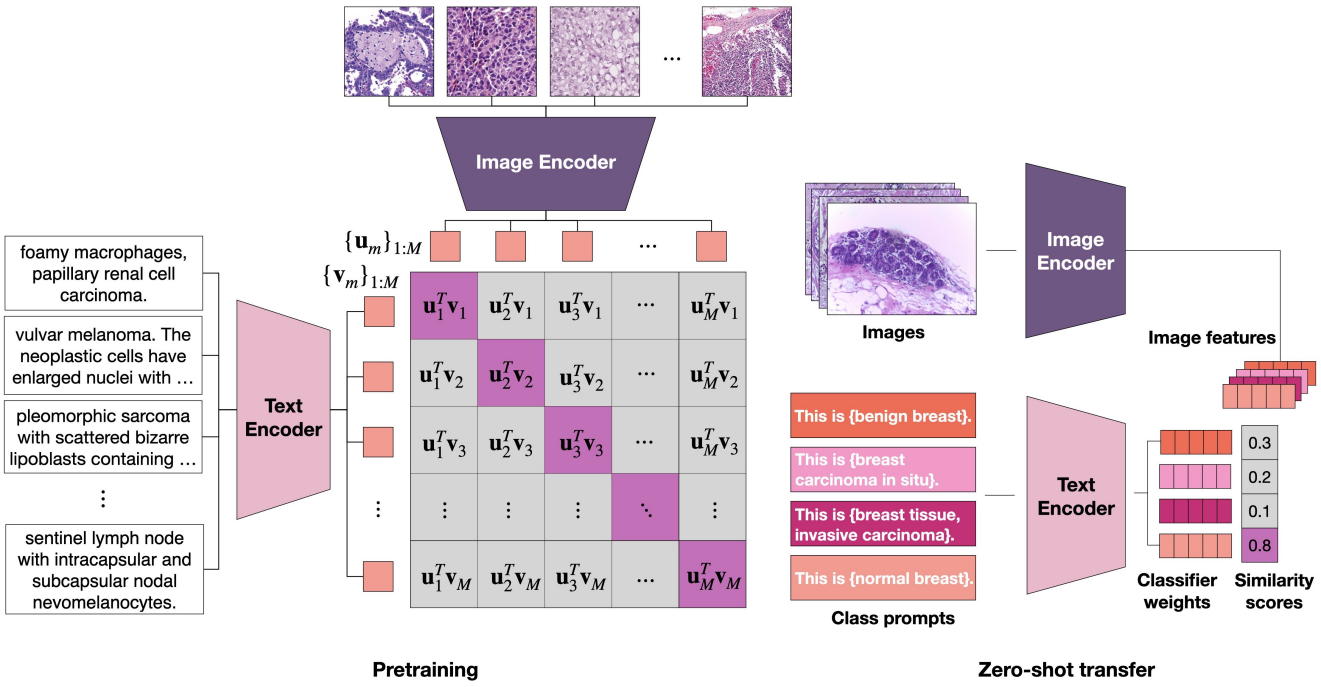

Figure 1. Illustration of contrasive **visual language pretraining** (left). Visual and language embeddings are aligned using a cross-modal contrastive loss. For our experiments, we curated the largest image-caption dataset in pathology consisting of 33,480 pairs. **Zero-shot transfer** for image classification (right), which we generalize for slide-level classification using MI-Zero (See Figure 2).

dataset. We perform cleaning and filtering, yielding a highly diverse dataset of 33,480 image-caption pairs covering a diverse set of tissue sites and morphologies (See **Supplementary Materials** for additional details).

3.2. Unsupervised pretraining of unimodal encoders.

While our paired dataset currently represents the largest of its kind in the domain of histopathology, it is still considerably smaller than what is feasible in the domain of radiology (*e.g.* MIMIC-CXR [34], 217k pairs) and what is used in representative works in general machine learning (*e.g.* LiT [86], 4B pairs, ALIGN [33], 1.8B pairs, CLIP [55], 400M pairs). Therefore, we initialize our encoders using pretrained weights before aligning their latent space using paired examples. For the text encoder, we collected a corpus specific to the domain of pathology, which notably includes the final diagnosis section of over 550k surgical pathology reports from Massachusetts General Hospital and over 400k histopathology-relevant PubMed abstracts. In-house diagnostic reports were cleaned and de-identified using regex. We pretrain a GPT-style autoregressive transformer (following the same architecture as GPT2-medium [57]) as the text encoder and will refer to it as HistPathGPT henceforth. Specifically, given a sequence of T word tokens $w_1, \ldots, w_T$, it is augmented to a sequence of length $T+2$: $\mathbf{t} = ([\text{BOS}], w_1, \ldots, w_T, [\text{EOS}])$. We maximize the log-likelihood of each token under an autoregressive generative model parameterized by ϕ:

$$\mathcal{L}_{\text{clm}}(\phi) = -\sum_{t=1}^{T+1} \log p\left(w_t \mid w_{0:t-1}; \phi\right) \tag{1}$$

Additionally, we also explore initializing from publicly available text encoders that have been trained on biomedical and clinical corpora such as PubMed abstracts and MIMIC [35]. We consider two pretrained models that fall into this category: BioClinicalBert [2] and PubMedBert [23]. For the image encoder, we explore 2 strategies: 1) initializing from ImageNet pretrained weights, and 2) using a SOTA publicly available encoder trained using self-supervised representation learning on a total of 15.5M unlabeled histopathology image patches [77].

3.3. Aligning vision and language embeddings.

We align the latent space of our visual and language encoders using the cross-modal contrastive loss formulated as a temperature scaled M-way classification [62], where M is the global batch-size of image-text pairs participating in the loss computation. Similar or analogous formulations of the contrastive loss are widely used for both self-supervised representation learning [14, 67] and visual-language pretraining [33, 55, 90]. Given a batch of M paired image and text samples $\{(\mathbf{x}_m, \mathbf{t}_m)\}_{m=1,\ldots,M}$, ℓ_2-normalized visual and text embeddings are computed via

the visual and text encoders $f(\cdot;\theta)$ and $g(\cdot;\phi)$ respectively as $\mathbf{u}_m = \frac{f(\mathbf{x}_m;\theta)}{\|f(\mathbf{x}_m;\theta)\|}$ and $\mathbf{v}_m = \frac{g(\mathbf{t}_m;\phi)}{\|g(\mathbf{t}_m;\phi)\|}$. The two directions of contrastive learning ($i \rightarrow t$) and $t \rightarrow i$) are viewed as symmetric and used jointly (with equal weight) to optimize the model during training, where τ is a temperature parameter:

$$\mathcal{L}_{i2t}(\theta,\phi) = -\sum_{i=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i\right)}{\sum_{j=1}^{M} \exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j\right)} \tag{2}$$

$$\mathcal{L}_{t2i}(\theta,\phi) = -\sum_{j=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_j\right)}{\sum_{i=1}^{M} \exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_i\right)} \tag{3}$$

For a batch of M image-text pairs, we see the connection to the aforementioned M-way classification problem by interpreting for each image (text), the index of its paired text (image) as the ground truth "target", and the remaining $M-1$ indices, which correspond to other texts (images) in the mini-batch as "negatives". Each direction of the contrastive loss is then equivalent to using the temperature scaled cosine similarity scores between embeddings as predicted logits, and minimizing the cross-entropy loss.

3.4. Zero-shot transfer for image classifcation.

We briefly describe the prompt-based approach to zero-shot classification popularized by [55]. For each class of interest, a prompt has two components, the classname (*e.g.* `"adenocarcinoma"`) and the template (*e.g.* `"an image showing {}."`), which collectively form the sequence of word tokens (*e.g.* `"an image showing adenocarcinoma."`) that are embedded by the trained text encoder to form the weights of a linear classifier. Formally, for an image $\mathbf{x}$, we compute its ℓ_2-normalized image embedding $\mathbf{u}$ using the image encoder. Given prompts $\{\mathbf{t}_m\}_{m=1,\ldots,C}$ where C is the total number of classes, the text encoder produces prompt embeddings $\{\mathbf{w}_m\}_{m=1,\ldots,C}$ where $\mathbf{w}_m = \frac{g(\mathbf{t}_m;\phi)}{\|g(\mathbf{t}_m;\phi)\|}$. The classification decision of the model is:

$$\hat{y} = \underset{m}{\operatorname{argmax}}\, \mathbf{u}^T \mathbf{w}_m \tag{4}$$

3.5. Zero-shot transfer for gigapixel WSIs.

There are several key challenges in performing zero-shot transfer for WSIs in the manner described in the previous section. First, each WSI can span up to $100,000 \times 100,000$ pixels, making it computationally intractable to directly compute an embedding vector using the image encoder [9]. Second, WSIs are known to be heterogeneous, comprising various tissue and cell types that can interact to form higher level architectures of both normal and diseased morphological patterns [1, 26, 27, 31]. In light of these challenges we propose MI-Zero, a zero-shot transfer framework for classifying WSIs inspired by the success of multiple instance learning for solving weakly-supervised learning tasks in computational pathology.

The approach entails first dividing each WSI into smaller tiles (called instances) more amenable to processing via our image encoder. We then consider the WSI as a collection of such instances by adopting either a permutation invariant set-based representation or a graph-based representation. Specifically, we consider dividing the tissue region of each WSI into N patches, and compute the ℓ_2-normalized embeddings of each patch independently using the image encoder to obtain $\{\mathbf{u}_i\}_{i=1,\ldots,N}$. We note that N is not a fixed constant, but instead varies depending on how large each WSI is, and therefore how many patches are obtained. Following the aforementioned prompt-based classification approach, we compute scores $\{\mathbf{s}_i\}_{i=1,\ldots,N}$:

$$\mathbf{s}_i = \mathbf{u}_i^T[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C], \quad \mathbf{s}_i \in \mathbb{R}^C \tag{5}$$

by measuring the cosine similarity between each patch embedding $\mathbf{u}_i$ and prompt embeddings $\{\mathbf{w}_m\}_{m=1,\ldots,C}$ defined earlier.

In the set-based representation, the set of scores $\mathcal{S} = \{\mathbf{s}_i\}_{i=1,\ldots,N}$ is directly passed to any permutation invariant operator $h(\cdot)$ such as the mean operator h_{mean} or topK max-pooling operator h_{topK} to produce the slide-level prediction scores (illustrated in Figure 2):

$$h_{\text{mean}}(\mathcal{S}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}_i \tag{6}$$

$$h_{\text{topK}}(\mathcal{S}) = \frac{1}{K}\left[\sum_{i=1}^{K} \tilde{s}_i^1, \sum_{i=1}^{K} \tilde{s}_i^2, \ldots, \sum_{i=1}^{K} \tilde{s}_i^C\right]^T \tag{7}$$

where $\mathcal{S}_{\text{topK}}^c = \{\tilde{s}_i^c\}_{i=1,\ldots,K}$ is the set of the K largest score values from $\mathcal{S}$ for class $c = 1,\ldots,C$. We note that in principle any permutation invariant pooling operator may be used here, as long as it is free of any learnable parameters (which are required in many attention-based methods) given the goal is to perform zero-shot transfer and no parameter update is allowed. In the graph-based representation, we take into account the spatial positions of each patch and build a directed KNN (k-nearest neighbors) graph $G = \{\mathcal{M}, \mathcal{E}\}$ connecting each patch (node) to its spatial neighbors, where the value at node i is its scores $\mathbf{s}_i$. Given the graph representation, we spatially smooth (*e.g.* average) the score values, by replacing $\mathbf{s}_i$ with $h_{\text{mean}}(\mathcal{S}_{\text{neighbors}})$, where $\mathcal{S}_{\text{neighbors}} = \{\mathbf{s}_j : j \in \{i\} \cup \mathcal{N}(i)\}$ and $\mathcal{N}(i) = \{j : (i,j) \in \mathcal{E}\}$ for each node i in the graph. We note that this is equivalent to applying a mean-filter with the receptive field size covering each patch's k-nearest neighbors. In principle, other filters such as the median or Gaussian filter can also

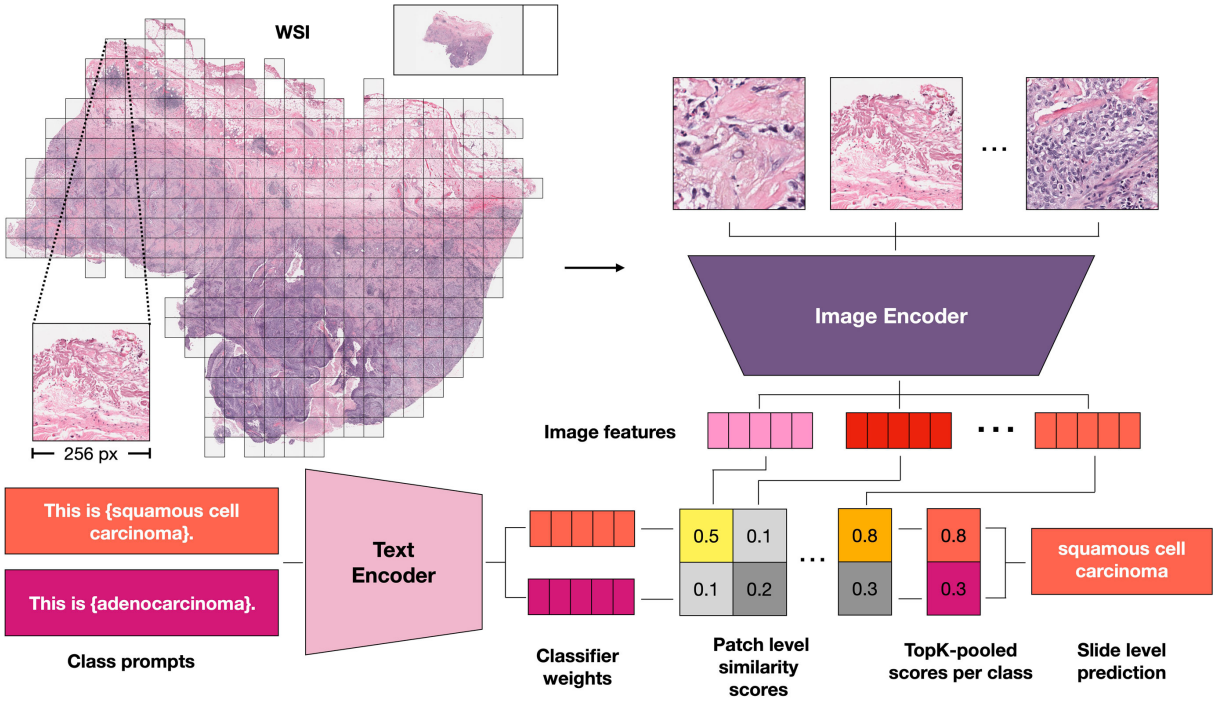

Figure 2. **Schematic of MI-Zero.** A gigapixel WSI is converted to a collection of patches (instances), each embedded into an aligned visual-language latent space. In the set-based representation, the similarity scores between patch embeddings and prompt embeddings are aggregated via a permutation invariant operator such as topK max-pooling to produce the WSI-level classification prediction. Alternatively, a graph-based representation may be used to incorporate spatial context by first aggregating predictions in local neighborhoods (Section 3.5).

be used, although we choose the simplest option (*i.e.* the mean) and leave other options for future exploration. After spatial smoothing, we then proceed by applying one of the possible permutation invariant pooling operators to the set of smoothed scores in the graph, $\mathcal{S}_{\text{smoothed}}$, and arrive at the slide-level prediction scores.

4. Experiments and results

4.1. Visual language pretraining.

For our HistPathGPT, we use a custom tokenizer trained on our dataset using Byte Pair Encoding (BPE) with a vocabulary size of 32,000. We use the hidden state corresponding to the position of the last word token to model the text embedding. For BioClinicalBert and PubmedBert we use the publicly available model weights and tokenizers as provided and use the [CLS] token as the text embedding. For the image encoder, by default we consider the SOTA CTransPath [77] (CTP) encoder trained using self-supervised representation learning on unlabeled histopathology patches, which has been shown to outperform ImageNet-initialized features by a wide margin on a range of different downstream tasks in CPATH. For all models, we use a linear projection head to map both the text and image embeddings into a 512-dimensional latent space for alignment. We align the representation space of our image and text encoders using the cross-view contrastive loss formulated in Section 3.3. We first preprocess all images to be of size 448×448 where images too large are first resized to 448 on the short side and then center cropped, while smaller images are zero-padded if needed. We use simple data augmentation techniques including horizontal and vertical flips applied to both images and captions. Each model is trained for 50 epochs on 8 A100 GPUs with a global batch size of 512. Other relevant hyperparameters and details on pretraining are included in the **Supplementary Materials**.

4.2. Downstream datasets.

After training on the image-text pairs as described in the previous section, we evaluate the zero-shot transfer performance for cancer subtype classification on 3 WSI datasets from Brigham and Women's Hospital. We used in-house independent datasets for zero-shot transfer evaluation because SSL encoders such as CTP are often trained on large public data repositories using all data that are available. This may result in information leakage if we use subsets of these public data sources for downstream evaluation, since their distribution was already exposed to the SSL encoder during its pretraining (transductive learning). Our in-house datasets are summarized below:

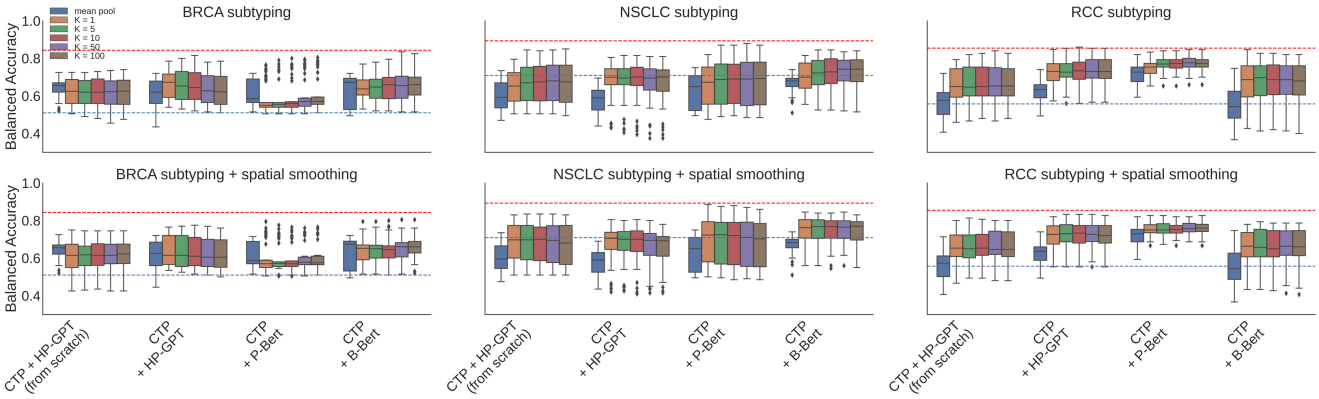

Figure 3. **Zero-shot transfer** performance of selected model configurations on independent test sets. Boxplots show distribution of model performance for 50 randomly sampled prompts. Columns show different subtyping tasks, rows show the absence or presence of spatial smoothing before pooling, and colors within each boxplot group show pooling methods (K indicates the number of patches selected by topK pooling). Red dashed line shows balanced accuracy of ABMIL trained on 100% of the corresponding TCGA cancer subsets averaged across 5 folds. Blue dashed line shows ABMIL performance trained on 1% of training data instead. **HP-GPT**: HistoPathGPT, **P-Bert**: PubMedBert, **B-Bert**: BioClinicalBert.

Independent BRCA is a dataset of invasive breast carcinoma (BRCA) histopathology WSIs. It consists of 100 slides of invasive ductal carcinoma (IDC) and 100 slides of invasive lobular carcinoma (ILC).

Independent NSCLC is a dataset of non-small cell lung cancer (NSCLC) histopathology WSIs. It consists of 100 slides of lung adenocarcinoma (LUAD) and 100 slides of lung squamous cell carcinoma (LUSC).

Independent RCC is a dataset of renal cell carcinoma (RCC) histopathology WSIs. It consists of 50 slides of chromophobe renal cell carcinoma (CHRCC), 50 slides of clear-cell renal cell carcinoma (CCRCC), 50 slides of papillary renal cell carcinoma (PRCC).

We include results on the publicly available datasets from The Cancer Genome Atlas (TCGA)[4] for the same 3 subtyping tasks in the **Supplementary Materials**.

4.3. Supervised baselines

To help contextualize the performance of zero-shot transfer models, we train supervised baselines using weakly-supervised attention-based MIL (ABMIL) [30] on the publicly available TCGA cohort of each task. Due to the relatively small size of these datasets (∼1000 WSIs each), we follow the study design of other weakly-supervised classification studies by performing 5-fold Monte Carlo cross-validation to train 5 different models and report their average performance on our in-house datasets. Each cross-validation training set includes on average 836 slides for BRCA, 838 for NSCLC, and 739 for RCC. We additionally study the data efficiency of ABMIL by restricting the number of training labels to be 1% and 10% of the full training set. We include more details about the training and results in the **Supplementary Materials**.

[4] portal.gdc.cancer.gov

4.4. Zero-shot transfer

Zero-shot evaluation methodology. Due to the reliance on prompts for zero-shot transfer, evaluation results vary with the choice of class names and prompt templates. For each task, we first curate a pool of relevant prompt templates and classnames (see **Supplementary Materials**). We then evaluate each model configuration on each task by randomly sampling 50 prompts and measuring the performance of each prompt. We plot the accuracy achieved using each prompt and report the median and interquartile range similar to [58]. This provides a more holistic view of the model's performance and demonstrates the degree of variation from using different prompts.

Zero-shot transfer for WSIs. For each of the 3 cancer subtyping classification tasks, we preprocess the WSIs by segmenting the tissue regions and dividing them into 256 × 256-sized patches at the 20× equivalent magnification. We then treat each WSI as a collection of its patches (instances) similar to MIL and use MI-Zero as described in Section 3.5 for zero-shot transfer. We compared performance between using a text encoder pretrained on in-domain pathology text data (HistPathGPT), encoders pretrained on non-pathology-specific medical data (PubMedBert and BioClinicalBert), as well as a text encoder trained from scratch. We also experimented with different pooling methods and spatial smoothing. Classification results comparing these setups are summarized in Table 1 and boxplots showing the performance distribution of each model on the set of 50 sampled prompts can be found in Figure 3. Overall, our models either perform on par or better than ABMIL baselines using 1% of training data for every task. In terms of pooling method for MI-Zero, we find that topK pooling performs better than mean pooling, while spatial smoothing does not change the

Model	Text Encoder & Pretraining	SS	Pooling	BRCA	NSCLC	RCC	Average
ABMIL (1% Data)	None	✗	attention	0.510	0.709	0.557	0.592
ABMIL (100% Data)	None	✗	attention	0.843	0.893	0.855	0.864
MI-Zero (Ours)	HistPathGPT (None)	✗	topK	0.625	0.680	0.653	0.653
	HistPathGPT (In-domain)	✗	topK	**0.673**	0.700	0.733	**0.702**
	PubmedBert (Out-of-domain)	✗	topK	0.570	0.693	**0.777**	0.680
	BioclinicalBert (Out-of-domain)	✗	topK	0.660	**0.742**	0.697	0.700
MI-Zero (Ours)	HistPathGPT (None)	✓	topK	0.623	0.700	0.653	0.659
	HistPathGPT (In-domain)	✓	topK	0.615	0.705	0.733	0.684
	PubmedBert (Out-of-domain)	✓	topK	0.577	0.725	**0.760**	0.688
	BioclinicalBert (Out-of-domain)	✓	topK	**0.660**	**0.770**	0.663	**0.698**
MI-Zero (Ours)	HistPathGPT (None)	✗	mean	0.655	0.593	0.577	0.608
	HistPathGPT (In-domain)	✗	mean	0.620	0.590	0.633	0.614
	PubmedBert (Out-of-domain)	✗	mean	0.585	0.650	**0.727**	**0.654**
	BioclinicalBert (Out-of-domain)	✗	mean	**0.672**	**0.680**	0.543	0.632
MI-Zero (Ours)	HistPathGPT (None)	✓	mean	0.655	0.595	0.573	0.608
	HistPathGPT (In-domain)	✓	mean	0.625	0.590	**0.637**	0.617
	PubmedBert (Out-of-domain)	✓	mean	0.587	0.650	0.730	**0.656**
	BioclinicalBert (Out-of-domain)	✓	mean	**0.675**	**0.682**	0.543	0.634

Table 1. **Slide-level zero-shot transfer.** All models shown here (including the supervised baseline ABMIL) use CTP as the image encoder. For MI-Zero, in-domain pretraining refers to pretraining on a corpus of pathology-specific text we collected while out-of-domain pretraining refers to non-pathology-specific corpora (See Section 3.2). SS means that spatial smoothing is used before pooling while topK and mean pooling refers to the pooling operator (Section 3.5). For each task, we report the median balanced accuracy across 50 sampled sets. For topK pooling, we report the highest performance across all $K \in \{1, 5, 10, 50, 100\}$. See Figure 3 for full distributions of results.

Dataset	BRCA	NSCLC	RCC	Average
CLIP [55]	0.500	0.500	0.333	0.444
ARCH [22]	0.625	0.593	0.540	0.586
Ours	**0.672**	**0.700**	**0.733**	**0.702**

Table 2. **Training data comparison**. We report balanced accuracy and only show results using topK pooling with no spatial smoothing. Since spatial smoothing yields the same trend, they are included in the **Supplementary Materials**. With the same MI-Zero setup, OpenAI's CLIP model [55] trained on 400M generic image-text pairs performs no better than random chance across all tasks. To assess the added value of our image-text pairs, we trained our best performing model configuration from Table 1 (CTP + HistPathGPT) on our full training dataset and compared to training only on ARCH (7,562 pathology pairs) [22], which is a subset of our training data (33,480 pathology pairs).

results significantly. We find that pretraining the text encoder improves performance over no pretraining, but pretraining on in-domain pathology text does not necessarily yield better performance. Example patches of highest and lowest similarity scores are visualized in Figure 4.

4.5. Ablation study

Training data comparison. To assess the benefit of pretraining with our expanded image-text dataset (compared to the smaller publicly available ARCH dataset originally proposed for representation learning via captioning), we train our best performing model configuration (CTP as image encoder and HistPathGPT pretrained on in-domain data as text encoder) on ARCH only. We find that training on our larger dataset improves performance across all tasks and raises the overall average performance by 11.6% (Table 2).

Image encoder pretraining. We experimented with the choice of image encoder by comparing CTP to encoders based on the ViT-S architecture, which has a similar parameter count. The encoders evaluated include both ImageNet initialization and pretraining with SSL (MoCo v3 [15]) on in-domain histology image data [77]. We also evaluate both the CTP and ViT-S encoders initialized fully from scratch with no pretraining as an additional ablation study. We find that pretraining both the image encoder and the text encoder performs the best across all tasks (Table 3).

Locked-image tuning. Zhai *et al.* [86] recently showed that "locking" a well-pretrained image encoder outperforms its unlocked counterpart during contrastive tuning. We therefore also explored locked-image tuning by freezing the parameters in the pretrained image encoder and only updating the text encoder. We find that when using the SSL-pretrained CTP as the image encoder and in-domain

Image Encoder	Text Encoder	Image Pretraining	Text Pretraining	BRCA	NSCLC	RCC	Average
CTP	HistPathGPT	SSL	In-domain	**0.672**	**0.700**	**0.733**	**0.702**
ViT-S	HistPathGPT	SSL	In-domain	0.617	0.625	0.673	0.639
ViT-S	HistPathGPT	ImageNet	In-domain	0.660	0.525	0.600	0.595
CTP	HistPathGPT	None	None	0.535	0.520	0.297	0.451
ViT-S	HistPathGPT	None	None	0.500	0.510	0.290	0.433

Table 3. **Pretraining comparison**. To assess the benefit of pretraining the image encoder, we compare our best performing model with a variation that uses ViT-S pretrained using SSL (MoCo v3), pretrained using supervised ImageNet, as well as variations with entirely randomly-initialized weights. We report balanced accuracy and we only show results using topK pooling with no spatial smoothing. Since spatial smoothing yields the same trend, we include those results in the **Supplementary Materials**.

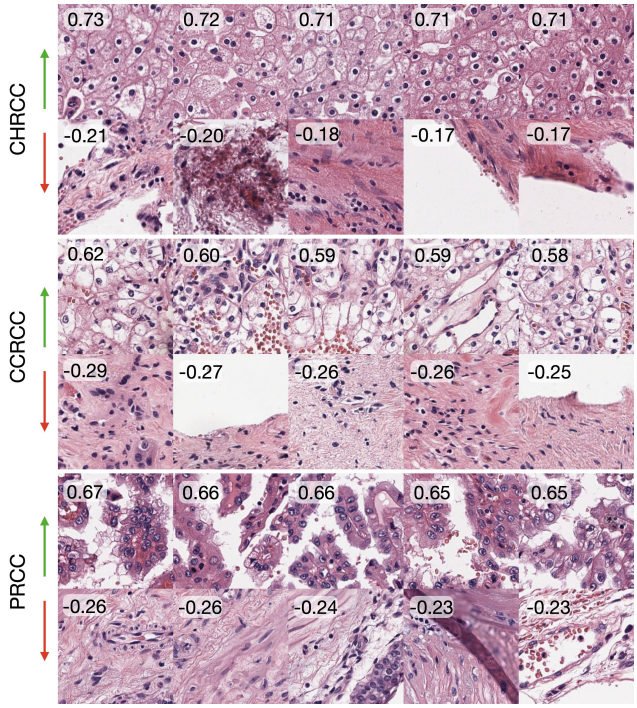

Figure 4. **Visualization of similarity scores.** A WSI of each RCC subtype (CHRCC, CCRCC and PRCC) is randomly selected from the in-house test set, and patches are ranked by their cosine similarity score with the class prompt embedding. The top (highest similarity scores) and bottom (lowest similarity scores) patches are displayed for each WSI. A board certified pathologist confirms relevant morphological patterns to each class embedding are selected by MI-Zero (high similarity scores), while low scores generally correspond to debris or normal tissue irrelevant to diagnosis. See **Supplementary Materials** for examples from other tasks.

pretrained HistPathGPT as the text encoder, locked-image text tuning only offers marginal improvement on zero-shot transfer performance. For all other configurations, locked-image tuning considerably lowers performance. We conjecture that by pretraining on in-domain data, the image and text features are easier to align in the latent space such that locked-image tuning was able to provide an improvement (See **Supplementary Materials**).

Additional experiments. Additional experimental results using TCGA WSIs, as well as run time analyses of MI-Zero, are included in **Supplementary Materials**.

5. Conclusion

In this work we introduce MI-Zero, the first method for zero-shot transfer in pathology, and apply it to gigapixel-scale whole slide images. The ability of our visual language pretrained model to retrieve relevant ROIs for a given class label (see Figure 4 with additional examples in **Supplementary Materials**) suggests potential usefulness for semi-supervised learning workflows in histopathology [8, 19, 54, 63] (e.g. as a way of performing pseudo-labeling). Our current results, however, are constrained by data limitations, as curating larger datasets of high quality image-caption pairs is a difficult task. Valuable future directions include collecting additional image caption datasets [22, 72], exploring methods that may improve the sample efficiency of visual language pretraining and also evaluating the capabilities of zero-shot transfer models on a large and diverse set of computational pathology benchmarks. We hope our work might inspire efforts to curate large scale pathology-specific image text datasets, and pave the way for a new generation of models in computational pathology capable of performing diverse visual language understanding tasks such as visual question answering, cross-modal retrieval, and captioning. Lastly, beyond pathology, many fields including satellite imaging and remote sensing involve high resolution images, similar to WSIs, in their workflow. MI-Zero can potentially be generalized and adapted to create effective solutions in such domains.

6. Acknowledgements

We thank Guillaume Jaume for his feedback. This work was supported in part by the BWH president's fund, BWH & MGH Pathology, and NIGMS R35GM138216 (F.M.). M.Y.L. was also supported by the Siebel Scholars program. R.J.C. was also supported by the NSF Graduate Fellowship.

References

[1] Shahira Abousamra, David Belinsky, John Van Arnam, Felicia Allard, Eric Yee, Rajarsi Gupta, Tahsin Kurc, Dimitris Samaras, Joel Saltz, and Chao Chen. Multi-class cell detection using spatial context representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4005–4014, 2021. 4

[2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019. 3

[3] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105, 2013. 2

[4] Jianpeng An, Yunhao Bai, Huazhen Chen, Zhongke Gao, and Geert Litjens. Masked autoencoders pre-training in multiple instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, 2022. 2

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 1

[6] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[7] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 1

[8] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020. 1, 8

[9] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. 1, 2, 4

[10] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1, 2

[11] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 339–349. Springer, 2021. 2

[12] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021. 2

[13] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022. 2

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

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 7

[16] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 1, 2

[17] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 2

[18] Donglin Di, Changqing Zou, Yifan Feng, Haiyan Zhou, Rongrong Ji, Qionghai Dai, and Yue Gao. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[19] Mohamed El Amine Elforaici, Emmanuel Montagnon, Feryel Azzi, Dominique Trudel, Bich Nguyen, Simon Turcotte, An Tang, and Samuel Kadoury. Semi-supervised tumor response grade classification from histology images of colorectal liver metastases. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022. 8

[20] Parsa Ashrafi Fashi, Sobhan Hemati, Morteza Babaie, Ricardo Gonzalez, and HR Tizhoosh. A self-supervised contrastive learning approach for whole slide image representation in digital pathology. *Journal of Pathology Informatics*, 13:100133, 2022. 2

[21] Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *arXiv preprint arXiv:2110.11316*, 2021. 2

[22] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology text-

books and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16549–16559, 2021. 2, 7, 8

[23] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021. 3

[24] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[26] Andreas Heindl, Sidra Nawaz, and Yinyin Yuan. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Laboratory investigation*, 95(4):377–384, 2015. 4

[27] Mahdi S Hosseini, Lyndon Chan, Gabriel Tse, Michael Tang, Jun Deng, Sajad Norouzi, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11747–11756, 2019. 4

[28] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016. 2

[29] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 2

[30] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 6

[31] Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical image analysis*, 63:101696, 2020. 4

[32] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, et al. Additive mil: Intrinsically interpretable multiple instance learning for pathology. In *Advances in Neural Information Processing Systems*. 2

[33] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2, 3

[34] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019. 3

[35] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 3

[36] Shivam Kalra, Mohammed Adnan, Graham Taylor, and Hamid R Tizhoosh. Learning permutation invariant representations using memory networks. In *European Conference on Computer Vision*, pages 677–693. Springer, 2020. 2

[37] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2

[38] Fanjie Kong and Ricardo Henao. Efficient classification of very large images with tiny objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2384–2394, 2022.

[39] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 2021. 1, 2

[40] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, pages 1–7, 2022. 1

[41] Ritika Kundra, Hongxin Zhang, Robert Sheridan, Sahussapont Joseph Sirintrapun, Avery Wang, Angelica Ochoa, Manda Wilson, Benjamin Gross, Yichao Sun, Ramyasree Madupuri, et al. Oncotree: a cancer classification system for precision oncology. *JCO Clinical Cancer Informatics*, 5:221–230, 2021. 1

[42] Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, et al. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, pages 1–15, 2022. 2

[43] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 1, 2

[44] Jiajun Li, Tiancheng Lin, and Yi Xu. Sslp: Spatial guided self-supervised learning on pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2021. 1

[45] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[47] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021. 1

[48] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 2

[49] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 2

[50] Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):1–10, 2019. 1

[51] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[52] Pushpak Pati, Guillaume Jaume, Antonio Foncubierta-Rodriguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, et al. Hierarchical graph representations in digital pathology. *Medical image analysis*, 75:102264, 2022. 2

[53] AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch, and Anelia Angelova. Answer-me: Multi-task open-vocabulary visual question answering. *arXiv preprint arXiv:2205.00949*, 2022. 2

[54] Linhao Qu, Siyu Liu, Xiaoyu Liu, Manning Wang, and Zhijian Song. Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. *Physics in Medicine & Biology*, 2022. 8

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 4, 7

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 3

[58] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. 6

[59] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision*, pages 153–170. Springer, 2020. 2

[60] Andrew B Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022. 2

[61] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 2

[62] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 3

[63] Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022. 8

[64] Chetan L Srinidhi and Anne L Martel. Improving self-supervised learning with hardness-aware dynamic curriculum learning: An application to digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 562–571, 2021. 1, 2

[65] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):567–578, 2019. 2

[66] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 699–715. Springer, 2022. 2

[67] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 2, 3

[68] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 2022. 2

[69] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn

for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12597–12606, 2019. 2

[70] Yuri Tolkach, Tilmann Dohmgörgen, Marieta Toma, and Glen Kristiansen. High-accuracy prostate cancer pathology using deep learning. *Nature Machine Intelligence*, 2(7):411–418, 2020. 1

[71] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019. 1

[72] Masayuki Tsuneki and Fahdi Kanavati. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, pages 1235–1250. PMLR, 2022. 8

[73] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021. 1

[74] Quoc Dang Vu, Kashif Rajpoot, Shan E Ahmed Raza, and Nasir Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical Image Analysis*, page 102743, 2023. 2

[75] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[76] Xiyue Wang, Jinxi Xiang, Jun Zhang, Sen Yang, Zhongyi Yang, Ming-Hui Wang, Jing Zhang, Yang Wei, Junzhou Huang, and Xiao Han. Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. In *Advances in Neural Information Processing Systems*. 2

[77] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022. 1, 2, 3, 5, 7

[78] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 2

[79] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[80] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data efficient language-supervised zero-shot recognition with optimal transport distillation. *arXiv preprint arXiv:2112.09445*, 2021. 2

[81] Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[82] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 2

[83] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 2

[84] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. 2022. 2

[85] Cui Yufei, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B Chan. Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[86] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2, 3, 7

[87] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 2

[88] Jingwei Zhang, Ke Ma, John Van Arnam, Rajarsi Gupta, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3776–3784, 2021. 2

[89] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 2

[90] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 2, 3

[91] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846, 2020. 1, 2

[92] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7234–7242, 2017. 1

[93] James Y Zou, Daniel J Hsu, David C Parkes, and Ryan P Adams. Contrastive learning using spectral methods. *Advances in Neural Information Processing Systems*, 26, 2013. 2