CLIPath: Fine-tune CLIP with Visual Feature Fusion for Pathology Image Analysis Towards Minimizing Data Collection Efforts

Zhengfeng Lai[1] Zhuoheng Li[2] Luca Cerny Oliveira[1]
Joohi Chauhan[1] Brittany N. Dugger[3] Chen-Nee Chuah[1]
University of California, Davis
[1]Department of Electrical and Computer Engineering [2]Department of Computer Science
[3]Department of Pathology and Laboratory Medicine
{lzhengfeng, pipli, lcernyo, jhichauhan, bndugger, chuah}@ucdavis.edu

Abstract

Contrastive Language-Image Pre-training (CLIP) has shown its ability to learn distinctive visual representations and generalize to various downstream vision tasks. However, its applicability in the classification of pathology images with limited labeled data is still under study due to the giant domain shift (between large natural image datasets in the source domain and small-scale target pathology images) and overfitting issues. In this work, we first explore the zero-shot transferability of CLIP on pathology classification tasks and benchmark the performance. Then, we propose Residual Feature Connection (RFC) to fine-tune CLIP with a small amount of trainable parameters. RFC aims to fuse the task-specific knowledge learned from the target domain and the original knowledge pre-trained from CLIP. We show that RFC can adapt pre-trained CLIP to downstream pathology tasks and achieve good performance with just a few annotated samples. Specifically, RFC achieves over 19% improvement in accuracy when only using 0.1% of labeled data in PCam with only 10 minutes of fine-tuning while running on a single GPU.

1. INTRODUCTION

Deep learning with better network designs and large-scale well-curated datasets has achieved significant performance improvement in pathology image analysis tasks [13, 14]. However, collecting high-quality datasets with reliable annotations for every vision task can be time-consuming and labor-extensive [15, 38]. This may prevent the broad adoption of advanced deep learning techniques. To relieve the reliance on such datasets, pre-training and fine-tuning methods have been studied in vision tasks: pre-train the model on a large-scale dataset and then fine-tune the model on different downstream tasks [6]. There are several challenges of such methods: 1) they may still require a large amount of labeled set to avoid the overfitting issue when fine-tuning the model for the downstream task [17, 31]; 2) the fine-tuning may not bring satisfactory performance in the target domain due to the existence of a large domain gap between the pre-trained data and pathology images [25].

To fill the performance gap due to domain shift, Contrastive Language-Image Pre-training (CLIP) [19] has shown its power in learning generic and distinctive visual representations via language supervision. It aligns images and texts in the same feature space and uses a contrastive loss to formulate the learning objective. CLIP uses two separate encoders for images and texts, then maximizes the similarity score of positive pairs of images and texts while minimizing for the negative pairs [19, 39]. It achieves promising results on various image classification tasks without needing any annotated data, i.e., zero-shot transfer settings. As a language-vision model, CLIP uses prompts as the supervision, where the visual labels are entered into the hand-crafted template. By pre-training the model at a large scale, models can learn the visual contents and easily be transferred to downstream tasks through the prompt-based zero-shot transfer.

However, the manual design of prompts can be a non-trivial and time-consuming task. In [39], the authors found that even a slight change in the prompt (e.g., one word) can make a big difference. They introduced Context Optimization (CoOp) to automate prompt engineering to generate continuous soft prompts instead of using hand-chosen hard prompts [39]. CoOp requires substantial computing resources, and the results of CoOp are not interpretable. Besides, CoOp faces performance degradation when there is a big domain shift, e.g., from natural to pathology images, making it hard to adapt to medical imaging tasks.

In this work, we aim to fine-tune CLIP efficiently with light computing resources for pathology image classifica-

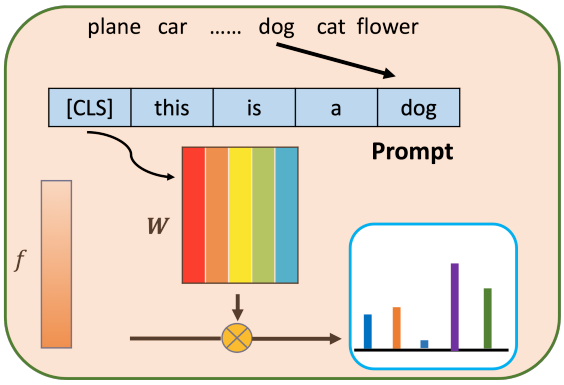

Figure 1. An overview on the inference stage of CLIP in computer vision tasks: f is the output from the vision encoder while W is the output from the text encoder.

tion tasks. There are several challenges in fine-tuning CLIP. First, overfitting is a severe issue if we directly adapt CLIP to the downstream tasks since CLIP is pre-trained on a 400M dataset while the new domain dataset can be small [7]. Second, it is unclear how to effectively learn the new knowledge while retaining the original pre-trained knowledge to maintain the generalization of CLIP. Third, there are limited studies on benchmarking CLIP's transferability in the pathology domain; hence, its applicability remains unclear.

To address the above issues, we first study the applicability of CLIP on two pathology datasets and benchmark the zero-shot ability of CLIP on them. Then, we propose Residual Feature Connection (RFC) as a lightweight approach for adapting CLIP to pathology images. It will fuse the original knowledge from CLIP and the new knowledge learned from the new pathological task with only a tiny number of trainable additional weights instead of optimizing the entire encoders in CLIP. Third, to further improve the fine-tuning ability, we propose using Language-Vision Alignment (LVA) in the fine-tuning stage to mimic contrastive learning in the pre-training stage.

We summarize our contributions as follows:

- We explore the applicability of CLIP on pathology images and benchmark its zero-shot transfer ability.
- We propose CLIPath to introduce CLIP in pathology image applications. In CLIPath, we propose Residual Feature Connection (RFC) and Language-Vision Alignment to fine-tune CLIP on pathology tasks with limited labeled data.
- We show that CLIPath has the potential to quickly adapt pre-trained CLIP to downstream tasks with good performance but light computational cost.

2. Related Work

2.1. Language-vision model

Language-vision models have exhibited promising performance in acquiring general visual representations [11, 18, 19, 35]. Recent advancements in these models involve text representation learning using large-scale Transformers [23] and training on extensive datasets from the web [39]. Transformer-based multimodal learning has achieved remarkable success on such vast datasets [22, 29]. For instance, CLIP [19] was trained on 400 million image-caption pairs and achieved state-of-the-art performance across various domains [19, 32–34]. CLIP comprises two encoders: a vision encoder, which can be ResNet [9] or ViT [4], and a text encoder, such as Transformer (e.g., BERT [3]).

In a recent study [20], CLIP was fine-tuned for video data and demonstrated competitive results compared to more complex methods specifically designed for video processing. Another application, PointCLIP [37], employed CLIP for 3D recognition. CLIP has also been utilized for image generation tasks [5] and exhibits the ability to reduce data collection. However, the optimal approach for adapting CLIP to downstream tasks is still under investigation, particularly when the new domain, such as the medical field, significantly differs from the pre-trained domains.

2.2. Language-vision training in medical domain

Language-vision pre-training, which involves training models to understand language and visual information, typically relies on vast amounts of web images and captions from diverse domains. For instance, the CLIP model utilizes a 400 million image-caption pairs dataset [19]. However, medical datasets are considerably smaller in comparison, posing a challenge in applying pre-training methods to the medical domain. Additionally, annotating medical images require specialized domain knowledge [15], further increasing the cost of training when dealing with multiple medical tasks. To illustrate, Lai *et al.* conducted a study on the distribution of Amyloid-β plaques, a prominent pathology in Alzheimer's disease, in grey and white matter. This investigation involved two learning tasks: image segmentation and object detection, each requiring separate datasets and their respective annotations [13].

Although expert annotation of medical images is already expensive, the situation becomes even more challenging when incorporating language-vision training due to the need for captions and prompts for the images. For example, MedCLIP [27] achieved 60% zero-shot accuracy by employing 570,000 image-text pairs. Acquiring datasets of such magnitude is particularly daunting in pathology image tasks, where each slide is at the gigapixel level. Efforts to adapt CLIP for the medical domain have been limited, pri-

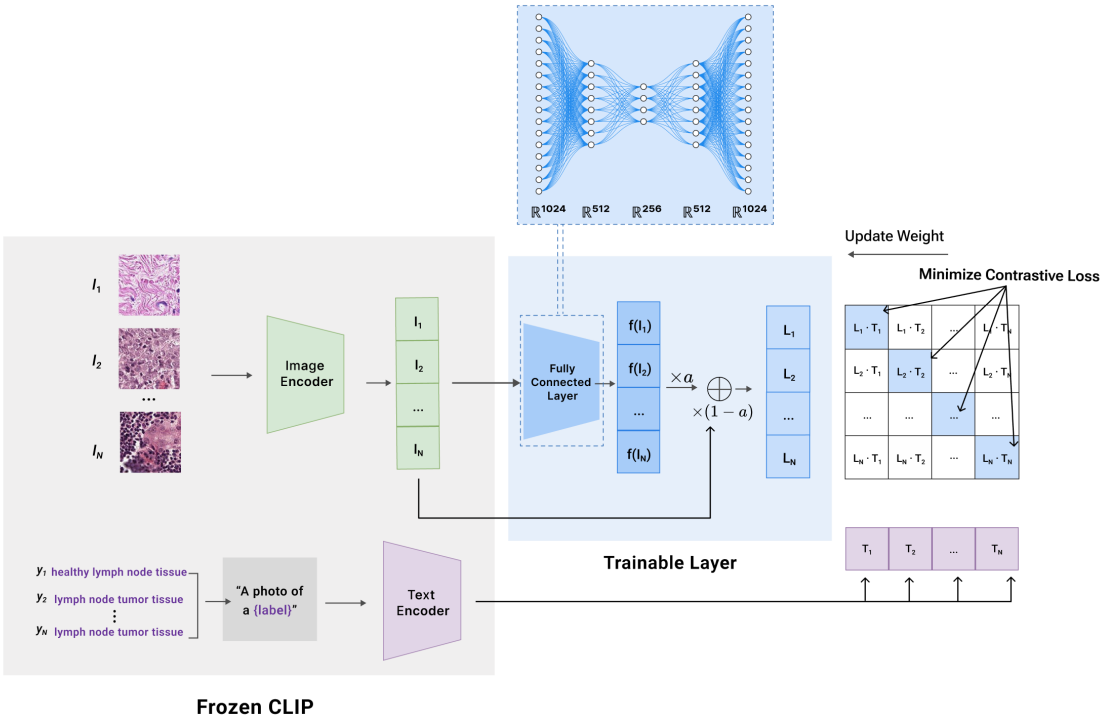

Figure 2. Overview of the proposed framework: the image and text encoders of CLIP are frozen while the Trainable Layer (RFC) is of a downsampling-upsampling architecture of linear layers. It blends the fine-tuned knowledge with the original knowledge from CLIP's vision encoder ($F(\cdot)$).

marily due to the substantial disparity between general images and medical images. In this study, our objective is to develop an efficient adaptation framework that can be easily applied to multiple downstream tasks in the medical field while addressing the scalability issue.

3. METHODS

3.1. CLIP and Setup

CLIP [19] has a vision encoder $F(\cdot)$ and a text encoder $G(\cdot)$. The vision encoder maps a high-dimensional image into low-dimensional image embeddings. The text encoder is built on Transformer [23] and generates text embeddings from the prompt. During training, CLIP jointly trains $F(\cdot)$ and $G(\cdot)$ to optimize the similarity score (e.g., symmetric cross-entropy loss [26]) between the visual and textual embeddings for each batch. Specifically, the input consists of an image and its corresponding prompt (e.g., "this is a dog"). Then given a batch of image-prompt pairs, CLIP maximizes the similarity score for positive pairs while minimizing it for negative pairs.

For the inference process, as shown in Fig. 1, an image I is transformed into a feature manifold $f \in \mathbb{R}^D$, where D is the feature dimension. Then, f is multiplied with a classifier weight matrix $\mathbf{W} \in \mathbb{R}^{D\times K}$, where K is the number of classes in the learning task. We get a K-dimensional logit after matrix multiplication. Then we apply softmax to convert this logit into a probability vector $p \in \mathbb{R}^K$ over the K classes. The whole process can be summarized as the following equation:

$$p_i = \frac{\exp(\mathbf{W_i}^T \times f)/\tau}{\sum_{i=1}^{K} \exp(\mathbf{W_i}^T \times f)/\tau}, \quad (1)$$

where τ is the temperature parameter learned by CLIP during training and $\mathbf{W_i}$ is the prototype weight vector for class-i.

3.2. Residual Feature Connection

In this subsection, we introduce Residual Feature Connection (RFC) to learn the task-relevant context when we adapt CLIP to the downstream learning tasks. Inspired by

CLIP-Adapter [6] that achieves promising results on computer vision few-shot benchmark, we argue the importance of preserving the original knowledge and fusing the new knowledge. Unlike CoOp's prompt tuning, which may not address the domain shift issue between the natural and pathology images, we focus on fine-tuning the visual features f. However, simple fine-tuning of the entire network may fail in a new pathological task due to the overfitting issue caused by a large number of parameters and a shortage of the training samples [10]. Inspired by [10] that fine-tunes the model with additional layers, we argue the importance of retaining the features from CLIP and propose a residual connection architecture to dynamically fuse the fine-tuned knowledge with the original CLIP's feature.

As shown in Fig. 2, given an image X, we get the visual feature f from the image encoder and compute the classifier weight $\mathbf{W}$ from the text encoder. Then we design trainable fine-tuning layers L to convert f into $L(f)$. L can be multiple layers of linear transformations in a "down-and-up" architecture. As shown in Fig. 2, we have four layers to transform the feature dimension as "1024-256-64-256-1024" so that $L(f)$ can be of the same size of f and blended with f as follows:

$$f^* = \alpha L(f) + (1 - \alpha) f, \tag{2}$$

where α is a residual ratio to balance the fine-tuned knowledge and the original CLIP's knowledge. Then we adopt Equation 1 with the new f^* to get the class probability vector and predict the category with the highest probability. During the fine-tuning, the weights of the trainable layers are optimized through the symmetric contrastive loss used in CLIP [19]. The prompt here is "this is a photo of []", where "[]" is filled with the class name. For example, "[]" can be "healthy lymph node tissue" or "healthy lymph tumor tissue" in PCam [24].

3.3. Language-vision Alignment

Although CLIP has shown promising zero-shot ability, there is a strong preference to enhance performance by engaging in supervised fine-tuning in many scenarios. This involves additional training and adjustments to the pretrained parameters using a limited set of labeled images, which can lead to further improvements [8]. While RFC can retain the pre-trained knowledge and learn new knowledge, it may still suffer from overfitting issues when the target dataset is too small. Goyal *et al.* [8] studies the importance of contrastive loss in the fine-tuning stage to alleviate this issue. Different from previous fine-tuning approaches that minimize a standard supervised loss (e.g., cross-entropy loss on an image classifier), they claim that keeping the contrastive loss used in the pre-training stage is more advantageous. Therefore, in CLIPath, we follow [8] and use a contrastive loss in the fine-tuning.

4. EXPERIMENTS

4.1. Data Preparation and Setup

The datasets used for validation of our framework are collected from two distinct pathology projects [24, 28]. Both projects make their data available as patches, extracted from H&E stained Whole Slide Images (WSI) digitized from Formalin-Fixed Paraffin-Embedded (FFPE) slides. Each dataset has a distinct binary classification task aimed at detecting different cancerous tissue.

Minimalist Histopathology Image Analysis Dataset (MHIST) [28]. MHIST contains 3,152 patches from colorectal regions at 224×224 pixel resolution. These patches were extracted from 328 WSIs scanned at $40\times$ resolution. Each patch may be labeled as Hyperplastic Polyp (HP) or Sessile Serrated Adenoma (SSA). HP is the majority class with 68.59% of the labels. The labeling of colorectal polyps between HP and SSA is a challenge due to high inter-pathologist disagreement. Seven pathologists contributed in the ground truth to ensure reliable labels.

PatchCamelyon (PCam) [24]. PCam patches were extracted from Camelyon16 challenge [1]. The original 400 WSIs from Camelyon16 were digitized breast tissue with potential metastasized cancerous tissue on the lymph nodes. The original WSIs were scanned at $40\times$ resolution but later downsampled to $10\times$. The WSIs were collected from two different centers. PCam extracted 327,680 patches at 96×96 resolution and labeled them as positive or negative. Positive labeled patches present tumor tissue in the central 32×32 patch region. There are an equal amount of positive and negative labeled samples.

For a fair comparison, we use ResNet-50 [9] as the backbone in the vision encoder $f(\cdot)$. We follow CLIP [19] to use gradient scaling in facilitating mixed-precision training. We set the learning rate as 0.0001. The batch size is 32. We use Adam [12] as the optimizer. All of the experiments are conducted on one piece of GPU (Nvidia RTX 2080Ti) to compare the computational complexity.

4.2. Main Results

In this subsection, we report the results on PCam [24] and MHIST [28] to show how RFC's fine-tuning improves CLIP. We select Accuracy, Recall, Precision, F1-score, and AUC to have a comprehensive comparison. The main results on PCam [24] when only using 0.5% data for fine-tuning are summarized in Table 1. First, CLIP shows its strong zero-shot ability: it can achieve 56.5% accuracy under this setting. Our proposed method can get better results than semi-supervised learning approaches that use extra unlabeled data. Therefore, CLIP and its efficient adaptation are promising for minimizing data collection efforts in clinical applications. Similar observations can also be found in MHIST, summarized in Table 2.

Table 1. Quantitative comparison on PCam. Precision, Recall, and F1-score refer to the macro-averaged values from all classes

Learning Manner	Labeled Ratio	Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Supervised	100%	-	92.8 ± 0.30	92.9 ± 0.20	92.6 ± 0.21	92.8 ± 0.20	0.95 ± 0.01
Supervised	0.5%	-	45.1 ± 2.27	42.0 ± 2.14	33.6 ± 3.30	37.3 ± 2.86	0.51 ± 0.01
Self-supervised	0.5%	SimCLR [2]	60.4 ± 1.45	63.3 ± 1.90	58.5 ± 2.01	60.8 ± 1.20	0.61 ± 0.04
Semi-supervised	0.5%	Pseudo-Label [16]	55.7 ± 2.05	59.2 ± 2.23	54.5 ± 1.97	56.8 ± 1.55	0.58 ± 0.03
		FixMatch [21]	73.2 ± 0.76	77.5 ± 0.50	73.7 ± 0.25	75.6 ± 0.37	0.84 ± 0.04
		Dash [30]	71.0 ± 0.98	75.3 ± 0.78	70.2 ± 0.45	72.7 ± 0.50	0.82 ± 0.02
		FlexMatch [36]	74.0 ± 0.80	78.1 ± 1.15	73.0 ± 0.90	75.5 ± 0.84	0.85 ± 0.03
Multimodal CLIP	-	Zero-shot [19]	56.5 ± 0.00	57.4 ± 0.00	50.3 ± 0.00	53.7 ± 0.00	0.60 ± 0.00
	0.5%	CoOp [39]	63.6 ± 0.25	63.9 ± 0.42	62.5 ± 0.35	63.0 ± 0.29	0.67 ± 0.02
		CLIP-Adapter [6]	72.3 ± 1.02	77.2 ± 0.95	63.2 ± 0.56	69.4 ± 0.84	0.81 ± 0.02
		Proposed	**81.5** ± 0.78	**79.4** ± 0.35	**85.0** ± 0.80	**82.1** ± 0.95	**0.89** ± 0.02

Table 2. Quantitative results on the hold-out test set of MHIST.

Algorithm	Data Usage	Accuracy	Recall	Precision	F1-score	AUC
CLIP [19]	Zero-shot	36.9	100.0	36.9	53.9	0.501
CLIP + RFC	1%	63.9	7.5	57.5	13.3	0.643
	5%	66.8	42.8	56.6	48.7	0.732
	10%	70.5	79.7	57.1	66.6	0.784
	20%	70.7	86.1	56.8	68.4	0.788
	50%	74.8	75.6	63.3	68.9	0.838

Table 3. Quantitative results on the hold-out test set of PCam.

Algorithm	Data Usage	Accuracy	Recall	Precision	F1-score	AUC
CLIP [19]	Zero-shot	56.5	50.3	57.4	53.7	0.600
CLIP + RFC	0.1%	76.4	90.0	70.7	79.2	0.849
	0.5%	81.5	85.0	79.4	82.1	0.894
	1%	81.9	82.9	81.3	82.1	0.900
	5%	82.9	77.1	87.2	81.8	0.918
	10%	82.8	79.2	85.4	82.1	0.914
	50%	81.4	71.0	89.6	79.3	0.918

4.3. A Closer Look at RFC

4.3.1 How Does RFC Improve CLIP?

In this subsection, we have a deeper look at RFC and study how it improves CLIP. As shown in Table 3, RFC can bring over 25% improvement in accuracy and 28.4% in F1-score by only using 0.5% of data in PCam [24] compared to the original CLIP. In Table 2, we get significant improvement in accuracy and AUC on MHIST [28], while the performance in F1-score and recall seems limited. We diagnose that MHIST is a more challenging task and has an inter-rate agreement issue, which may confuse the model during fine-tuning.

4.3.2 Compare with CoOp [39]

CoOp [39] is the recent state-of-the-art fine-tuning method for CLIP. Hence we mainly compare our proposed RFC with it in both performance and computational complexity. We summarize the results in Table 4. The training time refers to the time period from the start point to the time point when the validation set gets the best results. When we range the data usage from 0.1% to 1%, we find that RFC can get 5.5% improvement in accuracy while only introducing 1 minute of additional training time. However, CoOp [39] has the overfitting issue and gets a lower score while using 53 minutes, which is almost 5 times compared to the time used by RFC. The situation remains similar under the 10% of data usage. We conclude that our proposed RFC can get over 25% improvement on CLIP by only using 10 minutes for the fine-tuning on a single GPU, which is promising for digital pathology research.

Table 4. Performance and Complexity comparison on PCam.

Data Usage	Algorithm	Accuracy	Training Time
Zero-shot	CLIP [19]	56.5	-
0.1%	CLIP + CoOp [39]	64.3	**7 min 6 sec**
	CLIP + RFC	**76.4**	10 min 29 sec
1%	CLIP + CoOp [39]	61.9	53 min 21 sec
	CLIP + RFC	**81.9**	**11 min 56 sec**
10%	CLIP + CoOp [39]	59.9	2 h 23 min 45 sec
	CLIP + RFC	**82.8**	**27 min 18 sec**

5. Discussion

In this work, we explore the generalization of Contrastive Language-Image Pre-training (CLIP) in pathology image classification. We propose RFC to efficiently fine-tune CLIP using a small dataset and light computing resources. On the other hand, we use a contrastive loss in the fine-tuning stage to preserve the model's capacity. We show that RFC has the potential to bridge the domain shift between the pre-trained natural images and pathology images. However, we only evaluate our frameworks on two small-scale datasets. In the future, we aim to test it over more diverse pathology image tasks.

ACKNOWLEDGMENT

This work was supported by the Noyce Initiative UC Partnerships in Computational Transformation Grant and the UC Davis Center for Women's Cardiovascular and Brain

Health research program under the HEAL-HER (Heart, BrEast, and BrAin HeaLth Equity Research) award made possible by the Cy Pres funds. This work also received additional partial support from National Institutes of Health grants P30 AG072972 and R01 AG062517.

References

[1] 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. 4

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 5

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[5] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2

[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1, 4, 5

[7] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 2

[8] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023. 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 4

[10] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799. PMLR, 2019. 4

[11] 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 *ICML*, pages 4904–4916, 2021. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 4

[13] Zhengfeng Lai, Luca Cerny Oliveira, Runlin Guo, Wenda Xu, Zin Hu, Kelsey Mifflin, Charles Decarli, Sen-Ching Cheung, Chen-Nee Chuah, and Brittany N Dugger. Brainsec: Automated brain tissue segmentation pipeline for scalable neuropathological analysis. *IEEE Access*, 10:49064–49079, 2022. 1, 2

[14] Zhengfeng Lai, Chao Wang, Zin Hu, Brittany N Dugger, Sen-Ching Cheung, and Chen-Nee Chuah. A semi-supervised learning for segmentation of gigapixel histopathology images from brain tissues. In *2021 43nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021. 1

[15] Zhengfeng Lai, Chao Wang, Luca Cerny Oliveira, Brittany N Dugger, Sen-Ching Cheung, and Chen-Nee Chuah. Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling. In *ICCV Workshop*, pages 591–600, 2021. 1, 2

[16] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning (ICML)*, 2013. 5

[17] Min Liu, Lanlan Hu, Ying Tang, Chu Wang, Yu He, Chunyan Zeng, Kun Lin, Zhizi He, and Wujie Huo. A deep learning method for breast cancer classification in the pathology images. *IEEE J. Biomed. Health. Inf.*, 26(10):5025–5032, 2022. 1

[18] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, pages 5206–5215, 2022. 2

[19] 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 *ICML*, pages 8748–8763, 2021. 1, 2, 3, 4, 5

[20] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[21] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, volume 33, pages 596–608, 2020. 5

[22] Yuxin Tian, Shawn Newsam, and Kofi Boakye. Fashion image retrieval with text feedback by additive attention compositional learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1011–1021, 2023. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2, 3

[24] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, pages 210–218. Springer, 2018. 4, 5

[25] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *MICCAI*, pages 186–195. Springer, 2021. 1

[26] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019. 3

[27] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *Proceedings of EMNLP*, 2022. 2

[28] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Int. Conf. Arti. Intell. in Medi.*, pages 11–24. Springer, 2021. 4, 5

[29] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[30] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021. 5

[31] Jiawei Yang, Hanbo Chen, Yuan Liang, Junzhou Huang, Lei He, and Jianhua Yao. Concl: Concept contrastive learning for dense prediction pre-training in pathology images. In *ECCV*, pages 523–539. Springer, 2022. 1

[32] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[33] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[34] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, June 2022. 2

[35] 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 *CVPR*, pages 18123–18133, 2022. 2

[36] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NeurIPS*, 34, 2021. 5

[37] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 2

[38] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *CVPR*, pages 20666–20676, 2022. 1

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vision*, 130(9):2337–2348, 2022. 1, 2, 5