Co-HSF: Resource-Efficient One-Shot Semi-Supervised Adaptation of Histopathology Foundation Models

Luca Cerny Oliveira[1], Kartik Patwari[1], Xiaoguang Zhu[1], Sen-Ching Cheung[2], Brittany N. Dugger[3], Chen-Nee Chuah[1]

[1]Department of Electrical and Computer Engineering, University of California Davis, Davis, CA, USA
[2]Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY, USA
[3]Department of Pathology and Laboratory Medicine, University of California Davis, Sacramento, CA, USA
lcernyo@ucdavis.edu, kpatwari@ucdavis.edu, xgzhu@ucdavis.edu, sccheung@ieee.org, bndugger@ucdavis.edu, chuah@ucdavis.edu

Abstract

Automated analysis of histopathological samples has greatly augmented the ability of experts to perform deep phenotyping on biological samples. Current state-of-the-art (SOTA) methods for histopathology image classification rely on training deep neural networks with large annotated datasets, which can be costly to obtain. Recent studies propose to bypass annotated datasets by leveraging pre-trained foundation models (e.g. visual-language models) for zero-shot predictions. Moreover, fine-tuning these models enhances performance while requiring minimal labeled data (e.g. one-shot fine-tuning). However, one-shot fine-tuned performance of histopathology foundation models on image classification tasks is understudied. In this work, we first explore the use of semi-supervised few-shot learning (SSFSL) for fine-tuning histopathology foundation models on one-shot datasets with unlabeled samples. We find SOTA SSFSL methods improve fine-tuning performance, but their pseudo-labeling (i.e. assigning labels to unlabeled samples) strategies can increase inference times over zero-shot. We then propose a Co-filtered Histopathology Semi-Supervised Few-Shot (Co-HSF) pipeline: a dual-SSFSL (i.e. with teacher and student models) training loop followed by a co-filtering (CF) pseudo-labeling strategy to efficiently leverage unlabeled data for improved semi-supervised performance and reduced inference times. Using the National Center for Tumor Disease Colorectal Cancer Dataset (NCT-CRC-HE), we show our proposed module achieves 38.4% improvement in accuracy over zero-shot performance with only 9 labeled samples and over 53% faster inference times, while also outperforming other fine-tuning and SSFSL methods.

Introduction

Deep learning applied to histopathology images has enhanced digital histopathology analysis, augmenting the expert's deep phenotyping workflow through automated, scalable quantitative analysis of tissue (Scalco et al. 2024). For example, Jiang *et al.* achieved state-of-the-art brain tumor detection in stimulated Raman histology images with a fully supervised learning model trained on 229,000 annotated samples (Jiang et al. 2022). Additional examples can be found in diverse histopathological domains such as breast cancer (Tsiknakis et al. 2023), lymphoma (Xu et al. 2023), and Alzheimer's disease automated analysis (Lai et al. 2022). Despite the competitive performance reported, the aforementioned studies required significant labor-intensive data collection and expert-driven annotation efforts (Jiang et al. 2022; Tsiknakis et al. 2023; Xu et al. 2023; Lai et al. 2022). A goal of the artificial intelligence field is to generate models capable of extracting foundation feature representations that can be leveraged across different tasks with minimal or no further training. This goal is compounded when studying artificial intelligence applications in histopathology, where data collection and labeling are additionally challenging due to costly infrastructure (Scalco et al. 2023), the need for proper safeguards to ensure patient privacy, and the medical expertise required for annotations.

Contrastive Language-Image Pre-training (CLIP) emerged as a way to generate foundation models through pre-training on large image-caption datasets (Radford et al. 2021), displaying competitive zero-shot performance through the extraction of foundation visual representations from language supervision (Agarwal et al. 2021). However, time-consuming manual design of prompts is needed to achieve highest performance (Lai et al. 2023; Lozano et al. 2024), even in CLIP models pre-trained with histopathology-specific data such as CONCH (Lu et al. 2024). Additionally, due to the performance gap between zero-shot and state-of-the-art supervised learning approaches, there is a preference to fine-tune these models with small labeled sets (Huang et al. 2023; Lu et al. 2024; Lai et al. 2023). Studies propose linear probing with small task-specific labeled sets (Huang et al. 2023; Lu et al. 2024) as these methods provide a simple and efficient fine-tuning strategy. However, linear probing does not leverage unlabeled data, limiting its use in real-world histopathology datasets where unannotated data is often plentiful and can increase overall classification performance. Other alternative fine-tuning adaptation methods have been proposed through prompt-based (Zhou et al. 2022; Chen et al. 2022), adapter-based (Lai et al. 2023), and other adaptation methods (Javed et al. 2024). Although these methods can successfully adapt CLIP models to specific tasks and can leverage unlabeled data for improved performance, they exhibit significant limitations in three key areas: (1) overfitting to one-shot datasets (i.e. one labeled sample per class) (Lai et al. 2023; Zhou et al. 2022; Chen et al. 2022), (2) requiring detailed captions for

images in fine-tuning set (Mo et al. 2023; Javed et al. 2024; Mirza et al. 2024), and (3) requiring costly training of the entire CLIP architecture (Mo et al. 2023; Javed et al. 2024).

Semi-supervised few-shot learning (SSFSL) is an alternate approach to fine-tune pre-trained models with minimal labeled data. These methods leverage existing visual encoders trained on a large, annotated dataset and use metric-based models (Snell, Swersky, and Zemel 2017), pseudo-labeling techniques (i.e. assigning labels to unlabeled samples) (Huang et al. 2021; Wang et al. 2020), and other strategies to train classifiers capable of generalizing to novel downstream tasks given a small labeled dataset (Hu, Gripon, and Pateux 2021). Although these methods show promising results in natural images, their deployment in histopathology tasks remains understudied. Specifically, the pre-trained visual encoder selection criteria remain unclear in few-shot scenarios where no large, annotated dataset is available for pre-training. Additionally, these methods perform transductive pseudo-labeling, which requires substantial computing time for inference with larger unlabeled sets.

Therefore, we propose **Co-HSF** (**Co**-filtered **H**istopathology **S**emi-Supervised **F**ew-Shot), a combination of CLIP models and SSFSL for fine-tuning focused on utility (i.e. minimal labeling efforts and unlabeled data leveraging) and performance (i.e. faster inference times and lower GPU utilization). We show that the proposed framework enhances SSFSL performance through improved pseudo-labeling accuracy, while reducing inference time through inductive pseudo-labeling. We combine the CLIP-based model CONCH (Lu et al. 2024) and SSFSL due to their complementary nature: CONCH provides the histopathology-trained visual encoder needed by SSFSL, on the other hand, SSFSL combats overfitting issues when fine-tuning on small labeled datasets. Using one-shot fine-tuning, we show the proposed framework improves zero-shot inference time by over 53%, and improves zero-shot accuracy by up to 38.4%.

Our contributions can be summarized as follows:

- We combine CONCH (Lu et al. 2024) and SSFSL methods to develop a label-efficient dual-SSFSL training loop (i.e. with teacher and student models), addressing overfitting issues when fine-tuning CLIP-based models in small datasets.
- We design a Co-filtering (CF) pseudo-labeling strategy to improve semi-supervised performance during training. Additionally, we show CF's inductive inference allows for faster evaluation.
- We evaluate the proposed framework on two histopathology image classification datasets with hematoxylin and eosin (H&E)-stained colorectal (Kather, Halama, and Marx 2018) and breast tissue (Veeling et al. 2018). We show our proposed method outperforms competing SSFSL and CLIP fine-tuning methods.

Methods

Problem Formulation

We first introduce the basic notions and terminology of SSFSL problem formulation. In few-shot learning (FSL) settings, the training set includes a labeled data set X with the corresponding labels Y, sometimes referred as the support set. When utilizing unlabeled data (SSFSL setting), an additional unlabeled set U is available for pseudo-labeling. Metric-based frameworks use a pre-trained encoder $G(\cdot)$ to generate feature embeddings $G(X)$ and $G(U)$. The SSFSL frameworks train and evaluate on the embeddings generated by this pre-trained encoder rather than the images. Traditionally, $G(\cdot)$ is a deep neural network model trained on a task similar to the training set task. For our study, we evaluated SSFSL frameworks with the CLIP-based visual encoder from CONCH (Lu et al. 2024) as the pre-trained encoder $G(\cdot)$.

Algorithm 1: Dual-SSFSL Training Loop

Require: Initial labeled set X_0, unlabeled set U, encoder $G(\cdot)$, teacher M_t, student M_s, pseudo-labeling strategy $C(\cdot)$, iterations N, step s, pseudo-labeled set class-balance ratio α

1: **for** each iteration $i = 1$ to N **do**
2: Augment X_{i-1} to X_{aug} with Randaugment (Cubuk et al. 2020)
3: Featurize X_{aug} and X_{i-1} using $G(\cdot)$
4: Train M_t on $G(X_{\text{aug}})$ and M_s on $G(X_{i-1})$ to generate updated weights M_s' and M_t'
5: Featurize U and Randaugment (Cubuk et al. 2020) U_{aug} using $G(\cdot)$
6: Evaluate $G(U)$ with M_s' and $G(U_{\text{aug}})$ with M_t' for pseudo-labels $M_s'(G(U))$ and $M_t'(G(U_{\text{aug}}))$
7: Generate pseudo-labeled set $C(M_s'(G(U)), M_t'(G(U_{\text{aug}})))$ and update X with s randomly selected pseudo-labeled samples
8: **end for**
9: Featurize entire pseudo-labeled set $C(M_s'(G(U)), M_t'(G(U_{\text{aug}})))$ using $G(\cdot)$
10: Remove excess majority-class pseudo-labels to enforce class-balanced ratio of α and add to X_0 to generate X_{final}
11: Train M_s on $G(X_{\text{final}})$

Co-HSF Pipeline

The proposed Co-HSF pipeline is shown in Figure 1. Inspired by Co-teaching (Han et al. 2018), which achieves promising training results through using two models for label refinement, our pipeline emphasizes on the importance of leveraging predictions from different models trained with different data when assigning a pseudo-label. We first present the general dual-SSFSL training loop to enable teacher and student model training. Next, we introduce the Co-filtering (CF) pseudo-labeling strategy which can successfully leverage the predictions from both models for improved pseudo-labeling.

Dual-SSFSL Training Loop. The dual-SSFSL training loop used in this study is described in Algorithm 1. The proposed pipeline enables improved downstream pseudo-labeling (see Methods - CF Pseudo-labeling Section) by leveraging predictions from two distinct models (student and teacher). In each training iteration, we select the labeled data set X and generate an augmented training set X_{aug} using Randaugment transforms (Cubuk et al. 2020). Both X and X_{aug}

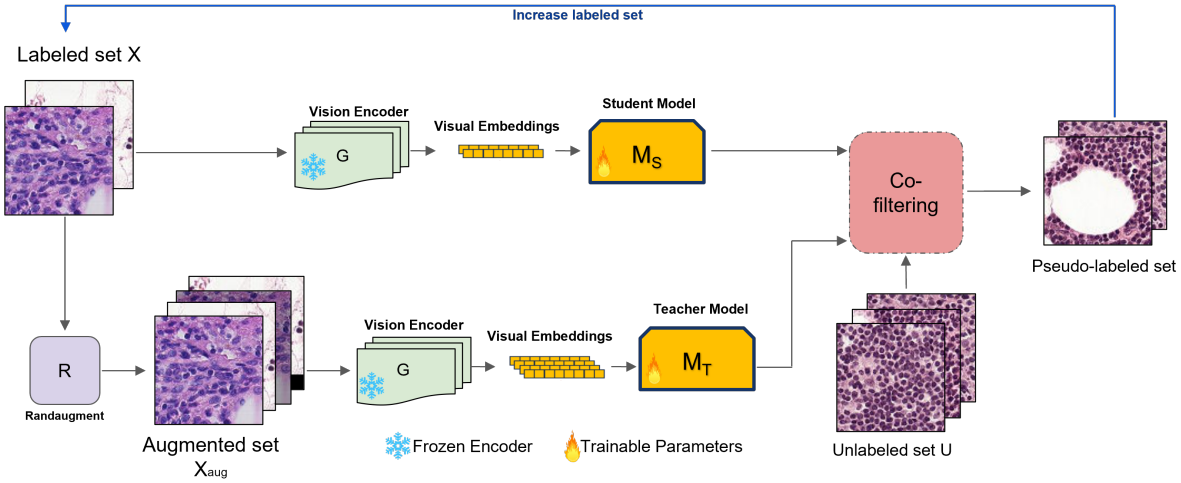


Figure 1: Overview of the proposed framework: the labeled set is augmented by Randaugment (Cubuk et al. 2020) and featurized by any CLIP-based pre-trained vision encoder $G(\cdot)$. Both the student and teacher models are trained using the CONCH(Lu et al. 2024) visual embeddings of X and X_{aug} respectively. The trained models and the unlabeled set are inputted to the Co-filtering algorithm (see Figure 2), which selects samples and pseudo-labels to be added to X for the next iteration.

are featurized using any CLIP pre-trained visual encoder $G(\cdot)$ (for this study we use CONCH (Lu et al. 2024)). The set $G(X)$ is fed to the student model M_s, and $G(X_{\text{aug}})$ is fed to the teacher model M_t. Both M_t and M_s are trained under any metric-based algorithm (for this study we use PLCM-FSL (Huang et al. 2021)) to generate updated weights M_t' and M_s'.

Sequentially, M_t' and M_s' evaluate the featurized unlabeled set $G(U)$. The student model evaluates only $G(U)$, while the teacher model evaluates an augmented training set $G(U_{\text{aug}})$, acquired through Randaugment transforms (Cubuk et al. 2020). The results are then inputted into the Co-filtering (CF) algorithm, which performs pseudo-labeling based on the student and teacher predictions on these unlabeled samples. The newly pseudo-labeled samples are then included in the labeled set X in the next training iteration. After the last iteration, the student model M_s is trained under an inductive metric-based algorithm (for this study we use class-weighted logistic probing) using the labeled set X and the added pseudo-labeled samples. Some classes may be under-represented in the generated pseudo-labels, thus we enforce a class-imbalance ratio (i.e. the ratio between the majority classes and minority class) of α, randomly removing excess majority class samples.

CF Pseudo-labeling. The CF pseudo-labeling strategy used in this study is depicted in Figure 2. Unlike traditional SSFSL pseudo-labeling, which may incorrectly label an unlabeled sample due to spurious correlations learned by a single model from a small training set, we focus on samples with consistent predictions agreed by both models: M_t' and M_s'. However, simply leveraging the predictions may overlook model uncertainty and lead to poor pseudo-labels. Therefore, we sort the prediction confidences P as follows:

$$P = \{p^{(0)} \leq p^{(1)} \leq \cdots \leq p^{(n)}\},$$

where $p^{(i)}$ is the class-prediction score for sample of the i^{th} sample in increasing order. We then establish the confidence threshold t based on hyperparameter T:

$$t = \left\lceil \left(1 - \frac{T}{100}\right)(n+1) \right\rceil,$$

such that we only assign pseudo-labels to samples in the group P' of most confident predictions:

$$P' = \{p^{(t)} \leq \cdots \leq p^{(n-1)} \leq p^{(n)}\}.$$

Additionally, samples on feature representation class-boundaries may still lead to incorrect predictions by M_t' and M_s', therefore, M_t' also generates predictions on U_{aug} samples for CF. A separate threshold t_{aug} is generated using hyperparameter T to calculate most confident predictions in U_{aug}. Similar to the FlexMatch pseudo-labeling (Zhang et al. 2021) strategy, where each class has different confidence threshold according to their class distribution, we set separate confidence thresholds t and t_{aug} for U and U_{aug} samples according to their confidence distribution. This separate threshold, based on the same hyperparameter T, allow pseudo-labeling to adapt to changes in confidence distribution caused by Randaugment transforms (Cubuk et al. 2020). Moreover, this inductive pseudo-labeling is performed during training, therefore reducing inference time.

Results

Evaluation Datasets and Setup

The datasets used for validation of our framework are acquired from two histopathology cancer studies analyzing H&E-stained tissue. Both datasets consist of images patches extracted from gigapixel Whole Slide Images (WSI) digitized tissue. All WSIs were digitized from Formalin-Fixed Paraffin-Embedded (FFPE) physical slides. For both datasets,

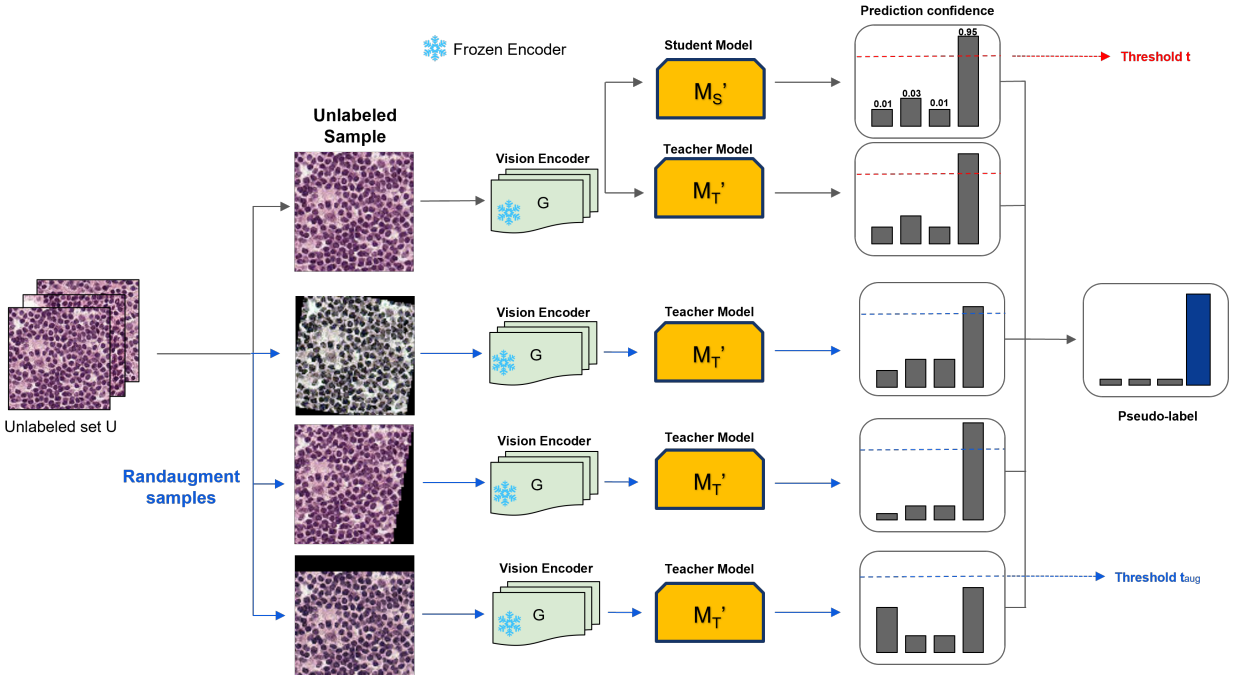


Figure 2: Overview of the proposed Co-filtering pseudo-labeling strategy: the unlabeled set U is augmented by Randaugment (Cubuk et al. 2020) and featurized by CLIP's pre-trained vision encoder $G(\cdot)$. Both student and teacher models evaluate on U and U_{aug} respectively. The prediction confidences from the evaluations are analyzed, if the prediction confidences for the evaluations of a single sample and their augmented samples are all above the thresholds t and t_{aug} (both calculated from hyperparameter T), we will add the sample to the pseudo-labeled set.

the training and evaluation sets are stratified by subject - meaning patches from a given subject are all either on training or evaluation sets. Both datasets are class-balanced. We remove labels from 10% of the training set for the unlabeled set.

PatchCamelyon (PCam) (Veeling et al. 2018). Pcam contains 327,680 patches from hematoxylin & eosin (H&E) lymph node regions at 96×96 pixel resolution. The patches were acquired from the Camelyon16 dataset (Bejnordi et al. 2017). The dataset was derived from 400 WSIs scanned from formalin-fixed paraffin-embedded (FFPE) samples from two independent datasets collected at Radboud University Medical Center and the University Medical Center Utrecht. There are two tissue classes. Each patch may be labeled as tumor positive or tumor free.

National Center for Tumor Disease Colorectal Cancer Dataset (NCT-CRC-HE) (Kather, Halama, and Marx 2018). NCT-CRC-HE contains 100,000 patches from hematoxylin & eosin (H&E) colorectal regions at 224×224 pixel resolution. These patches were extracted from 86 WSIs scanned at 0.5 microns per pixel resolution from formalin-fixed paraffin-embedded (FFPE) samples from the NCT Biobank (National Center for Tumor Diseases, Heidelberg, Germany) and the UMM pathology archive (University Medical Center Mannheim, Mannheim, Germany). All images are color-normalized using Macenko's method (Macenko et al. 2009). There are nine tissue classes. Each patch may be labeled as Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM).

Evaluation Setup. All of the computational performance experiments are conducted on one piece of GPU (NVIDIA Tesla T4) and CPU (Intel Xeon 4210) to compare the average inference time on the entire test set (repeated 5 times), and total GPU memory consumption. The experiments with linear probing, fine-tuning, FSL, and SSFSL learning types were conducted with CONCH (Lu et al. 2024) vision-language model (VLM) under one-shot setting. The zero-shot experiments were conducted with the listed VLMs. Quantitative experiments comparing fine-tuning performance were done on five different seeds, using different one-shot labeled sets. For zero-shot evaluation, we report the highest performance metrics out of six different tested prompts. Prompts include, but are not limited to: "this is a photo of []", "An H&E image of []", "An image of []", "A pathology image of []"; where "[]" is filled with the class name. For SSFSL algorithms (Huang et al. 2021; Wang et al. 2020) we evaluate both FSL (leverage query data) and SSFSL (leverage query and unlabeled data) settings. We used hyperparameters/prompts proposed by each framework when fine-tuning or generating zero-shot predictions to the evaluation datasets.

Learning Type	Algorithm	PCam		NCT	
		Accuracy (%)	F-1 Score (%)	Accuracy (%)	F-1 Score (%)
Zero-shot	CLIP	56.57	56.56	28.08	23.64
	CLIP (ViT-B-L16)	57.02	53.15	45.46	39.91
	Zhang et al. 2023	53.35	40.39	50.83	46.82
	PLIP (Huang et al. 2023)	68.91	68.76	47.06	46.70
	CONCH (Lu et al. 2024)	82.66	82.57	51.70	47.83
One-shot Linear Probing	SGD (Huang et al. 2023)	80.55±3.61	80.41±4.03	82.26±5.30	81.54±1.46
	Logistic (Lu et al. 2024)	80.75±3.51	80.53±4.25	84.10±1.38	83.49±1.19
One-shot Fine-tuning	CoOp (Zhou et al. 2022)	57.80±0.01	47.04±0.65	24.30±0.02	31.58±6.45
	CLIPath (Lai et al. 2023)	72.18±4.04	70.00±8.44	69.24±1.46	65.65±4.76
One-shot FSL	Hu et al. 2021	80.15±0.78	80.15±0.79	68.22±2.05	68.12±2.10
	Snell et al. 2017	80.75±3.51	80.53±3.66	83.29±5.18	82.19±5.92
	ICI (Wang et al. 2020)	81.69±2.58	81.69±2.59	84.40±4.16	83.20±4.77
	PLCM (Huang et al. 2021)	81.54±1.79	81.49±1.82	87.49±3.62	86.62±4.71
One-shot SSFSL	ICI (Wang et al. 2020)	81.06±2.07	80.99±2.02	87.11±4.07	86.71±4.60
	PLCM (Huang et al. 2021)	82.02±3.43	81.93±3.49	88.76±2.96	88.42±3.35
	Co-HSF (proposed)	**84.41**±**1.38**	**84.40**±**1.38**	**90.10**±**0.52**	**89.82**±**1.86**

Table 1: Quantitative comparison on PCam and NCT-CRC-HE (NCT). CLIP refers to OpenAI CLIP (Radford et al. 2021).

Performance on Histopathology Tasks

We evaluate the proposed semi-supervised few-shot learning (SSFSL) framework on two H&E-stained histopathology datasets (PCam (Veeling et al. 2018) and NCT-CRC-HE (Kather, Halama, and Marx 2018)) under one-shot scenarios. Table 1 presents classification accuracies for zero-shot and four fine-tuning approaches: linear probing (Lu et al. 2024; Huang et al. 2023), prompt (Zhou et al. 2022) and adapter-based (Lai et al. 2023) fine-tuning, conventional few-shot learning (FSL) (Hu, Gripon, and Pateux 2021; Snell, Swersky, and Zemel 2017; Wang et al. 2020; Huang et al. 2021), and SSFSL (Huang et al. 2021; Wang et al. 2020). We first observe CONCH shows its strong zero-shot capabilities, often outperforming fine-tuning methods, especially in the simpler binary classification task in PCam (healthy vs. tumor). Next, we observe our proposed method consistently outperforms both zero-shot and competing fine-tuning approaches in one-shot settings. While prompt-based and adapter-based fine-tuning methods report promising results when moderate amounts of labeled data are available (Lai et al. 2023; Zhou et al. 2022), their performance degrades under one-shot conditions due to overfitting (e.g., CLIPath reports 69.24% accuracy on NCT-CRC-HE). By contrast, SSFSL/FSL approaches often surpass linear probing and other fine-tuning methods, highlighting the advantages of combining CLIP-based visual features with SSFSL/FSL pipelines under extreme low-data constraints.

On the binary classification task of PCam (i.e. healthy vs tumor), our framework achieves 84.41% ± 1.38 accuracy, surpassing the strongest zero-shot baseline (CONCH (Lu et al. 2024)) by over 1.7%. SSFSL baselines such as PLCM-SSFSL (Huang et al. 2021) are slightly below our approach by more than 2% in accuracy. Fine-tuning performance gains were notably higher in tasks with additional complexity such as the NCT-CRC-HE classification task, where there are nine classes, and two of those classes are different types of tumors. In that setting, our method attains 90.10% ± 0.52 accuracy. This marks a 38.4% improvement over the best zero-shot baseline, a much higher increase than in simpler binary classification. Additionally, our method shows a 1.34% edge over PLCM-SSFSL, the second highest fine-tuning baseline. Although competing SSFSL shows strong overall fine-tuning performance - approaching our proposed method's performance, they have larger inference times (see Results - Resource Utilization Section). Moreover, we observe our method displays smaller variability (i.e. lower standard deviations score) compared to baselines, which we attribute to the accurate proposed pseudo-labeling (see Results - CF Pseudo-labeling Section), which can mitigate the variability from learning from only one labeled sample per class or an inaccurate pseudo-labeled set.

Algorithm	Pseudo-label performance	
	Accuracy (%)	F-1 Score (%)
PLCM	88.04±1.53	84.46±1.74
ICI	89.33±0.89	86.03±1.69
Co-HSF (proposed)	**99.96**±**0.10**	**99.20**±**1.60**

Table 2: Quantitative comparison on pseudo-label performance for NCT-CRC-HE

CF Pseudo-labeling

The proposed pipeline leverages unlabeled data via an inductive CF pseudo-labeling strategy that utilizes predictions from two models. As reported in Table 1, the resulting accuracy gains over the best SSFSL baselines range from 2.39% on PCam to 1.34% on NCT-CRC-HE. This improvement stems from CF's higher pseudo-label accuracy compared to other algorithms (see Table 2 and Table 3). For instance, CF achieves over 3% higher pseudo-label accuracy than ICI-SSFSL on PCam. We observe even larger benefits in NCT-CRC-HE pseudo-labeling, where CF achieves more than 10% higher pseudo-label accuracy and over 13% higher F-1 score compared to ICI-SSFSL. Despite high mean accuracy and F-1

Algorithm	Pseudo-label performance	
	Accuracy (%)	F-1 Score (%)
PLCM	80.17±2.56	80.01±2.58
ICI	96.03±7.94	**96.04±7.92**
Co-HSF (proposed)	**99.05±0.54**	92.60±10.80

Table 3: Quantitative comparison on pseudo-label performance for PCam

Algorithm	Inference Performance		
	Min	Acc(%)	VRAM(MiB)
CONCH	5.20±0.01	51.70	2222
CLIPath	9.22±0.07	69.24±1.46	2222
Linear Probing	**2.35±0.01**	84.10±1.38	952
ICI	11.86±0.14	87.11±4.07	952
PLCM	2.60±0.01	88.76±2.96	952
Co-HSF	2.44±0.01	**90.10±0.52**	**952**

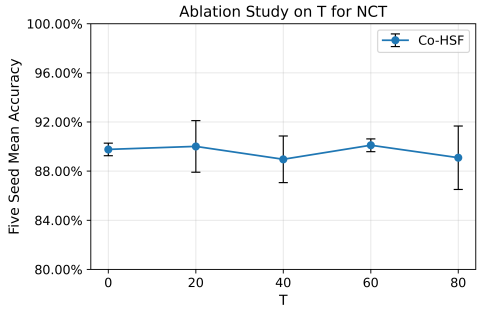
Table 4: Quantitative comparison on test set inference time in minutes (Min), accuracy (Acc), and model VRAM utilization on NCT-CRC-HE dataset

scores, a class-balancing step after pseudo-labeling is recommended as Co-HSF can produce class-imbalanced pseudo-label sets (see higher standard deviation in F-1 scores in Table 1 from one seed where pseudo-label set generated for healthy class was small and imprecise).

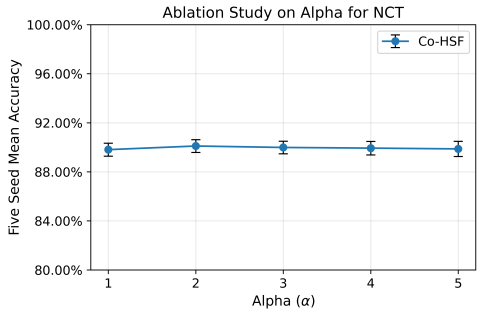
Resource Utilization

Table 4 shows that Co-HSF processes all 7180 NCT-CRC-HE test set samples in 2.44 $\pm$ 0.01 minutes on average, significantly faster than SSFSL pipelines such as ICI-SSFSL (Wang et al. 2020) (11.86 $\pm$ 0.14 minutes), and CONCH's zero-shot baseline (5.20 $\pm$ 0.01 minutes). Additionally, our method is 6.15% faster than PLCM-SSFSL (Huang et al. 2021) while showing improved accuracy. The speed-up over other SSFSL arises because no additional pseudo-labeling or query data leveraging happens during inference. Moreover, our proposed method is faster than zero-shot as it does not require tokenization or text encoding during inference, allowing for larger inference batches. Our proposed method presents 3.67% slower inference times than the linear probing, however it has 6% increase in accuracy.

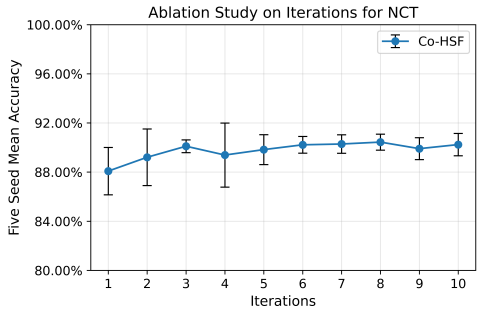
In addition to fast inference, our method demonstrates lower memory usage than compared baselines (see Table 4). Its VRAM usage peaks at 952MiB—comparable to PLCM but less than half that of CLIPath (2222MiB). This reduced memory footprint stems from not requiring tokenizers or text encoders during inference. Together, the efficient time and memory profile makes the proposed method well-suited for clinical or research environments where edge-device deployment with limited compute resources is critical.



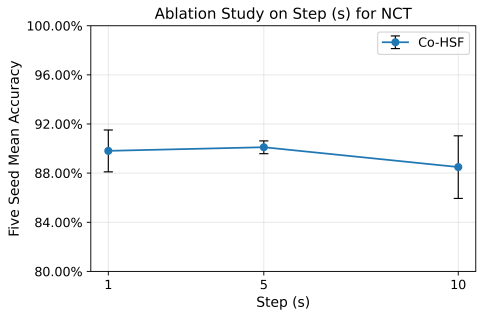
(a) Ablations study on NCT-CRC-HE (NCT) performance different T values.



(b) Ablations study on NCT-CRC-HE (NCT) performance different α values.



(c) Ablations study on NCT-CRC-HE (NCT) performance different amounts of iterations.



(d) Ablations study on NCT-CRC-HE (NCT) performance different step values.

Figure 3: Ablation studies showing the effect of different hyperparameter configurations on NCT-CRC-HE dataset. Panels (a)–(d) illustrate studies on T, α, iterations, and step hyperparameters respectively.

Ablation Study

Figure 3 shows the ablation study on all Co-HSF hyperparameters. We perform our ablation studies on the NCT-CRC-HE dataset, the most challenging evaluation dataset in this study with nine different classes (including two different tumor classes). First, we observe the different hyperparameters have little effect on the overall accuracy, showing Co-HSF's ability to perform well with minimal tuning. Figure 3a shows T has a minor effect on overall accuracy, and our study selected $T = 60$ due to highest mean accuracy. Figure 3b shows the metric-based learning is fairly robust to class-imbalance on pseudo-labeled set, given α has little effect on accuracy. Figure 3c shows Co-HSF has benefits from using three pseudo-labeling iterations, but has diminished returns from additional increases in iteration count. Lastly, Figure 3d shows the mean averages and standard deviation for different step values are stable as mean accuracy values range from 90.10% to 88.49%. We choose step $s = 5$ as it displayed the highest mean accuracy.

Conclusion

Overall, these experiments show the proposed Co-HSF pipeline, incorporating a novel Co-filtering (CF) mechanism through dual-SSFSL training loop, addresses important challenges in histopathology image classification when labeled samples and computational resources are limited. Through rigorous one-shot evaluations on two histopathology datasets, we show accuracy and inference time improvements over existing baselines. The robustness of the proposed method against overfitting can be attributed to an effective teacher-student pipeline capable of leveraging strong feature representations from CONCH (Lu et al. 2024) during training and pseudo-labeling with the proposed CF strategy.

In conclusion, our results show the promise of the proposed framework and CLIP-SSFSL ensemble approaches for real-world computational histopathology settings, where labeling costs are high but unlabeled image repositories are abundant. Additionally, our method is better suited for deployability (i.e. using edge-devices) due to its lower GPU utilization and inference times (see Table 4). However, our study only evaluates on two datasets, and do not study scenarios such as class-imbalanced unlabeled sets. Additionally, our study does not discuss selection criteria for one-shot labeled dataset. In the future, we aim to test the proposed method over additional, diverse histopathology datasets and evaluation scenarios.

Acknowledgments

This project was made possible by a grant from the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R01AG062517 and U24NS133949, Noyce Initiative UC Partnerships in Computational Transformation Program, NIH-National Institute On Aging awards #2R01-AG0652517, a 2023-2024 UC Davis Chancellor's Fellowship and Child Family Endowed Professorship Funds.

References

Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. In *The Eleventh International Conference on Learning Representations*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Hu, Y.; Gripon, V.; and Pateux, S. 2021. Leveraging the feature distribution in transfer-based few-shot learning. In *International Conference on Artificial Neural Networks*, 487–499. Springer.

Huang, K.; Geng, J.; Jiang, W.; Deng, X.; and Xu, Z. 2021. Pseudo-loss confidence metric for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8671–8680.

Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T. J.; and Zou, J. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9): 2307–2316.

Javed, S.; Mahmood, A.; Ganapathi, I. I.; Dharejo, F. A.; Werghi, N.; and Bennamoun, M. 2024. CPLIP: Zero-Shot Learning for Histopathology with Comprehensive Vision-Language Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11450–11459.

Jiang, C.; Chowdury, A.; Hou, X.; Kondepudi, A.; Freudiger, C.; Conway, K.; Camelo-Piragua, S.; Orringer, D.; Lee, H.; and Hollon, T. 2022. OpenSRH: optimizing brain tumor surgery using intraoperative stimulated Raman histology. *Advances in neural information processing systems*, 35: 28502–28516.

Kather, J. N.; Halama, N.; and Marx, A. 2018. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281.

Lai, Z.; Li, Z.; Oliveira, L. C.; Chauhan, J.; Dugger, B. N.; and Chuah, C.-N. 2023. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2374–2380.

Lai, Z.; Oliveira, L. C.; Guo, R.; Xu, W.; Hu, Z.; Mifflin, K.; Decarli, C.; Cheung, S.-C.; Chuah, C.-N.; and Dugger, B. N. 2022. BrainSec: Automated Brain Tissue Segmentation Pipeline for Scalable Neuropathological Analysis. *IEEE Access*, 10: 49064–49079.

Lozano, A.; Nirschl, J.; Burgess, J.; Gupte, S. R.; Zhang, Y.; Unell, A.; and Yeung-Levy, S. 2024. {\mu}-Bench: A Vision-Language Benchmark for Microscopy Understanding. *arXiv preprint arXiv:2407.01791*.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3): 863–874.

Macenko, M.; Niethammer, M.; Marron, J. S.; Borland, D.; Woosley, J. T.; Guan, X.; Schmitt, C.; and Thomas, N. E. 2009. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, 1107–1110. IEEE.

Mirza, M. J.; Karlinsky, L.; Lin, W.; Possegger, H.; Kozinski, M.; Feris, R.; and Bischof, H. 2024. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. *Advances in Neural Information Processing Systems*, 36.

Mo, S.; Kim, M.; Lee, K.; and Shin, J. 2023. S-clip: Semi-supervised vision-language learning using few specialist captions. *Advances in Neural Information Processing Systems*, 36: 61187–61212.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Scalco, R.; Hamsafar, Y.; White III, C. L.; Schneider, J. A.; Reichard, R. R.; Prokop, S.; Perrin, R. J.; Nelson, P. T.; Mooney, S.; Lieberman, A. P.; et al. 2023. The status of digital pathology and associated infrastructure within Alzheimer's Disease Centers. *Journal of Neuropathology & Experimental Neurology*, 82(3): 202–211.

Scalco, R.; Oliveira, L. C.; Lai, Z.; Harvey, D. J.; Abujamil, L.; DeCarli, C.; Jin, L.-W.; Chuah, C.-N.; and Dugger, B. N. 2024. Machine learning quantification of Amyloid-β deposits in the temporal lobe of 131 brain bank cases. *Acta neuropathologica communications*, 12(1): 134.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Tsiknakis, N.; Tzoras, E.; Zerdes, I.; Manikis, G. C.; Acs, B.; Hartman, J.; Hatschek, T.; Foukakis, T.; and Marias, K. 2023. Multiresolution Self-Supervised Feature Integration via Attention Multiple Instance Learning for Histopathology Analysis. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1–4. IEEE.

Veeling, B. S.; Linmans, J.; Winkens, J.; Cohen, T.; and Welling, M. 2018. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, 210–218. Springer.

Wang, Y.; Xu, C.; Liu, C.; Zhang, L.; and Fu, Y. 2020. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12836–12845.

Xu, J.; Xin, J.; Shi, P.; Wu, J.; Cao, Z.; Feng, X.; and Zheng, N. 2023. Lymphoma Recognition in Histology Image of Gastric Mucosal Biopsy with Prototype Learning. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1–4. IEEE.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.