Prompt-MIL: Boosting Multi-Instance Learning Schemes via Task-specific Prompt Tuning

Jingwei Zhang[1], Saarthak Kapse[1], Ke Ma[2], Prateek Prasanna[1], Joel Saltz[1], Maria Vakalopoulou[3], and Dimitris Samaras[1]

[1] Stony Brook University, USA
[2] Snap Inc., USA
[3] CentraleSupélec, University of Paris-Saclay, France
{jingwezhang, kemma, samaras}@cs.stonybrook.edu
{saarthak.kapse, prateek.prasanna}@stonybrook.edu
Joel.Saltz@stonybrookmedicine.edu
maria.vakalopoulou@centralesupelec.fr

Abstract. Whole slide image (WSI) classification is a critical task in computational pathology, requiring the processing of gigapixel-sized images, which is challenging for current deep-learning methods. Current state of the art methods are based on multi-instance learning schemes (MIL), which usually rely on pretrained features to represent the instances. Due to the lack of task-specific annotated data, these features are either obtained from well-established backbones on natural images, or, more recently from self-supervised models pretrained on histopathology. However, both approaches yield task-agnostic features, resulting in performance loss compared to the appropriate task-related supervision, if available. In this paper, we show that when task-specific annotations are limited, we can inject such supervision into downstream task training, to reduce the gap between fully task-tuned and task agnostic features. We propose Prompt-MIL, an MIL framework that integrates prompts into WSI classification. Prompt-MIL adopts a prompt tuning mechanism, where only a small fraction of parameters calibrates the pretrained features to encode task-specific information, rather than the conventional full fine-tuning approaches. Extensive experiments on three WSI datasets, TCGA-BRCA, TCGA-CRC, and BRIGHT, demonstrate the superiority of Prompt-MIL over conventional MIL methods, achieving a relative improvement of 1.49%-4.03% in accuracy and 0.25%-8.97% in AUROC while using fewer than 0.3% additional parameters. Compared to conventional full fine-tuning approaches, we fine-tune less than 1.3% of the parameters, yet achieve a relative improvement of 1.29%-13.61% in accuracy and 3.22%-27.18% in AUROC and reduce GPU memory consumption by 38%-45% while training 21%-27% faster.

Keywords: Whole slide image classification · Multiple instance learning · Prompt tuning.

1 Introduction

Whole slide image (WSI) classification is a critical task in computational pathology enabling disease diagnosis and subtyping using automatic tools. Owing to the paucity of patch-level annotations, multiple instance learning (MIL) [9,18,24] techniques have become a staple in WSI classification. Under an MIL scheme, WSIs are divided into tissue patches or instances, and a feature extractor is used to generate features for each instance. These features are then aggregated using different pooling or attention-based operators to provide a WSI-level prediction. ImageNet pretrained networks have been widely used as MIL feature extractors. More recently, self-supervised learning (SSL), using a large amount of unlabeled histopathology data, has become quite popular for WSI classification [13,5] as it outperforms ImageNet feature encoders.

Most existing MIL methods do not fine-tune their feature extractor together with their classification task; this stems from the requirement for far larger GPU memory than is available currently due to the gigapixel nature of WSIs, e.g. training a WSI at 10x magnification may require more than 300 Gb of GPU memory. Recently, researchers have started to explore optimization methods to enable end-to-end training of the entire network and entire WSI within GPU memory [25,21,29]. These methods show better performance compared to conventional MIL; they suffer, however, from two limitations. First, they are ImageNet-pretrained and do not leverage the powerful learning capabilities of histology-trained SSL models. Second, these are mostly limited to convolutional architectures rather than more effective attention-based architectures such as vision transformers [7].

Motivation: To improve WSI-level analysis, we explore end-to-end training of the entire network using SSL pretrained ViTs. To achieve this, we use the patch batching and gradient retaining techniques in [25]. However, we find that conventional fine-tuning approaches, where the entire network is fine-tuned, achieve low performance. For example, on the BRIGHT dataset[2], the accuracy drops more than 5% compared to the conventional MIL approaches. The poor performance is probably caused by the large network over-fitted to the limited downstream training data, leading to suboptimal feature representation. Indeed, especially for weakly supervised WSI classification, where annotated data for downstream tasks is significantly less compared to natural image datasets, conventional fine-tuning schemes can prove to be quite challenging.

To address the subpar performance of SSL-pretrained vision transformers, we utilize the prompt tuning techniques. Initially proposed in natural language processing, a prompt is a trainable or a pre-defined natural language statement that is provided as additional input to a transformer to guide the neural network towards learning a specific task or objective [3,12]. Using prompt tuning we *fine-tune only the prompt and downstream network without re-training the large backbone* (e.g. GPT-3 with 17B parameters). This approach is parameter efficient [12,15] and has been shown to better inject task-specific information and reduce the overfitting in downstream tasks, particularly in limited data scenarios [23,8]. Recently, prompts have also been adopted in computer vision and demonstrated superior performance compared to conventional fine-tuning methods [10]. Prompt tuning performs well even when only limited labeled data is available for training, making it particularly attractive in computational pathology. The process of prompt tuning thus involves providing a form of limited guidance during the training of downstream tasks, with the goal of minimizing the discrepancy between feature representations that are fully tuned to the task and those that are not task-specific.

In this paper, we propose a novel framework, Prompt-MIL, which uses prompts for WSI-level classification tasks within an MIL paradigm. Our contributions are:

- **Fine-tuning:** Unlike existing works in histopathology image analysis, Prompt-MIL is fine-tuned using prompts rather than conventional full fine-tuning methods.
- **Task-specific representation learning:** Our framework employs an SSL pretrained ViT feature extractor with a trainable prompt that calibrates the representations making them task-specific. By doing so, only the prompt parameters together with the classifier, are optimized. This avoids potential overfitting while still injecting task-specific knowledge into the learned representations.

Extensive experiments on three public WSI datasets, TCGA-BRCA, TCGA-CRC, and BRIGHT demonstrate the superiority of Prompt-MIL over conventional MIL methods, achieving a relative improvement of 1.49%-4.03% in accuracy and 0.25%-8.97% in AUROC by using only less than 0.3% additional parameters. Compared to the conventional full fine-tuning approach, we fine-tune less than 1.3% of the parameters, yet achieve a relative improvement of 1.29%-13.61% in accuracy and 3.22%-27.18% in AUROC. Moreover, compared to the full fine-tuning approach, our method reduces GPU memory consumption by 38%-45% and trains 21%-27% faster. To the best of our knowledge, this is the first work where prompts are explored for WSI classification. While our method is quite simple, it is versatile as it is agnostic to the MIL scheme and can be easily applied to different MIL methods. Our code is available at https://github.com/cvlab-stonybrook/PromptMIL.

2 Method

Our Prompt-MIL framework consists of three components: a frozen feature model to extract features of tissue patches, a classifier that performs an MIL scheme of feature aggregation and classification of the WSIs, and a trainable prompt. Given a WSI and its label y, the image is tiled into n tissue patches/instances $\{x_1, x_2, \ldots, x_n\}$ at a predefined magnification. As shown in Fig. 1, the feature model $F(\cdot)$ computes n feature representations from the corresponding n patches:

$$\begin{aligned} h &= [h_1, h_2, \ldots, h_n] \\ &= [F(x_1, \mathbb{P}), F(x_2, \mathbb{P}), \ldots, F(x_n, \mathbb{P})], \end{aligned} \tag{1}$$

where h_i denotes the feature of the i^{th} patch, h is the concatenation of all h_i, and $\mathbb{P} = \{p_i, i = 1, 2, \ldots, k\}$ is the trainable prompt consisting of k trainable tokens. The classifier $G(\cdot)$ applies an MIL scheme to predict the label $\hat{y}$ and calculate the loss $\mathcal{L}$ as:

$$\mathcal{L} = \mathcal{L}_{cls}(\hat{y}, y) = \mathcal{L}_{cls}(G(h), y), \tag{2}$$

where the $\mathcal{L}_{cls}$ is a classification loss.

2.1 Visual prompt tuning

The visual prompt tuning is the key component of our framework. As shown in Fig. 1(b), our feature model $F(\cdot)$ is a ViT based architecture. It consists of a patch embedding

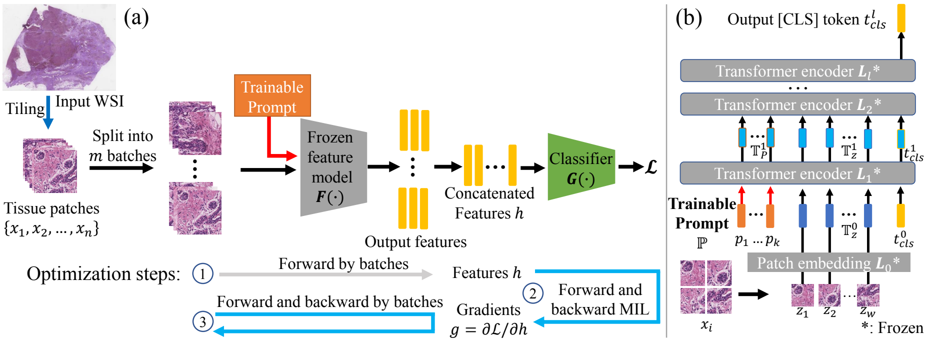

Fig. 1. Overview of the proposed method. (a) Overall structure of our training pipeline. Tissue patches tiled from the input WSI are grouped into separate batches, which are fed into a frozen feature model $F(\cdot)$ to compute their respective features. The features are subsequently concatenated into the feature h and a classifier $G(\cdot)$ applies an MIL scheme on h to predict the label and calculate the loss $\mathcal{L}$. (b) Structure of the feature model $F(\cdot)$ with the additional prompt. An input image x_i is cropped into w small patches $z_1, \ldots, z_w$. k trainable prompt tokens, together with the embedding of small patches and a class token t_{cls}^0, are fed into l layers of Transformer encoders. The output feature corresponding to x_i is the last class token t_{cls}^l. The feature model $F(\cdot)$ is frozen and only the prompt is trainable.

layer L_0 and l sequential encoding layers $\{L_1, L_2, \ldots, L_l\}$. The ViT first divides an input image x_i into w smaller patches $[z_1, z_2, \ldots, z_w]$ and embeds them into w tokens:

$$\mathbb{T}_z^0 = L_0([z_1, z_2, \ldots, z_w]) = \{t_1^0, t_2^0, \ldots, t_w^0\}, \tag{3}$$

where t_i^0 is the embedding token of z_i and $\mathbb{T}_z^0$ is the collection of such tokens. These tokens $\mathbb{T}_z^0$ are concatenated with a class token t_{cls}^0 and a prompt $\mathbb{P}$: The class token is used to aggregate information from all other tokens. The prompt consists of k trainable tokens $\mathbb{P} = \{p_i | i = 1, 2, \ldots, k\}$. The concatenation is fed into l layers of the Transformer encoders:

$$[\mathbb{T}_z^1, \mathbb{T}_P^1, t_{cls}^1] = L_1([\mathbb{T}_z^0, \mathbb{P}, t_{cls}^0]) \tag{4}$$

$$[\mathbb{T}_z^i, \mathbb{T}_P^i, t_{cls}^i] = L_i([\mathbb{T}_z^{i-1}, \mathbb{T}_P^{i-1}, t_{cls}^{i-1}]), i = 2, 3, \ldots, l \tag{5}$$

$$\mathbb{T}_P^i = \{p_j^i | j = 1, 2, \ldots, k\}, \tag{6}$$

where p_j^i is the j^{th} output prompt token of the i^{th} Transformer encoder and $\mathbb{T}_P^i$ is the collection of all k such output prompt tokens, which are not trainable. The output feature of x_i is defined as the last class token: $h_i = t_{cls}^l$.

2.2 Optimization

Our overall loss function is defined as

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{cls}(G(H), y) \\ &= \mathcal{L}_{cls}(G([F(x_1, \mathbb{P}), F(x_2, \mathbb{P}), \ldots, F(x_n, \mathbb{P})]), y), \end{aligned} \tag{7}$$

where only the parameters of the $G(\cdot)$ and the prompt $\mathbb{P}$ are optimized, while the feature extractor model $F(\cdot)$ is frozen.

Training the entire pipeline in an end-to-end fashion on gigapixel images is infeasible using the current hardware. To address this issue, we utilize the patch batching and gradient retaining techniques from [25]. As shown in Fig. 1(a), to reduce the GPU memory consumption, the n tissue patches $\{x_1, x_2, \ldots, x_n\}$ are grouped into m batches. The first step (step① in the figure) of our optimization is to sequentially feed m batches of tissue patches forward to the feature model to compute its respective features which are subsequently concatenated into the h matrix. In this step, we just conduct a forward pass like the inference stage, without storing the memory-intensive computational graph for back-propagation.

In the second step (step②), we feed h into the classifier $G(\cdot)$ to calculate the loss $\mathcal{L}$ and update the parameters of $G(\cdot)$ by back-propagate the loss. The back-propagated gradients $g = \partial\mathcal{L}/\partial h$ on h are retained for the next step.

Finally (step③), we feed the input batches into the feature model $F(\cdot)$ again and use the output h and the retained gradients g from the last step to update the trainable prompt tokens. In particular, the gradients on the j^{th} prompt token p_j are calculated as:

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial p_j} &= \frac{\partial\mathcal{L}}{\partial h}\frac{\partial h}{\partial p_j} \\ &= \sum_i \frac{\partial\mathcal{L}}{\partial h_i}\frac{\partial h_i}{\partial p_j} = \sum_i g_i \frac{\partial h_i}{\partial p_j}, \end{aligned} \tag{8}$$

where g_i is the gradient calculated with respect to h_i.

To sum up, in each step, we only update either F or G given the current batch, which avoid storing the gradients of the whole framework for all the input patches. This patch batching and gradient retaining techniques make the end-to-end training feasible.

In this study, we use DSMIL [13] as the classifier and binary cross entropy as the classification loss $\mathcal{L}_{cls}$ when the task is a tumor sub-type classification or cross entropy otherwise.

3 Experiments and Discussion

3.1 Datasets

We assessed Prompt-MIL using three histopathological WSI datasets: TCGA-BRCA [14], TCGA-CRC [19], and BRIGHT [2]. These datasets were utilized for both the self-supervised feature extractor pretraining and the end-to-end fine-tuning (with or without prompts), including the MIL component. Note that the testing data were not used in the SSL pretraining. **TCGA-BRCA** contains 1034 diagnostic digital slides of two breast cancer subtypes: invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). We used the same training, validation, and test split as that in the first fold cross validation in [5]. The cropped patches (790K training, 90K test) were extracted at 5× magnification. **TCGA-CRC** contains 430 diagnostic digital slides of colorectal cancer for a binary classification task: chromosomal instability (CIN) or genome stable (GS). Following the common 4-fold data split [1,16], we used the first three folds

Table 1. Comparison of accuracy and AUROC on three datasets. Reported metrics (in $\%$age) are the average across 3 runs. "Num. of Parameters" represents the number of optimized parameters

Dataset	TCGA-BRCA		TCGA-CRC		BRIGHT		Num. of
Metric	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Parameters
Conventional MIL	92.10	96.65	73.02	69.24	62.08	80.96	70k
Full fine-tuning	88.14	93.78	74.53	56.63	56.13	75.87	5.6M
Prompt-MIL (ours)	**93.47**	**96.89**	**75.47**	**75.45**	**64.58**	**81.31**	70k+192

for training (236 GS, 89 CIN), and the fourth for testing (77 GS, 28 CIN). We further split 20% (65 slides) training data as a validation set. The cropped patches (1.07M training, 370K test) were extracted at 10× magnification. **BRIGHT** contains 503 diagnostic slides of breast tissues. We used the official training (423 WSIs) and test (80 WSIs) splits. The task involves classifying non-cancerous (196 training, 25 test) vs. pre-cancerous (66 training, 23 test) vs. cancerous (161 training, 32 test). We further used 20% (85 slides) training slides for validation. The cropped patches (1.24M training, 195K test) were extracted at 10× magnification.

3.2 Implementation Details

We cropped non-overlapping 224 × 224 sized patches in all our experiments and used ViT-Tiny (ViT-T/16) [7] for feature extraction. For SSL pretraining, we leveraged the DINO framework [4] with the default hyperparameters, but adjusted the batch size to 256 and employed the global average pooling for token aggregation. We pretrained separate ViT models on the TCGA-CRC datasets for 50 epochs, on the BRIGHT dataset for 50 epochs, and on the BRCA dataset for 30 epochs. For TCGA-BRCA, we used the AdamW [17] optimizer with a learning rate of $1e-4$, $1e-2$ weight decay, and trained for 40 epochs. For TCGA-CRC, we also used the AdamW optimizer with a learning rate of $5e-4$ and trained for 40 epochs. For Bright, we used the Adam [11] optimizer with a learning rate of $1e-4$, $5e-2$ weight decay and trained for 40 epochs. We applied a cosine annealing learning rate decay policy in all our experiments. For the MIL baselines, we employed the same hyperparameters as above. For all full fine-tuning experiments, we used the learning rate in the corresponding prompt experiment as the base learning rate. For parameters in the feature model $F(\cdot)$, which are SSL pretrained, we use 1/10 of the base learning rate. For parameters in the Classifier $G(\cdot)$, which are randomly initialized, we use the base learning rate. We train the full tuning model for 10 more epochs than our prompt training to allow full convergence. This training strategy is optimized using the validation datasets. All model implementations were in PyTorch [20] on a NVIDIA Tesla V100 or a Nvidia Quadro RTX 8000.

3.3 Results

We chose overall accuracy and Area Under Receiver Operating Characteristic curve (AUROC) as the evaluation metrics.

Table 2. Comparison of GPU memory consumption and training speed per slide benchmarked on the BRIGHT dataset between the full fine-tuning and our prompt tuning on four slides with different sizes. Our prompt method requires far less memory and is significantly faster.

	WSI size	$44k \times 21k$	$26k \times 21k$	$22k \times 17k$	$11k \times 16k$
	#Tissue patches	9212	4765	2307	1108
GPU Mem.	Full fine-tuning	21.81G	18.22G	16.37G	12.71G
	Prompt (ours)	**12.04**G	**10.66**G	**10.00**G	**7.90**G
	Reduction percentage	44.79%	41.50%	38.92%	37.84%
Time per slide	Full fine-tuning	17.73s	8.92s	4.37s	2.15s
	Prompt (ours)	**13.92**s	**7.09**s	**3.35**s	**1.56**s
	Reduction percentage	21.49%	20.51%	23.32%	27.27%

Evaluation of prompt tuning performance: We compared the proposed Prompt-MIL with two baselines: 1) a conventional MIL model with a frozen feature extractor [13], 2) fine-tuning all parameters in the feature model (full fine-tuning). Table 1 highlights that our Prompt-MIL consistently outperformed both. Compared to the conventional MIL method, Prompt-MIL added negligible parameters (192, less than 0.3% of the total parameters), achieving a relative improvement of 1.49% in accuracy and 0.25% in AUROC on TCGA-BRCA, 3.36% in accuracy and 8.97% in AUROC on TCGA-CRC, and 4.03% in accuracy and 0.43% in AUROC on BRIGHT. The observed improvement can be attributed to a more optimal alignment between the feature representation learned during the SSL pretraining and the downstream task, i.e., the prompt explicitly calibrated the features toward the downstream task.

The computationally intensive full fine-tuning method under-performed conventional MIL and Prompt-MIL. Compared to the full fine-tuning method, our method achieved a relative improvement of 1.29% to 13.61% in accuracy and 3.22% to 27.18% in AUROC on the three datasets. Due to the relatively small amount of slide-level labels (few hundred to a few thousands) fully fine tuning 5M parameters in the feature model might suffer from overfitting. In contrast, our method contained less than 1.3% of parameters compared to full fine-tuning, leading to robust training.

Table 3. Comparison of accuracy and AUROC on three datasets for a pathological foundation model.

Dataset	TCGA-BRCA		BRIGHT	
Metric	Accuracy	AUROC	Accuracy	AUROC
ViT-small [27]	91.75	97.03	54.17	76.76
ViT-small w/ Prompt-MIL	**92.78**	**97.53**	**57.50**	**78.29**

Evaluation of time and GPU memory efficiency: Prompt-MIL is an efficient method requiring less GPU memory to train and running much faster than full fine-tuning methods. We evaluated the training speed and memory consumption of our method and compared to the full fine-tuning baseline on four different sized WSIs in the BRIGHT dataset. As shown in Table 2, our method consumed around 38% to 45% less GPU memory compared to full fine-tuning and was 21% to 27% faster. As we scaled up the WSI size (i.e. WSIs with more number of patches), the memory cost difference between Prompt-MIL and full fine-tuning further widened.

Evaluation on the pathological foundation models: We demonstrated our Prompt-MIL also had a better performance when used with the pathological foundation model. Foundational models refer to those trained on large-scale pathology datasets (e.g. the entire TCGA Pan-cancer dataset [28]). We utilized the publicly available [27,26] ViT-Small network pretrained using MoCo v3 [6] on all the slides from TCGA [28] and PAIP [22]. In Table 3, we showed that our method robustly boosted the performance on both TCGA (the same domain as the foundation model trained on) and BRIGHT (a different domain). The improvement is more prominent in BRIGHT, which further confirmed that Prompt-MIL aligns the feature extractor to be more task-specific.

Table 4. Performance with a different number of prompt tokens. For two different WSI classification tasks, one token was enough to boost the performance of the conventional MIL schemes.

Dataset	TCGA-BRCA		BRIGHT	
#prompt tokens k	Accuracy	AUROC	Accuracy	AUROC
$k = 1$	**93.47**	**96.89**	**64.58**	**81.31**
$k = 2$	93.13	**96.93**	60.41	79.74
$k = 3$	**93.47**	96.86	59.17	76.75

Ablation study: An ablation was performed to study the effect of the number of trainable prompt tokens on downstream tasks. Table 4 shows the accuracy and AUROC of our Prompt-MIL model with 1, 2 and 3 trainable prompt tokens ($k = 1, 2, 3$) on the TCGA-BRCA and the BRIGHT datasets. On the TCGA-BRCA dataset, our Prompt-MIL model with 1 to 3 prompt tokens reported similar performance. On the BRIGHT dataset, the performance of our model dropped with the increased number of prompt tokens. Empirically, this ablation study shows that for classification tasks, one prompt token is sufficient to boost the performance of conventional MIL methods.

4 Conclusion

In this work, we introduced a new framework, Prompt-MIL, which combines the use of Multiple Instance Learning (MIL) with prompts to improve the performance of WSI classification. Prompt-MIL adopts a prompt tuning mechanism rather than a conventional full fine-tuning of the entire feature representation. In such a scheme, only a small fraction of parameters calibrates the pretrained representations to encode task-specific information, so the entire training can be performed in an end-to-end manner. We applied our proposed method to three publicly available datasets. Extensive experiments demonstrated the superiority of Prompt-MIL over the conventional MIL as well as the conventional fully fine-tuning methods. Moreover, by fine-tuning much fewer parameters compared to fully fine-tuning, our method is GPU memory efficient and fast. Our proposed approach also showed promising potentials in transferring foundation models. We will further explore the task-specific features that are captured by our prompt toward explainability of these models.

Acknowledgements This work was partially supported by the ANR Hagnodice ANR-21-CE45-0007, the NSF IIS-2212046, the NSF IIS-2123920, the NIH 1R21CA258493-01A1, the NCI UH3CA225021 and Stony Brook University Provost Funds. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., Rajpoot, N.M.: Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. The Lancet Digital Health (2021)
2. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022** (2022)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
5. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16144–16155 (June 2022)
6. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. Gu, Y., Han, X., Liu, Z., Huang, M.: Ppt: Pre-trained prompt tuning for few-shot learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8410–8423 (2022)

9. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2424–2433 (2016)
10. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII. pp. 709–727. Springer (2022)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
12. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059 (2021)
13. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
14. Lingle, W., Erickson, B., Zuley, M., Jarosz, R., Bonaccio, E., Filippini, J., Gruszauskas, N.: Radiology data from the cancer genome atlas breast invasive carcinoma [tcga-brca] collection. The Cancer Imaging Archive **10**, K9 (2016)
15. Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68 (2022)
16. Liu, Y., Sethi, N.S., Hinoue, T., Schneider, B.G., Cherniack, A.D., Sanchez-Vega, F., Seoane, J.A., Farshidfar, F., Bowlby, R., Islam, M., et al.: Comparative molecular analysis of gastrointestinal adenocarcinomas. Cancer cell (2018)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
18. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
19. Network, C.G.A., et al.: Comprehensive molecular characterization of human colon and rectal cancer. Nature (2012)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
21. Pinckaers, H., Van Ginneken, B., Litjens, G.: Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1581–1590 (2020)
22. Platform, P.A.: Paip (2021), data retrieved from PAIP, http://http://www.wisepaip.org/paip/
23. Schucher, N., Reddy, S., de Vries, H.: The power of prompt tuning for low-resource semantic parsing. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 148–156 (2022)
24. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
25. Takahama, S., Kurose, Y., Mukuta, Y., Abe, H., Fukayama, M., Yoshizawa, A., Kitagawa, M., Harada, T.: Multi-stage pathological image classification using semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10702–10711 (2019)
26. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X.: Transpath: Transformer-based self-supervised learning for histopathological image classification. In:

Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. Springer (2021)

27. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis **81**, 102559 (2022)
28. Weinstein, J.N., Collison, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Noushmehr, H.: The cancer genome atlas pan-cancer analysis project. Nature Genetics **45**(10), 1113–1120 (2013)
29. Zhang, J., Zhang, X., Ma, K., Gupta, R., Saltz, J., Vakalopoulou, M., Samaras, D.: Gigapixel whole-slide images classification using locally supervised learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 192–201. Springer (2022)