Path-CLIP: Efficient Adaptation of CLIP for Pathology Image Analysis with Limited Data

Zhengfeng Lai[1], Joohi Chauhan[1], Zhuoheng Li[2], Luca Cerny Oliveira[1], Brittany N. Dugger[3], and Chen-Nee Chuah[1], *Fellow, IEEE*

***Abstract*—Contrastive Language-Image Pre-training (CLIP) has shown its ability of learning distinctive visual representations and generalization to various downstream vision tasks. However, its applicability in pathology image analysis with limited labeled data is still under study due to the giant domain shift and catastrophic forgetting issues. Furthermore, pathological tasks may need to be separated into multiple learning tasks. Hence, it is critical to study efficient adaptation schemes in this domain for scalable analysis. In this work, we propose Path-CLIP to quickly adapt CLIP to multiple pathology tasks. First, we propose Residual Feature Connection (RFC) with a self-adaptive ratio to fuse and balance the source and task-specific knowledge. Second, we propose Hidden Representation Perturbation (HRP) and Dual-view Vision Contrastive (DVC) to alleviate the overfitting issue. Lastly, we propose Doublet Multimodal Contrastive Loss (DMCL) to adapt CLIP to pathology tasks. We show that Path-CLIP can adapt pre-trained CLIP to downstream pathology tasks and achieve competitive results. Specifically, Path-CLIP 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.**

***Index Terms*—Language-vision model, medical image analysis, multimodal applications**

I. INTRODUCTION

DEEP learning with better network designs and large-scale well-curated datasets has achieved significant performance improvement in pathology image analysis tasks [1], [2]. However, collecting high-quality datasets with reliable annotations for each individual vision task can be time-consuming and labor-intensive [3], [4]. This may prevent the broad adoption of advanced deep learning techniques. To relieve the reliance on such datasets, pre-training and fine-tuning methods (TFT) 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 [5]. 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 [6], [7]; 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 [8]. Therefore, a natural question is how to efficiently fit the model to the new task with minimal overfitting when the new dataset is small.

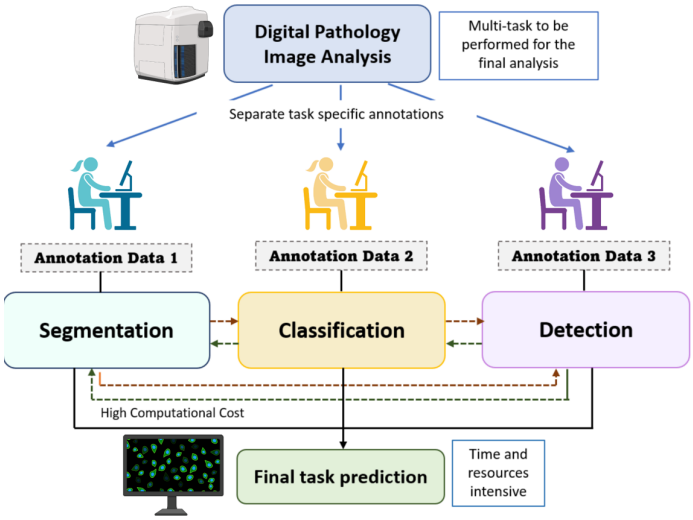


Fig. 1. Overview of the traditional computer vision training in medical image analysis task.

To fill the performance gap due to domain shift, Contrastive Language-Image Pre-training (CLIP) [9] 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 it of negative pairs [9], [10]. 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 [10], 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 [10].

[1]Z. Lai, J. Chauhan, L. C. Oliveira, and C.-N. Chuah are with the Department of Electrical and Computer Engineering, University of California Davis, Davis, CA 95616 USA. {lzhengfeng, jhichauhan, lcernyo, chuah}@ucdavis.edu

[2]Z. Li is with the Department of Computer Science, University of California Davis, Davis, CA 95616 USA. pipli@ucdavis.edu

[3]B. N. Dugger is with the Department of Pathology and Laboratory Medicine, University of California Davis, Sacramento, CA 95817 USA. bndugger@ucdavis.edu

However, CoOp requires substantial computing resources and the results of CoOp are not interpretable. Furthermore, CoOp faces performance degradation when there is a significant domain shift, e.g., from natural images to pathology images, making it hard to be adapted to medical imaging tasks. On the other hand, medical image analysis may involve multiple learning tasks simultaneously. It is impractical if the fine-tuning requires heavy computational resources when the learning tasks are separated from the analysis objective. As shown in Figure 1, multi-tasks need to be performed for the final analysis of the digital pathology domain: this may involve segmentation, classification, and/or detection tasks. For example, a recent work [2] quantifies and visualizes each type of pathology in grey and white matter regions in gigapixel images, respectively. To achieve such an objective, we need to separate it into two learning tasks: pathology detection and grey/white matter (GM/WM) segmentation. They collected and annotated two datasets, then trained two models from scratch to achieve it. Therefore, it will be extremely beneficial if we can propose an effective way to adapt a big model into multiple downstream tasks quickly so the final analysis objective can be reached in a reasonable time manner.

In this work, we aim to fine-tune CLIP efficiently with light computing resources for pathology image analysis tasks. There are several challenges in fine-tuning CLIP. First, overfitting is a severe issue if we adapt CLIP to the downstream tasks directly since CLIP is pre-trained on a 400M dataset while the new domain dataset can be small [11]. Second, CLIP has shown its promising zero-shot generalization ability, directly fine-tuning it may lose it and face the "catastrophic forgetting" issue. In other words, it is unclear how to effectively learn the new knowledge while retaining the original pre-trained knowledge so that the generalization of CLIP can be maintained. 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 public pathology datasets and benchmark the zero-shot ability of CLIP on them. Then we did empirical experiments to verify the importance of the CLIP's pre-trained knowledge. To address the "catastrophic forgetting" issue, we propose Residual Feature Connection (RFC) as a lightweight approach for adapting CLIP to pathology images. In RFC, we design a self-adaptive module to automatically fuse/balance the original knowledge from CLIP and the new knowledge learned from the new pathological task with only a few additional trainable weights instead of optimizing the entire encoders in CLIP. To further alleviate the overfitting, we also incorporate Hidden Representation Perturbation (HRP) and Dual-view Vision Contrastive (DVC). Lastly, we build an end-to-end adaptation of CLIP with a proposed loss, Doublet Multimodal Contrastive Loss (DMCL), to pathology tasks, named Path-CLIP.

We summarize our contributions as follows:

- We explore the applicability of CLIP on pathology images and highlight the catastrophic forgetting issues when adapting CLIP to downstream pathology tasks
- We propose Path-CLIP to adapt CLIP to pathology tasks with limited data. In Path-CLIP, we propose Residual Feature Connection (RFC) with Hidden Representation Perturbation (HRP) and a self-adaptive ratio to alleviate the overfitting and catastrophic forgetting issues. We also propose Dual-view Vision Contrastive (DVC) and Doublet Multimodal Contrastive Loss (DMCL) to fine-tune CLIP as an end-to-end pipeline.
- Extensive experiments demonstrate the efficiency and robustness of Path-CLIP. It has the potential to quickly adapt pre-trained CLIP to multi-tasks with competitive performance and light computational cost.

II. Related Work

A. Language-vision model

Language-vision models have shown promising performance in learning generic visual representations [9], [12]–[14]. Recent models advance via text representation learning with large-scale Transformers [15] and even web-scale training datasets [10]. Transformer-based multimodal learning has achieved great success in such large-scale datasets [16], [17]. For example, CLIP [9] was trained on 400 million image-caption pairs and achieved state-of-the-art results in many fields [9], [18]–[20]. In CLIP's architecture, it has two encoders: the vision encoder can be ResNet [21] or ViT [22] while the text encoder can be Transformer (e.g., BERT [23]). A recent work [24] fine-tuned CLIP for video data and performed competitively to more complicated methods designed for video processing. PointCLIP [25] applied CLIP to 3D recognition. CLIP is also applied for image generation tasks [26] and shows its power to reduce data collection. However, the best way to adapt CLIP for downstream tasks is still being studied, especially when the new domain is far from the pre-trained domains, e.g., the medical domain.

B. Language-vision training in medical domain

Language-vision pre-training typically requires near-infinite web images and captions from general domains, e.g., CLIP uses 400 million image-caption pairs [9]. Such datasets dwarf the scale of medical datasets. Another challenge is the annotation cost: medical images typically require domain knowledge from trained personnel [4]. This makes the training cost-expensive if we have multiple medical tasks. For example, Lai *et al.* studies the distribution of Amyloid-β plaques (a hallmark pathology with Alzheimer's disease) in grey and white matter, which involves two learning tasks: image segmentation and object detection. These two tasks require two different datasets with their required annotation to train two models [2].

Although expert annotation of medical images can be expensive, preparing the captions and prompts for images will exacerbate the situation if we use language-vision training. For example, MedCLIP [27] used 570 thousands of image-text pairs to achieve 60% zero-shot accuracy. Datasets with such scale are challenging to curate in pathology image tasks, where each slide is at gigapixel level. The efforts to adapt CLIP for the pathology domain remain limited due to the large gap

between the general and medical image domains. In this work, we aim to design an efficient adaption framework that can be scalable for multiple downstream tasks.

C. CLIP Fine-tuning and adaptation

Deep learning methods are data-hungry, which can prevent their wide adoption in real-world scenarios. Pre-training and fine-tuning offer a promising direction for adapting the model to downstream tasks. Although recent multimodal models (e.g., CLIP [9]) perform astonishing standard vision tasks without fine-tuning, the gap between CLIP and supervised learning is still noticeable. Therefore, many works focus on fine-tuning CLIP with a few samples from the new task provided. For example, CoOp [10] kept the entire pre-trained parameters fixed but optimized the continuous prompt through a learnable vector. This method closes the gap between CLIP and supervised learning but the performance gain is limited as it only focuses on the perspective of prompt design instead of the vision part for the image recognition tasks. On the other hand, CLIP-Adapter [5] and Tip-Adapter [28] freeze the pre-trained weights and train additive module, which significantly improves the CLIP's performance with limited training data in the new task. These methods require a lower computational cost. However, both CLIP-Adapter [5] and Tip-Adapter [28] involve tuning a ratio to balance the original and fine-tuned knowledge manually, which may not be practical when we have multiple target learning tasks in the medical domain. Therefore, in this work, we propose a scalable approach to automate this process.

III. Methodology

In this section, we introduce the proposed Path-CLIP. We summarize the end-to-end pipeline in Fig. 2: a self-adaptive residual ratio α will be incorporated into Residual Feature Connection (Section III-B) to fuse and balance the original knowledge with the task-specific knowledge from the new downstream task. We introduce Hidden Representation Perturbation (HRP) and Dual-view Vision Contrastive (Section III-C) to alleviate the overfitting issue. Finally, we propose Doublet Multimodal Contrastive Loss (DMCL) in Section III-D as the loss function for fine-tuning.

A. Language-Vision Pre-training

We first revisit language-vision pre-training and use CLIP [9] as one example. Note that our framework is applicable to broader language-vision models. CLIP [9] 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 vision encoder can be a CNN like ResNet [21] or a Vision Transformer (ViT) [22]. The text encoder is built on Transformer [15] and generates text embeddings from the prompt. Specifically, "prompt" here refers to a sequence of words (tokens), such as "a neuropathological image of grey matter".

Training. During training, CLIP jointly trains $F(\cdot)$ and $G(\cdot)$ to optimize the similarity score (e.g., symmetric cross-entropy loss [29]) between the visual and textual embeddings for each batch. The input consists of an image and its corresponding prompt (e.g., "this is a photo of grey matter"). They formulate the learning objective as a contrastive loss to align the visual and textual embedding spaces. Given a batch of image-prompt pairs, CLIP maximizes the similarity score for positive pairs while minimizing it for negative pairs in that batch. To learn diverse and wide visual concepts, CLIP's team collects a dataset of 400 million image-prompt pairs to achieve better transferability to downstream tasks.

Contrastive loss function. In CLIP training, the agreement between a positive pair of vision and text embeddings is maximized by InfoNCE loss [30]:

$$l_i^{v\to t} = -\log \frac{\exp\left(\mathbf{sim}(v_i, t_i)/\gamma\right)}{\sum_{k=1}^{B} \exp\left(\mathbf{sim}(v_i, t_k)/\gamma\right)}, \tag{1}$$

where v, t are normalized vectors from vision and text encoders. (v_i, t_i) is a positive pair. **sim** is a function to measure the similarity of two vectors. B is the mini-batch size of the image-text pairs and γ is the learnable temperature. Therefore, there is only 1 positive text for the input ith image but $N-1$ negative texts. $l_i^{v\to t}$ means the InfoNCE loss from an image i to texts, so as the text loss $l_i^{t\to v}$. Then the loss function for CLIP is as follows:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2B} \sum_{i=1}^{B} \left(l_i^{v\to t} + l_i^{t\to v}\right). \tag{2}$$

Inference for classification tasks. For the inference process, an image x 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:

$$p_i = \frac{\exp(\mathbf{W_i}^T \times f)/\gamma}{\sum_{i=1}^{K} \exp(\mathbf{W_i}^T \times f)/\gamma}, \tag{3}$$

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

Transferability. Compared to traditional classifier learning methods where the supervision is a fixed one-hot vector, language-vision pre-training utilizes an open-set visual concept that can be explored through the text encoder to learn a broader semantic space, which in turn makes the learned representations more diverse and adaptable to various downstream tasks (e.g., zero-shot learning). Therefore, we argue that such language-vision pre-training is helpful for medical problems where the domain is very distinctive from the natural images.

B. 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. 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

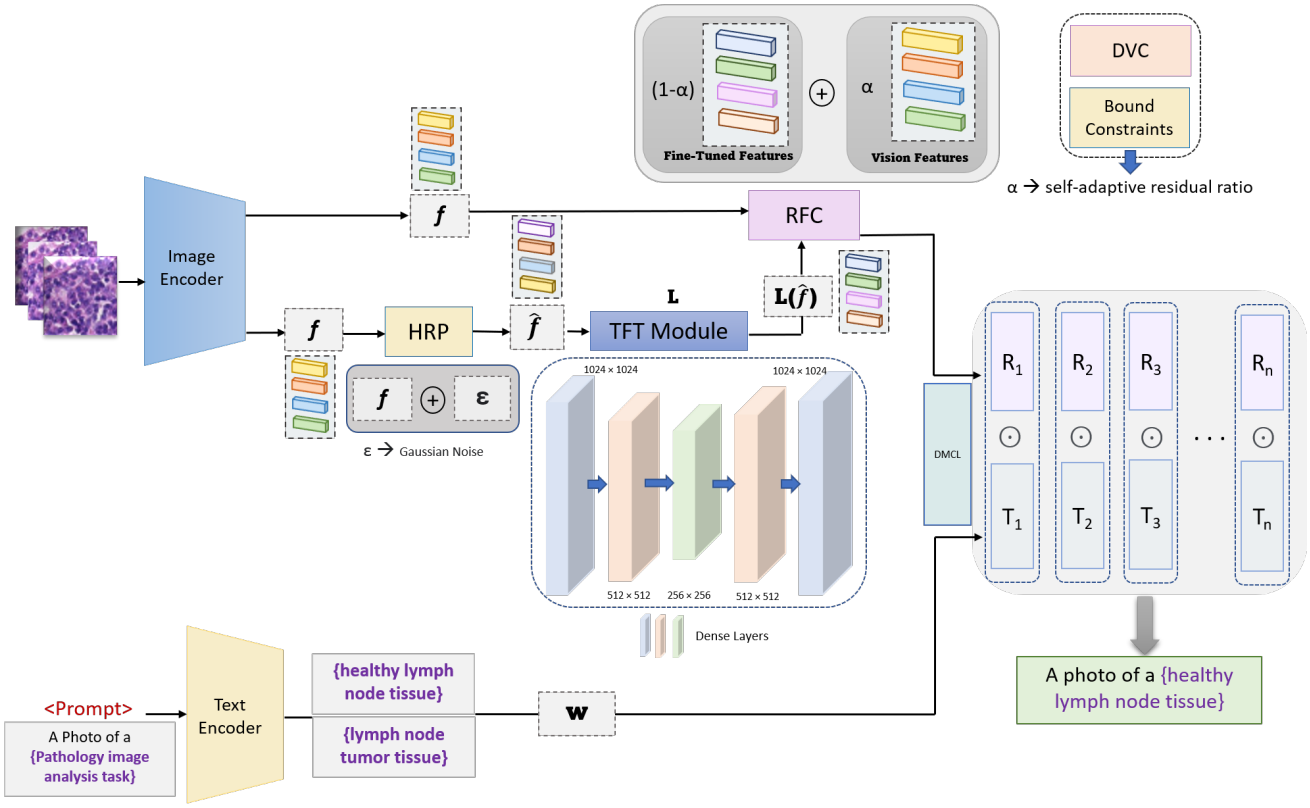


Fig. 2. Overview of Path-CLIP: the image and text encoders are frozen while RFC (Residual Feature Connection) is a downsampling-upsampling architecture of fully connected layers. RFC blends the fine-tuned knowledge with the original knowledge from CLIP's vision encoder ($F(\cdot)$) via a self-adaptive residual ratio α. HRP refers to Hidden Representation Perturbation.

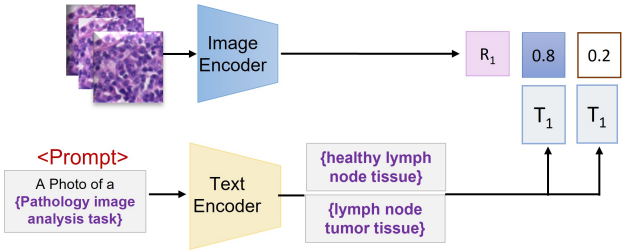


Fig. 3. Zero-shot pathology image classification.

visual features f. However, simple fine-tuning of the entire network may fail in a new pathological task due to 1) the overfitting issue caused by a large amount of parameters in CLIP and a shortage of training samples [31]; 2) catastrophic forgetting issue.

Importance of the CLIP's pre-trained knowledge. CLIP is pre-trained on 400 million of image-text pairs and has shown its zero-shot ability and transferability to a wide range of downstream tasks. We argue preserving pre-trained knowledge of CLIP is crucial to maintain its generalization and alleviate the overfitting issue when the new dataset is small. There are two reasons behind this. First, one recent NLP work [32] shows large pre-trained language models (PLMs) are well-calibrated on the masked language modeling tasks. However, such property cannot be retained through direct fine-tuning due to catastrophic forgetting. Specifically, the pre-trained features are only used for the initialization and will be distorted by the fine-tuning [32]. Inspired by this, we investigate whether this issue also appears in language-vision models. As shown in Table I, "no fine-tuning" but zero-shot gets the best performance in these datasets compared to fin-tune CLIP on 1% of training data in PCam [33]. We also attempt to lower the learning rate to 1×10^{-6}, but the results are still far from satisfactory. Therefore, we conclude direct fine-tuning will distort the pre-trained knowledge and result in a severe catastrophic forgetting issue.

Besides the catastrophic forgetting issue, overfitting is another common challenge when adapting a big model to a new task. Recent work shows that CLIP's original knowledge is the key to preserving its generalization [9]. Therefore, to alleviate the above two issues, we propose a residual connection architecture to preserve the pre-trained features and acquire rich semantic information for pathology image analysis.

TABLE I
BASELINE. WE FINE-TUNE CLIP (RESNET-50) ON PCAM [33] WITH 1% TRAINING DATA. "NO FINE-TUNING" IS THE BEST BASELINE DUE TO THE CATASTROPHIC FORGETTING ISSUE.

#	Configuration	Accuracy
1	No fine-tuning	56.5%
2	Fine-tune 10 epochs	40.3%
3	Fine-tune 100 epochs	35.9%
4	With a lower learning rate on #3	45.6%

Architecture overview. Inspired by [32] that preserving the pre-trained knowledge help calibrate the language models in NLP tasks, we argue the importance of retaining the features from CLIP for computer vision tasks 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 adaptive 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". This is one example: the layers are flexible with other dimensions. After this architecture, $L(f)$ can be of the same size of f and blended with f as follows:

$$f_i{}^* = (1 - \alpha_i)L(f_i) + \alpha_i f_i, \tag{4}$$

where, $i = \{1, 2, 3, ..n\}$ and α is the self-adaptive residual ratio to balance the fine-tuned knowledge and the original CLIP's knowledge. Specifically, the first term adaptively learns knowledge from the new dataset, while the second term preserves the original knowledge from CLIP.

Then we adopt Eq. (3) 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 [9]. The prompt can be "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 [33]. We can incorporate pathological descriptions into the prompt to provide more domain information, e.g., "this is a **pathological breast** photo of []".

Hidden Representation Perturbation (HRP). When we fine-tune a pre-trained language-vision model on a small dataset, overfitting or representation collapse can be a serious issue [34]. To alleviate this issue, we propose Hidden Representation Perturbation (HRP) in this subsection to complement the trainable layers. We select a "four-layer" architecture with HRP instead of more profound architectures so that the number of trainable parameters can remain small. More trainable parameters may tend to overfit the downstream dataset. This issue can be aggravated in limited-resource scenarios, e.g., only a few data samples are available in the new task [35].

To alleviate the overfitting issue in low-resource settings, we focus on adding noise during the training process, a widely-used NLP strategy for large language models to smoothen the optimization landscape. For example, NoisyTune [36] adds the perturbations to the pre-trained weights to improve the model's generalization. Inspired by these NLP strategies [34], [36], we propose to add random noise to the feature f before trainable adapting layers fine-tuning it. The random noise can be Gaussian noise ε generated from $\mathcal{N}(0, \sigma^2)$. After this, the perturbed feature f is converted into $\hat{f} = f + \varepsilon$. Therefore, Eq. (4) is converted into:

$$f_i^* = (1 - \alpha_i)L(\hat{f}_i) + \alpha_i f_i. \tag{5}$$

Therefore, the residual feature that must be fine-tuned will be perturbed with the Gaussian noise, designed to alleviate the overfitting issue.

C. Self-adaptive residual ratio

In Section III-B, we introduce residual feature connection and the residual ratio α. α helps modify the degree of preserving the original knowledge for better performance. Empirically, α should be small if the domain gap between the source data and the downstream task is enormous, so more knowledge from the downstream task ($L(f)$) can be considered. However, if α is too small, the model can be overfitting as it will heavily rely on the fine-tuned features and disregard the pre-trained knowledge from CLIP. On the other hand, the model can also be underfitting if α is too large, as it may prevent learning from the fine-tuned features. Therefore, it is crucial to quantify the residual ratio α so the model can be well-trained instead of underfitting or overfitting.

Recent works such as Tip-Adapter [28] and CLIP-Adapter [5] involve feature connection while they set such ratio as a fixed hyper-parameter that needs to be tuned during the training process. We argue the residual ratio should be dynamic and self-adaptive during training instead of fixed, as the learning status changes along the training process. The learning status is the key to measuring the model's fit for the new tasks. Besides, when we have multiple downstream tasks, it is time-consuming and computation-inefficient to manually tune the hyper-parameter for every single task. Hence we propose a self-adaptive residual ratio in this subsection.

DVC: Dual-view Vision Contrastive. Inspired by self-supervised learning (e.g., SimCLR [37]), the composition of data augmentations is relevant to define effective predictive tasks. Therefore, we propose Dual-view Vision Contrastive (DVC) to measure the model's learning status. Given an image x, we leverage two types of augmentations: "weak" and "strong" [38]. "weak" refers to the standard flip-and-shift augmentation strategy while "strong" refers to perturb the appearance (RandAugment [39] and CTAugment [40]). We propose to measure the distance between representations from both augmentations, e.g., cosine similarity has shown good performance in both visual contrastive [37] and textual contrastive [41] tasks.

With the input x, we will get $w(x)$ from the weak augmentation and $s(x)$ from the strong augmentation. We obtain $F(w(x))$ and $F(s(x))$ after the vision encoder F. We define the distance as τ:

$$\tau = \frac{1}{N}\sum_{i=1}^{N} cos(F(w(x_i)), F(s(x_i))), \tag{6}$$

where N is the number of samples in the training set, and cos is a cosine similarity. The model's learning status is relevant to its consistency across different perturbations. We traverse the entire training set and calculate its mean value to represent the model's performance.

Bound constraints. In addition to DVC that measures the predictive consistency of the model, we also select "accuracy"

on the training set as a second metric to measure the learning status, denoted as T. Empirically, "accuracy" alone cannot fully tell if the model is well-fit: 100% of accuracy may indicate overfitting, which is the biggest challenge when we adapt a large model to a downstream task where the training set is very small. However, "accuracy" is a good metric to show if the model is underfitting. Therefore, we propose to fuse DVC and "accuracy" to represent the model's status during training.

If we look at α and set it fixed, we can find the model is prone to overfit if α is too small as it will heavily rely on the fine-tuned knowledge; the model tends to underfit if α is too large as it prevents the model from learning new knowledge (directly use the original pre-trained knowledge). Therefore, we propose a dynamic way with lower and upper bounds to obtain α during training:

$$\alpha = \min(\max(\tau, T), 1-\tau). \tag{7}$$

The bounds are associated with predictive consistency (DVC) and predictive accuracy on the training set. For example, if the model is overfitting, T will be large and τ will be relatively small. Then α will be smaller in the next epoch to use less fine-tuned features. Therefore, our proposed bound constraints can effectively adjust α to alleviate the overfitting issue.

Algorithm 1 Path-CLIP: Adapt CLIP for pathology tasks.

Input: N: size of the dataset, $F(\cdot)$: image encoder, $G(\cdot)$: text encoder, x: input image, t: input text, $w(\cdot)$: weak augmentation, $s(\cdot)$: strong augmentation.

Freeze $F(\cdot)$ and $G(\cdot)$
for epoch $= 1, 2, \ldots$ **do**
  for image $= x$ **do**
    Extract image features: $f = F(x)$
    Generate the Gaussian noise ϵ
    Perturbed the image features: $\hat{f} = f + \epsilon$
    Fine-tune $\hat{f}$ with TFT module: $L(\hat{f})$
    for $i = 1, 2, ..$ **do**
      Compute self-adaptive residual ratio from (7): $\alpha_i = \min(\max(\tau, T), 1-\tau)$, where, T→Accuracy, and $\tau = \frac{1}{N}\sum_{i=1}^{N} cos(F(w(x_i)), F(s(x_i)))$
      calculate $f_i^* = (1-\alpha_i)L(\hat{f}_i) + \alpha_i f_i$
    end for
  end for
  Perform text embedding: $g = G(t)$
  Define the contrastive loss l_i^c from (8)
  Optimize (9) for Doublet Multimodal Contrastive Loss (DMCL):
    $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{SWAPP}}$, where,
    $\mathcal{L}_{\text{CLIP}} \rightarrow \frac{1}{2B}\sum_{i=1}^{B}\left(l_i^{v\rightarrow t} + l_i^{t\rightarrow v}\right)$,
    $\mathcal{L}_{\text{SWAPP}} \rightarrow \frac{1}{2B}\sum_{i=1}^{2B} l_i^c$, B is the mini-batch size of the image-text pair.
end for

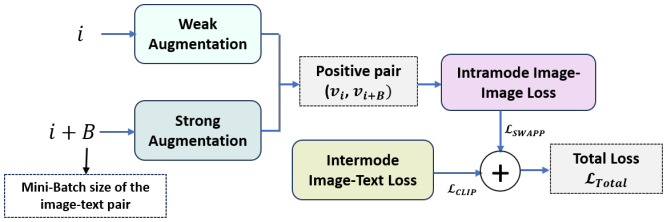


Fig. 4. Doublet Multimodal Contrastive Loss (DMCL) that utilizes the traditional image-text pair and the proposed image-image pair with the strong and weak augmentation criteria.

D. Path-CLIP: End-to-end adaptation of CLIP to pathology tasks

To efficiently adapt CLIP to pathology image tasks, we design a residual feature connection with a self-adaptive residual ratio to balance the original and fine-tuned knowledge based on the learning status. To alleviate the overfitting issue, we incorporate Hidden Representation Perturbation and design bound constraints for the residual ratio. Besides these, to distill more visual representations from the new image domain, we add a contrastive loss to the loss function so that the trainable portion has a better classifying ability. We follow SimCLR [37] to define the Contrastive Loss (CL) for one image in the batch as:

$$l_i^c = \begin{cases} -\log \frac{\exp(\textbf{sim}(v_i, v_{i+B})/\gamma)}{\sum_{k\neq i}^{2B}\exp(\textbf{sim}(v_i, v_k)/\gamma)} & \text{, if } i < B, \\ -\log \frac{\exp(\textbf{sim}(v_i, v_{B-i})/\gamma)}{\sum_{k\neq i}^{2B}\exp(\textbf{sim}(v_i, v_k)/\gamma)} & \text{, otherwise.} \end{cases} \tag{8}$$

Differing from SimCLR [37], as shown in Fig. 4, here we set image i as the weakly augmented image and image $i+B$ as the strongly augmented image instead of two randomly augmented images as the positive pair. We named this loss as $\mathcal{L}_{\textbf{SWAPP}}$. **SWAPP** here refers to **S**trong-**W**eak **A**ugmented **P**ositive **P**air. Finally, the total loss for fine-tuning CLIP on our pathology datasets is as follows:

$$\begin{aligned} &\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{SWAPP}}. \\ &\mathcal{L}_{\text{CLIP}} \rightarrow \frac{1}{2B}\sum_{i=1}^{B}\left(l_i^{v\rightarrow t} + l_i^{t\rightarrow v}\right), \\ &\mathcal{L}_{\text{SWAPP}} \rightarrow \frac{1}{2B}\sum_{i=1}^{2B} l_i^c \end{aligned} \tag{9}$$

We define this loss as **Doublet Multimodal Contrastive Loss (DMCL)** because the items involve both inter-contrastive (image and text) and intra-contrastive (image and image). The entire end-to-end pipeline is summarized in Algorithm 1.

IV. EXPERIMENTS

A. Datasets and setup

Pathology image classification tasks. The datasets used to validate our framework are collected from two distinct pathology projects [33], [45]. Both projects make their data available as patches, which are extracted from H&E stained Whole Slide Images (WSI) digitized from Formalin-Fixed Paraffin-Embedded (FFPE) slides. Each dataset aims to detect different cancerous tissue.

TABLE II
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 [37]	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 [42]	55.7 ± 2.05	59.2 ± 2.23	54.5 ± 1.97	56.8 ± 1.55	0.58 ± 0.03
		FixMatch [38]	73.2 ± 0.76	77.5 ± 0.50	73.7 ± 0.25	75.6 ± 0.37	0.84 ± 0.04
		Dash [43]	71.0 ± 0.98	75.3 ± 0.78	70.2 ± 0.45	72.7 ± 0.50	0.82 ± 0.02
		FlexMatch [44]	74.0 ± 0.80	78.1 ± 1.15	73.0 ± 0.90	75.5 ± 0.84	0.85 ± 0.03
Multimodal CLIP	-	Zero-shot [9]	56.5 ± 0.00	57.4 ± 0.00	50.3 ± 0.00	53.7 ± 0.00	0.60 ± 0.00
	0.5%	CoOp [10]	63.6 ± 0.25	63.9 ± 0.42	62.5 ± 0.35	63.0 ± 0.29	0.67 ± 0.02
		CLIP-Adapter [5]	72.3 ± 1.02	77.2 ± 0.95	63.2 ± 0.56	69.4 ± 0.84	0.81 ± 0.02
		TIP-Adapter [28]	73.9 ± 0.50	70.5 ± 0.85	78.0 ± 0.49	74.0 ± 0.95	0.84 ± 0.02
		Path-CLIP	**81.9 ± 0.78**	**79.4 ± 0.35**	**85.0 ± 0.80**	**82.1 ± 0.95**	**0.89 ± 0.02**
Supervised	1.5%	-	62.8 ± 1.31	69.5 ± 0.67	54.1 ± 2.27	60.8 ± 1.68	0.63 ± 0.02
Self-supervised	1.5%	SimCLR [37]	64.9 ± 1.02	65.2 ± 1.45	60.1 ± 1.70	62.5 ± 1.05	0.64 ± 0.03
Semi-supervised	1.5%	Pseudo-Label [42]	67.9 ± 1.85	70.5 ± 1.90	62.2 ± 1.76	66.1 ± 1.42	0.67 ± 0.04
		FixMatch [38]	83.9 ± 1.29	85.1 ± 0.85	84.1 ± 0.95	84.6 ± 0.90	0.87 ± 0.01
		Dash [43]	82.8 ± 1.09	83.0 ± 0.55	81.2 ± 0.60	82.1 ± 0.78	0.85 ± 0.03
		FlexMatch [44]	84.1 ± 1.02	84.9 ± 0.75	84.4 ± 0.80	84.6 ± 0.75	0.86 ± 0.04
Multimodal CLIP	1.5%	CoOp [10]	61.5 ± 0.75	60.5 ± 1.03	65.6 ± 0.90	63.1 ± 0.85	0.65 ± 0.02
		CLIP-Adapter [5]	74.1 ± 0.88	79.1 ± 0.34	65.4 ± 0.89	71.6 ± 1.02	0.82 ± 0.01
		TIP-Adapter [28]	75.2 ± 0.94	72.3 ± 0.69	79.4 ± 0.86	75.7 ± 0.73	0.83 ± 0.02
		Path-CLIP	**83.1 ± 0.67**	**82.8 ± 0.95**	**82.0 ± 0.75**	**82.4 ± 0.46**	**0.91 ± 0.02**

PatchCamelyon (PCam) [33]. PCam is a patch dataset extracted from Camelyon16 challenge [46]. 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× resolution but later downsampled to 10×. The WSIs were collected from two different centers. PCam extracted 327,680 patches at 96×96 resolution and labeled them 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.

MHIST [45]. MHIST contains 3,152 patches from colorectal regions at 224×224 pixel resolution. These patches were extracted from 328 WSIs scanned at 40× resolution. Each patch may be labeled Hyperplastic Polyp (HP) or Sessile Serrated Adenoma (SSA). HP is the majority class with 68.59% of labels. Labeling colorectal polyps between HP and SSA is challenging due to high inter-pathologist disagreement. Seven pathologists contributed to the ground truth to ensure reliable labels.

GM/WM segmentation task [2]. Besides the two pathology classification tasks above, we also evaluate our proposed approach on a grey/white matter segmentation task in pathological slides. For this dataset, we have 30 gigapixel slides with an average resolution of $50,000 \times 60,000$ by an Aperio AT2 scanner at 20X magnification. They were collected and annotated by a trained researcher and expert at the University of California Davis, Alzheimer's Disease Research Center (UCD-ADRC). We follow the previous work [2] to split 20 slides into the training/validation set while 10 slides into the hold-out test set. We tile these gigapixel slides into patches at 256×256 and convert it into a weakly-supervised task [47].

Experimental setup. For a fair comparison, we use ResNet-50 [21] as the backbone in the vision encoder $f(\cdot)$ and BERT [23] as the textual encoder. We follow CLIP [9] to use gradient scaling to facilitate mixed-precision training. We set the learning rate as 0.0001. The batch size is 32. We use Adam [48] as the optimizer. To compare the computational complexity, we conduct all experiments on a single GPU (Nvidia RTX 2080Ti).

Baselines. We first implement traditional supervised learning (SL) methods with the entire dataset to explore the upper-bound performance for each task. Then we reduce the data usage and investigate the performance change of these SL methods. Then we compare Path-CLIP with original CLIP [9] and three recent CLIP-based fine-tuning methods: CoOp [10], Tip-Adapter [28] and CLIP-Adapter [5]. We compare them under the limited data scenarios to test their ability to do downstream tasks when the new dataset is small. Besides that, we also incorporate recent self-supervised learning (SimCLR [37]) and semi-supervised learning (Pseudo-Label [42], FixMatch [38], Dash [43], and FlexMatch [44]). The semi-supervised training also requires a large amount of unlabeled data, which involves data collection efforts. We compare our proposed CLIP-based method with semi-supervised learning to show the advantages of reducing data collection efforts.

B. Main results

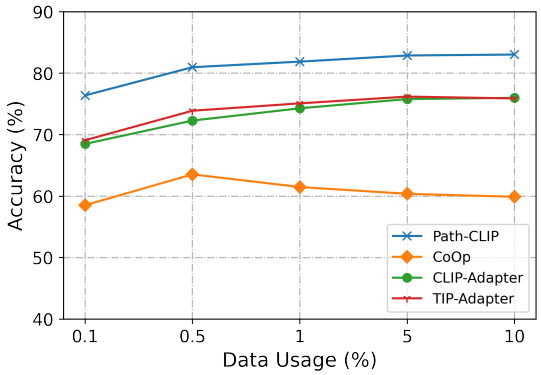
Classification task: PCam [33] and MHIST [45]. The main results for two classification benchmarks are summarized in Table II and Table III. First, CLIP shows its promising zero-shot ability: when we directly test CLIP on PCam [33], it can get 56.5% accuracy. Then, we compare Path-CLIP with other CLIP-based fine-tuning methods. In PCam [33],

TABLE III
QUANTITATIVE COMPARISON ON MHIST. PRECISION, RECALL, AND F1-SCORE REFER TO THE MACRO-AVERAGED VALUES FROM ALL CLASSES

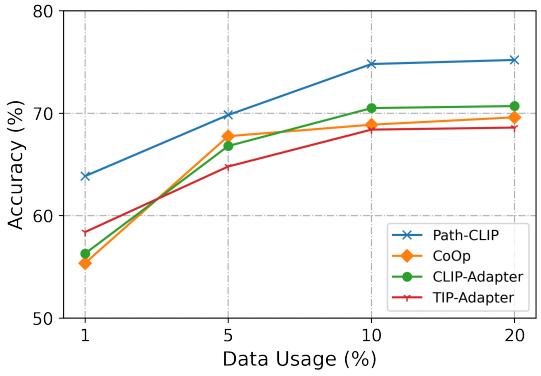
Learning Manner	Labeled Ratio	Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Supervised	100%	-	86.0 ± 0.36	84.8 ± 0.61	83.5 ± 0.55	84.1 ± 0.58	0.89 ± 0.01
Self-supervised	10%	SimCLR [37]	63.2 ± 0.95	63.8 ± 1.20	62.2 ± 1.70	63.0 ± 0.85	0.66 ± 0.04
Semi-supervised	10%	Pseudo-Label [42]	65.4 ± 1.04	66.5 ± 0.98	65.9 ± 0.85	66.2 ± 1.42	0.75 ± 0.03
		FixMatch [38]	70.5 ± 1.55	68.8 ± 2.76	69.5 ± 2.97	69.1 ± 2.86	0.79 ± 0.02
		Dash [43]	70.2 ± 1.21	70.0 ± 0.85	71.3 ± 0.69	70.6 ± 1.48	0.81 ± 0.04
		FlexMatch [44]	73.1 ± 0.84	74.6 ± 1.05	75.1 ± 0.90	74.8 ± 0.89	0.82 ± 0.03
Multimodal CLIP	10%	Zero-shot [9]	36.9 ± 0.00	36.9 ± 0.00	100 ± 0.00	53.9 ± 0.00	0.50 ± 0.00
		CoOp [10]	68.9 ± 0.40	56.0 ± 0.65	72.8 ± 0.53	63.3 ± 0.86	0.76 ± 0.01
		CLIP-Adapter [5]	65.9 ± 1.45	52.2 ± 1.23	87.2 ± 0.85	65.3 ± 0.95	0.77 ± 0.02
		TIP-Adapter [28]	63.1 ± 0.78	49.9 ± 0.90	94.7 ± 1.02	65.4 ± 0.56	0.78 ± 0.03
		Path-CLIP	**74.8 ± 1.95**	**63.3 ± 1.40**	**75.6 ± 0.98**	**68.9 ± 1.50**	**0.84 ± 0.02**

we present 0.5% and 1.5% data usage in training set to fine-tune CLIP, respectively. We find all of these fine-tuning algorithms [5], [10], [28] can improve the performance of the original CLIP with the new data from the downstream task presented. Our proposed Path-CLIP can outperform them in both settings. Specifically, Path-CLIP outperforms TIP-Adapter [28] with 8% and 7.9% accuracy when only 0.5% and 1.5% data are used, respectively. Path-CLIP also shows similar improvement in MHIST [45] (Table III). Since MHIST [45] is small, we use 10% of data for fine-tuning. On the other hand, we implemented several recent semi-supervised learning where the rest of the data in the training set are used as an unlabeled set. Although Path-CLIP does not use unlabeled data, it can achieve competitive results with semi-supervised learning. This indicates the potential of Path-CLIP to reduce the efforts of data collection.

Besides these ratios of data usage, we also visualize the results by varying the ratio to show the performance of each fine-tuning algorithm. As shown in Fig. 5, CoOp [10] may not bring improvement to PCam when more data are used to fine-tune the prompt. CoOp freezes the encoders and only fine-tunes the language prompt part, while other algorithms also fine-tune and vision part. We conclude that visual fine-tuning is more critical compared to language fine-tuning, which is also reported in [5]. In Fig. 5, we also show that our proposed CLIP-Path can consistently outperform the other three recent fine-tuning methods when we vary the data usage.



(a) PCam



(b) MHIST

Fig. 5. Comparison of different fine-tuning methods by varying data usage for the adaptation.

TABLE IV
PERFORMANCE COMPARISON ON GM/WM SEGMENTATION.

Data Usage	Algorithm	GM_IoU	GM_DICE	WM_IoU	WM_DICE
100% (SL)	FCN [49]	71.72	83.21	49.09	66.08
	U-Net [50]	91.52	95.17	82.02	88.15
0.1% (SL)	FCN [49]	48.25	25.01	24.54	30.02
	U-Net [50]	46.13	50.10	5.75	12.23
0.1% (CLIP)	CoOp [10]	62.23	65.70	39.92	48.65
	CLIP-Adapter [5]	67.80	74.32	43.25	54.61
	TIP-Adapter [28]	69.92	75.50	46.23	56.60
	Path-CLIP	**78.05**	**83.70**	**54.50**	**63.23**
1% (SL)	FCN [49]	51.34	30.20	31.02	35.12
	U-Net [50]	50.24	53.23	15.43	25.58
1% (CLIP)	CoOp [10]	63.45	64.12	37.85	45.34
	CLIP-Adapter [5]	69.99	76.54	47.95	58.34
	TIP-Adapter [28]	72.12	77.30	55.65	63.60
	Path-CLIP	**80.21**	**88.70**	**60.50**	**70.34**

Segmentation task: GM/WM segmentation [2]. Besides two classification tasks, we follow a recent work [2] to generalize our method to a weakly-supervised segmentation problem. We select IoU and DICE scores as the measuring metrics for the grey matter (GM) and white matter (WM) regions. The results are summarized in Table IV. We select two supervised learning (SL) approaches for the segmentation task: FCN [49] and U-Net [50]: we first use all data in the training set (100%) to train them in a supervised manner, then we reduce the dataset to 0.1% and 1% setting and train the model again. We can see that both SL algorithms face performance degradation when the training data is scarce. They require an extensive

and well-curated labeled dataset to get satisfactory results. We then focus on multi-modal training and fine-tuning CLIP with recent algorithms. Table IV shows Path-CLIP can achieve superior results among other fine-tuning techniques.

TABLE V
VISION VS PROMPT FINE-TUNING ON PCAM.

Data Usage	Algorithm	Accuracy	Training Time
Zero-shot	CLIP [9]	56.5	-
0.1%	CoOp [10]	58.5	**7 min 6 sec**
	Path-CLIP	**76.4**	10 min 29 sec
1%	CoOp [10]	61.5	53 min 21 sec
	Path-CLIP	**81.9**	**11 min 56 sec**
10%	CoOp [10]	59.9	2 h 23 min 45 sec
	Path-CLIP	**83.0**	**27 min 18 sec**

C. Ablation studies

(a) CLIP

(b) CLIP with RFC

Fig. 6. Visualization of different learned feature manifolds via t-SNE.

1) Vision vs prompt fine-tuning: CoOp [10] is the recent state-of-the-art fine-tuning method for CLIP. Hence, we mainly compare our proposed RFC with its performance and computational complexity. We summarize the results in Table V. When we range the data usage from 0.1% to 1%, we reveal Path-CLIP can obtain 5.5% improvement in accuracy while only introducing 1 minute of additional training time. However, CoOp [10] has the overfitting issue and obtains a lower score while using 53 minutes, which is almost five times compared to the time used by RFC. The situation remains similar under the 10% of data usage. We conclude 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 V demonstrates results revealing vision fine-tuning for a medical domain may be more effective than prompt fine-tuning.

TABLE VI
ABLATION STUDY ON PCAM USING 0.5% DATA FOR FINE-TUNING.

#	RFC	HRP	DVC	DMCL	Accuracy	F1-score	AUC
1	✗	✗	✗	✗	56.5	53.7	0.60
2	✓	✗	✗	✗	72.0	70.5	0.69
3	✓	✓	✗	✗	74.2	75.5	0.78
4	✓	✓	✓	✗	78.5	79.3	0.84
5	✓	✓	✓	✓	81.9	82.1	0.89

2) Effect of Residual Feature Connection: To show the effect of the Residual Feature Connection (RFC), we use t-SNE [51] to visualize the manifold of the original CLIP and CLIP with RFC after training it on PCam [33]. The t-SNE results are presented in Fig. 6: it illustrates RFC can show clear separations of image features belonging to different classes in the high-dimensional classification space. In addition, as shown in Table VI, RFC with the self-adaptive ratio can significantly improve with +15.5% accuracy.

D. Importance of alleviating overfitting

Since the new downstream task may have a very limited training set, it is crucial to alleviate the potential overfitting issue. Besides the RFC with the self-adaptive ratio, we also propose HRP (Hidden Representation Perturbation), DVC (Dual-view Vision Contrastive), and DMCL (Doublet Multimodal Contrastive Loss) to cope with the overfitting issue. Table VI summarizes the contribution of each proposed module: we find HRP can improve with +2.2% accuracy and DVC can improve with +4.3%. DMCL inspired by self-supervised learning can also aid in great improvements. We conclude contrastive learning and perturbations on data and hidden representations are helpful to the overfitting issue when the new dataset is small.

V. DISCUSSION

Automated analysis of the pathology image task is an important aspect in providing scalable objective evaluations of medical images. Limited data availability limits the intensive end-to-end training of deep networks. Also, the requirement of multiple learning tasks for the complete pathology image analysis is time-consuming and can be resource-hungry. In this work, we explore the generalization of CLIP in pathology image classification. The Contrastive Language-Image Pre-training (CLIP) model has shown powerful capabilities in diverse downstream applications. However, its applicability in pathology image analysis with limited labeled data is still under study due to the giant domain shift and overfitting issues. We propose Path-CLIP to efficiently fine-tune CLIP image encoder that extracts the rich semantic information using small datasets and light computing resources.

We introduce the self-adaptive residual ratio α into Residual Feature Connection (RFC), which changes dynamically during training and overcomes the overfitting issue. The ablation

studies show the importance of self-adaptive residual ratio, an improvement of 15.5% and 16.8% is observed in accuracy and F1 score, respectively. Additionally, other modules of the image encoder of the vision-language pre-trained model significantly contribute to improving the prediction ability of the model and address the overfitting issue as well. The combination of these modules resulting in an overall improvement of 9.9% in accuracy over RFC. The results reveal the superiority and effectiveness of the proposed Path-CLIP in comparison to the existing fine-tuning methods, with respect to the evaluation criteria considered. It is interesting to note the Path-CLIP outperforms the traditional CLIP by 19.9% when using only 0.1% of the labeled data in PCam dataset, in addition, the less computational time compliments the efficiency of the method. Hence, we conclude Path-CLIP is more robust and has the potential to bridge the domain shift between the pre-trained natural images and pathology images in an efficient and cost-effective manner.

ACKNOWLEDGMENTS

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] Z. Lai, C. Wang, Z. Hu, B. N. Dugger, S.-C. Cheung, and C.-N. 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.

[2] Z. Lai, L. C. Oliveira, R. Guo, W. Xu, Z. Hu, K. Mifflin, C. Decarli, S.-C. Cheung, C.-N. Chuah, and B. N. Dugger, "Brainsec: Automated brain tissue segmentation pipeline for scalable neuropathological analysis," *IEEE Access*, vol. 10, pp. 49 064–49 079, 2022.

[3] W. Zhang, L. Zhu, J. Hallinan, S. Zhang, A. Makmur, Q. Cai, and B. C. Ooi, "Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation," in *CVPR*, 2022, pp. 20 666–20 676.

[4] Z. Lai, C. Wang, L. C. Oliveira, B. N. Dugger, S.-C. Cheung, and C.-N. Chuah, "Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling," in *ICCV Workshop*, 2021, pp. 591–600.

[5] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.

[6] M. Liu, L. Hu, Y. Tang, C. Wang, Y. He, C. Zeng, K. Lin, Z. He, and W. Huo, "A deep learning method for breast cancer classification in the pathology images," *IEEE J. Biomed. Health. Inf.*, vol. 26, no. 10, pp. 5025–5032, 2022.

[7] J. Yang, H. Chen, Y. Liang, J. Huang, L. He, and J. Yao, "Concl: Concept contrastive learning for dense prediction pre-training in pathology images," in *ECCV*. Springer, 2022, pp. 523–539.

[8] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, and X. Han, "Transpath: Transformer-based self-supervised learning for histopathological image classification," in *MICCAI*. Springer, 2021, pp. 186–195.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[11] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang, "Domain adaptation via prompt learning," *arXiv preprint arXiv:2202.06687*, 2022.

[12] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

[13] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *CVPR*, 2022, pp. 18 123–18 133.

[14] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *CVPR*, 2022, pp. 5206–5215.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[16] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *arXiv preprint arXiv:2206.06488*, 2022.

[17] Y. Tian, S. Newsam, and K. 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*, 2023, pp. 1011–1021.

[18] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 12 104–12 113.

[19] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[20] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] H. Rasheed, M. U. khattak, M. Maaz, S. Khan, and F. S. Khan, "Finetuned clip models are efficient video learners," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[25] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8552–8562.

[26] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

[27] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," *Proceedings of EMNLP*, 2022.

[28] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *ECCV*. Springer, 2022, pp. 493–510.

[29] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *ICCV*, 2019.

[30] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[31] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *ICML*. PMLR, 2019, pp. 2790–2799.

[32] G. He, J. Chen, and J. Zhu, "Preserving pre-trained features helps calibrate fine-tuned language models," in *ICLR*, 2023.

[33] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *MICCAI*. Springer, 2018, pp. 210–218.

[34] H. Yuan, Z. Yuan, C. Tan, F. Huang, and S. Huang, "Hype: Better pre-trained language model fine-tuning with hidden representation perturbation," *arXiv preprint arXiv:2212.08853*, 2022.

[35] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting few-sample bert fine-tuning," in *ICLR*, 2021.

[36] C. Wu, F. Wu, T. Qi, and Y. Huang, "Noisytune: A little noise can help you finetune pretrained language models better," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, pp. 680–685.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[38] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, vol. 33, 2020, pp. 596–608.

[39] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshop*, 2020, pp. 702–703.

[40] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *NeurIPS*, vol. 32, 2019.

[41] F. Chen, H. Zhang, Z. Li, J. Dou, S. Mo, H. Chen, Y. Zhang, U. Ahmed, C. Zhu, and M. Savvides, "Unitail: Detecting, reading, and matching in retail scene," in *ECCV*. Springer, 2022, pp. 705–722.

[42] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning (ICML)*, 2013.

[43] Y. Xu, L. Shang, J. Ye, Q. Qian, Y.-F. Li, B. Sun, H. Li, and R. Jin, "Dash: Semi-supervised learning with dynamic thresholding," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 525–11 536.

[44] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *NeurIPS*, vol. 34, 2021.

[45] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, N. Tomita, L. Torresani *et al.*, "A petri dish for histopathology image analysis," in *Int. Conf. Arti. Intell. in Medi.* Springer, 2021, pp. 11–24.

[46] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[47] D. Liu, Y. Cui, L. Yan, C. Mousas, B. Yang, and Y. Chen, "Densernet: Weakly supervised visual localization using multi-scale feature aggregation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 6101–6109.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[49] P. Bándi, R. van de Loo, M. Intezar, D. Geijs, F. Ciompi, B. van Ginneken, J. van der Laak, and G. Litjens, "Comparison of different methods for tissue segmentation in histopathological whole-slide images," in *Proc. 2017 IEEE Int. Symp. Biomed. Imaging*, 2017, pp. 591–595.

[50] K. R. Oskal, M. Risdal, E. A. Janssen, E. S. Undersrud, and T. O. Gulsrud, "A U-net based approach to epidermal tissue segmentation in whole slide histopathological images," *SN Appl. Sci.*, vol. 1, no. 7, pp. 1–12, 2019.

[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

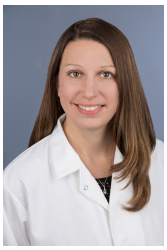
VI. Biography Section

Zhengfeng Lai received his B.S. in Information Engineering from Zhejiang University in 2019, and his M.S. degree in Electrical and Computer Engineering from the University of California, Davis in 2021. He is currently pursuing his Ph.D. degree at the University of California, Davis. He has served as a reviewer for the IEEE Transactions on Image Processing, Medical Physics, ICML, NeurIPS, CVPR, ICCV, and their workshops.

Joohi Chauhan is currently a postdoctoral researcher in Electrical and Computer Engineering Department, University of California, Davis. She received the B.Tech. and M.Tech. degrees in computer science and the Ph.D. degree from IIT Ropar, India, in 2021. She was also selected as a Newton Bhabha Ph.D. Placement fellow, 2019-20. She works in an interdisciplinary domain and socially impactful problems and her research interests include applied deep learning, image processing, and healthcare apps and analytics. She has served as a reviewer for the IEEE Access, Pattern Recognition, Expert System with Application, Artificial Intelligence in Medicine, ICCV, MICCAI, and CVPR workshops.

Zhuoheng Li is currently in his third year of pursuing a Bachelor's degree in Computer Science at the University of California, Davis. His research interests include Computer Vision, Multimodal Learning, and Semi-Supervised Learning.

Luca Cerny Oliveira received his B.S. in Electrical Engineering and Computer Engineering from University of California, Davis in 2021. He is currently a graduate student at the University of California, Davis.



Brittany N. Dugger is an associate professor in the Department of Pathology and Laboratory Medicine at the University of California, Davis (UCD) and co-leader of the neuropathology core of the UCD Alzheimer's Disease Research Center. She received her B.S. From Michigan State University and Ph.D. from the Mayo Clinic and completed a postdoctoral fellowship at the Banner Sun Health Research Institute. Her research interests include understanding the heterogeneity within neurodegenerative diseases, understanding the interaction of peripheral changes to aging and neurodegenerative diseases, and creating easy to use tools to aid in scalable deeper disease phenotyping.

Chen-Nee Chuah (M'01-SM'06-F'15) is currently the Child Family Professor in Engineering in the Electrical and Computer Engineering Department, University of California, Davis. She received the B.S. degree in Electrical Engineering from Rutgers University, and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Sciences from the University of California, Berkeley. Her research interests include Internet measurements, cyber security, and applying data science and intelligent learning techniques to societal-scale networked systems and applications, including smart health domain and intelligent transportation systems. She is a Fellow of the IEEE and an ACM Distinguished Scientist. She has served as an Associate Editor for the IEEE/ACM Transactions on Networking and the IEEE Transactions on Mobile Computing.