PATE: Enhancing Few-Shot Pathological Image Classification via Prompt-Based Text-Image Embedding Adaptation

Shenghao Chen[1,2], Zhen Huang[2,3,5⋆], Xiaoqian Zhou[1,2], Han Li[1,2,4], Chunjiang Wang[1,2], and S. Kevin Zhou[1,2,6,7⋆⋆]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei Anhui, 230026, China
[2] Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advance Research, USTC, Suzhou Jiangsu, 215123, China
[3] School of Computer Science and Technology, USTC, Hefei, Anhui, 230026, China
[4] Computer Aided Medical Procedures (CAMP), TU Munich, 80333, Germany
[5] Eastern Institute of Technology, Ningbo Zhejiang, 315200, China
[6] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou Jiangsu, 215123, China
[7] State Key Laboratory of Precision and Intelligent Chemistry, USTC, Hefei Anhui 230026, China

Abstract. Automated pathological image classification remains a critical challenge, particularly due to the scarcity of annotated data and the complexity of disease-specific features. Existing methods, such as CLIP-based prompt tuning, struggle with limited few-shot learning and poor integration of multimodal information in medical contexts. In this study, we introduce PATE (**P**rompt-based **A**daptation for **T**ext-Image **E**mbedding), a novel framework to enhance CLIP's adaptability for few-shot pathological image classification. Our approach incorporates deep learnable prompts in both vision and language encoders, enabling effective use of visual and textual information. We also propose a dynamic bridging function for bidirectional information exchange and a Gaussian-weighted Prompt Integration (GPI) strategy to adjust prompt contributions across epochs, enhancing generalization and reducing overfitting. Extensive experiments on the PatchGastric dataset, which includes 179,285 histopathological patches across three gastric adenocarcinoma subtypes, demonstrate that PATE consistently outperforms state-of-the-art methods, achieving superior performance in both low-data and full-data settings. Ablation studies validate the effectiveness of each component, marking a significant advancement in few-shot medical image analysis, particularly in rare disease diagnosis and digital pathology workflows.

Keywords: Pathological image classification · Few-shot · Prompt tuning.

⋆ S. Chen and Z. Huang contributed equally to this work.
⋆⋆ Corresponding author: skevinzhou@ustc.edu.cn, tum_han.li@tum.de

1 Introduction

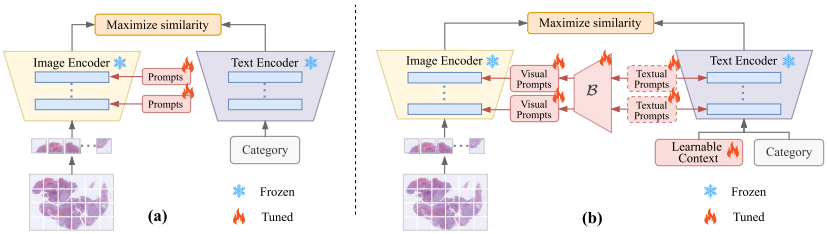

Fig. 1. Comparison of PATE with the state-of-the-art method. (a) The existing SOTA method CITE only applies prompt tuning to the vision modality of CLIP. (b) PATE introduces hierarchical multimodal prompts and a vision-language bridge.

The rapid advancement of deep learning has significantly impacted medical imaging, especially in pathological image analysis. The automated analysis of pathology Whole Slide Images (WSIs) [4, 16] is critical in cancer diagnosis and predicting treatment response. However, analyzing WSIs is challenging due to their gigapixel resolution [11], which makes them unsuitable for direct input into deep learning models. To address this, a common approach involves dividing WSIs into non-overlapping small patches for processing [20]. Moreover, the scarcity of annotated data [17, 30, 13, 38], particularly for rare diseases, limits the performance of traditional deep learning methods. In response, few-shot weakly supervised learning for WSIs classification (FSWC) [22] has emerged. This approach enables models to learn from limited labeled data, with a common and effective method using pre-trained [6] models combined with parameter-efficient fine-tuning techniques [10, 25]. In the context of FSWC, this approach enables effective AI-assisted diagnosis of WSIs [7] even with limited annotated data, ensuring robust performance in data-scarce scenarios [26, 12, 27].

A promising direction for improving model performance is multimodal learning [24, 31], which integrates visual and textual data to enhance understanding and generalization. CLIP (Contrastive Language-Image Pretraining) [23] aligns images and text in a shared semantic space, enabling zero-shot learning [32]. However, applying CLIP to medical imaging, such as pathological classification, is challenging due to the fine-grained, domain-specific nature of medical images [3, 18] and the need for precise alignment with clinical texts [21]. To address this, Prompt Tuning has emerged as an efficient adaptation method, allowing models like CLIP to adapt to new tasks with minimal computational cost [8] while preserving generalization and reducing overfitting [5, 10]. As shown in Figure 1(a), methods like CITE [33] for pathological image classification introduce learnable prompts only at the image encoder stage. However, this approach fails to exploit the intricate interactions between visual and textual modalities. By optimizing the image encoder independently, it misses the potential of leverag-

ing multimodal relationships, which are critical for improving diagnostic performance in medical contexts [2].

In this study, we propose **P**rompt-based **A**daptation for **T**ext-Image **E**mbedding (PATE), see Figure 1(b), a novel framework for FSWC. PATE incorporates deep prompts across multiple transformer layers in vision and language encoders. This dual-modal approach enables the model to learn from visual prompts, capturing spatial and appearance features, and textual prompts, learning contextual and semantic representations, thus improving multimodal feature extraction. By integrating these prompts at different depths, PATE enhances low-level and high-level feature learning, promoting better multimodal representation for pathological image analysis.

Inspired by CoOp [36] and CoCoOp [35], PATE introduces learnable context to model specific categories, which is particularly beneficial for fine-grained classification tasks. To strengthen the synergy between vision and language, PATE includes a bridging function that conditions visual prompts on textual prompts and vice versa, ensuring bidirectional information exchange. This mechanism leverages the strengths of both modalities, enhancing multimodal adaptation.

We propose a Gaussian-weighted Prompt Integration (GPI) strategy, which adaptively integrates prompts across training epochs using Gaussian-distributed weights to improve prompt robustness and generalization. This prioritizes prompts from intermediate epochs, which are more task-relevant while reducing the influence of less relevant initial and final prompts, enhancing performance in few-shot learning and cross-domain adaptation [34]. In summary, the main contributions of this work include:

1. **Hierarchical Multimodal Prompting Strategy(HMPS):** We integrate deep multi-layer prompts into vision and language encoders, enhancing structural-semantic feature extraction and improving multimodal representation learning under data scarcity through hierarchical cross-modal alignment.
2. **Bridging Function(BF):** We propose a cross-modal interaction mechanism that explicitly establishes bidirectional conditioning between visual and textual prompts. This mechanism facilitates synergistic information flow to enhance vision-language representation alignment and improve classification accuracy.
3. **Gaussian-weighted Prompt Integration(GPI):** We introduce a dynamic prompt integration strategy that employs Gaussian-weighted attention to prioritize prompts from intermediate epochs. This strategy significantly enhances few-shot learning and cross-domain adaptation in medical imaging tasks.

2 Methodology

Our approach leverages prompt tuning on the pre-trained CLIP [23] model to improve its performance on pathological image classification tasks. As shown in Figure 2(a), we introduce joint prompt tuning for both the textual and visual encoders by adding trainable tokens as prompts within the encoder layers of

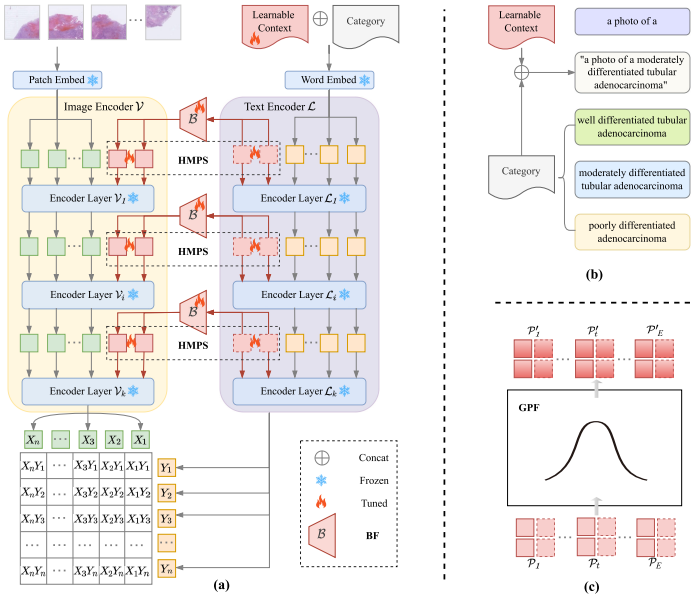

Fig. 2. Overview of PATE. (a) PATE framework with deep prompting and a bridging function. (b) Text input construction by concatenating learnable context tokens with labels. (c) GPI for adaptive prompt aggregation across epochs.

both branches and injecting these context prompts at different depths within the transformer blocks to capture hierarchical contextual features. Additionally, a bridging function is proposed to model the interactions between image and text representations, enhancing mutual information and improving multimodal learning during training. Only the context prompts and bridging function parameters are updated during fine-tuning, while the rest of the model remains frozen. In the following sections, we first revisit CLIP, followed by a detailed description of the design of our approach.

2.1 Preliminaries

In this section, we revisit CLIP's structure and its application to Whole Slide Imaging (WSI) pathology classification.

Vision and Text Encoding. Given a pathology image $I \in \mathbb{R}^{H\times W\times 3}$, the vision encoder in CLIP uses a Vision Transformer (ViT) model [28], where the image is divided into M fixed-size patches, each projected into embeddings $E_0 \in \mathbb{R}^{M\times d_v}$. These patch embeddings are processed through K transformer layers, each incorporating a class token c_i to integrate features from the entire image. The final class token c_K is projected into a shared latent space, yielding the image feature $x_v = \text{ImageProj}(c_K) \in \mathbb{R}^{d_v}$. For the text description, the words are tokenized and projected into word embeddings $W_0 = [w_0^1, w_0^2, \ldots, w_0^N] \in \mathbb{R}^{N\times d_l}$,

which are processed by a separate text transformer model, generating the final text representation $z_c = \text{TextProj}(w_K) \in \mathbb{R}^{d_l}$.

Image-Text Feature Matching and Classification. For pathology image classification, the image feature x_v and the text feature z_c are compared by calculating their cosine similarity. Given C classes, the prediction process is as follows:

$$p(\hat{y} = c|I) = \frac{\exp\left(\frac{\text{sim}(x_v, z_c)}{\tau}\right)}{\sum_{c'=1}^{C} \exp\left(\frac{\text{sim}(x_v, z_{c'})}{\tau}\right)} \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature scaling parameter that adjusts the magnitude of the similarity [9]. By maximizing the cosine similarity between image features and class text features, CLIP [23] can assign pathology images to the correct category.

2.2 Deep Prompting and Bridging Function

To optimize the alignment between the vision and language branches of CLIP, we introduce deep prompting in both modalities by incorporating learnable context tokens at multiple layers of the respective encoders. The prompts are initialized using a truncated normal distribution ($\mu = 0$, $\sigma = 0.02$), consistent with CLIP's original implementation. This approach enhances the model's ability to progressively adapt the representations to downstream tasks by learning rich contextual information from each stage of the transformer blocks.

Language Prompting. For the text encoder, we introduce a set of b learnable prompt tokens $\{P_i \in \mathbb{R}^{d_l}\}_{i=1}^{b}$, where b denotes the number of prompt tokens. These tokens are concatenated with the initial word embeddings $W_0 = [w_1, w_2, \ldots, w_N] \in \mathbb{R}^{N \times d_l}$. These tokens are injected into the transformer layers, where at each layer i, the prompt tokens are concatenated with the word embeddings W_i and processed through the corresponding language transformer block L_i:

$$[P_i, W_i] = L_i([P_{i-1}, W_{i-1}]) \quad i = 1, 2, \ldots, J \tag{2}$$

where J represents the injection depth, indicating the maximum layer index for prompt token integration. After the J-th layer, the remaining layers continue processing the previous prompts, leading to the final text representation z:

$$[P_j, W_j] = L_j([P_{j-1}, W_{j-1}]) \quad j = J+1, \ldots, K \tag{3}$$

$$z = \text{TextProj}(w_K^N) \tag{4}$$

In particular, as shown in Figure 2(b), the text input for the word embedding is formed by concatenating learnable context tokens with the category labels. These learnable context tokens belong to the dynamic prompt and are initialized with descriptions such as 'a photo of a poorly differentiated adenocarcinoma.'

Vision Prompting. Similarly, in the vision branch, we introduce b learnable prompt tokens $\{\tilde{P}_i \in \mathbb{R}^{d_v}\}_{i=1}^{b}$ alongside the initial patch embeddings of the

image. These context tokens are inserted into the image encoder's transformer layers. At each layer i, the prompt tokens are concatenated with the image patch embeddings E_{i-1} and processed by the transformer block V_i:

$$[c_i, E_i, \tilde{P}_i] = V_i([c_{i-1}, E_{i-1}, \tilde{P_{i-1}}]) \quad i = 1, 2, \ldots, J \tag{5}$$

At deeper layers, the tokens continue to propagate, and the final image representation x is obtained:

$$[c_i, E_i] = V_i([c_{i-1}, E_{i-1}]), \quad J+1 \leq i \leq K. \tag{6}$$

$$x = \text{ImageProj}(c_K) \tag{7}$$

Bridging Function. To align the vision and language branches in the model, we introduce a bridging function $F(\cdot)$ that projects the language prompts into the vision prompt space. This ensures effective interaction and synchronized learning between the two modalities. Specifically, the bridging function applies a linear layer to the language prompts $P_k \in \mathbb{R}^{d_l}$, mapping them to the vision prompt space $\tilde{P}_k \in \mathbb{R}^{d_v}$, where d_l and d_v are the respective embedding dimensions. At each layer i of the vision encoder, the transformed language prompt $F_{i-1}(P_{i-1})$ is concatenated with the previous vision embeddings and processed through the transformer block V_i:

$$\tilde{P}_i = F_i(P_i) \quad i = 1, 2, \ldots, J \tag{8}$$

2.3 Gaussian-weighted Prompt Integration

As illustrated in Figure 2(c), we introduce "Gaussian-weighted Prompt Integration (GPI)" to integrate prompt tokens across training epochs using Gaussian-distributed weights. The weights are designed to emphasize intermediate epochs more heavily, as initial and final prompts receive relatively lower weights due to their reduced relevance for task-specific representation. The integrated prompt $P_t{}'$ is weighted by a Gaussian distribution $w_t \sim \mathcal{N}(\mu, \sigma^2)$, where t denotes the $t-th$ epoch, μ and σ^2 are hyper-parameters and $\sum_{t=1}^{E} w_t = 1$:

$$P_t{}' = \sum_{t=1}^{E} w_t \cdot P_t \tag{9}$$

3 Experiments

Dataset. We use the PatchGastric [29] dataset, which contains a total of 262,777 patches, each of size 300x300 pixels, extracted from 991 whole slide images (WSIs) of gastric adenocarcinoma biopsy specimens, captured at magnifications of $\times 20$. The dataset includes nine different gastric adenocarcinoma subtypes, from which we select three major subtypes—well-differentiated tubular adenocarcinoma, moderately differentiated tubular adenocarcinoma, and poorly differentiated adenocarcinoma—to form a three-class grading classification task.

This subset includes 179,285 patches from 693 WSIs, with the dataset split into training and validation sets in a 2:8 ratio.

Implementation Details. We experiment with a few-shot learning setup to simulate scenarios with limited labeled data, selecting 1, 2, 4, 8, or 16 WSIs per class for training. We use a 16-shot setting for all experiments, where 16 samples are randomly chosen for each class to form the training set. We employ the pre-trained ViT-B/16 [23] CLIP model as the backbone for prompt tuning, where the text encoder embedding dimension is set to $d_l = 512$ and the vision encoder embedding dimension is $d_v = 768$. For PATE, the prompt depth J is set to 3 for few-shot settings with 1 or 2 WSIs, 5 for few-shot settings with 4 or 8 WSIs, and 9 for few-shot settings with 16 WSIs or the full dataset. We use 2 learnable prompt tokens for both modalities at each injected layer. The network is trained using the AdamW optimizer [19] with a learning rate of 0.0001 for 150 epochs and a batch size of 16. Training is conducted on eight NVIDIA RTX 3090 GPUs. The classification performance is evaluated using accuracy, averaged over three independent runs to ensure reliability. The text prompts follow the template "a photo of a <category>," while all other model parameters are randomly initialized from a normal distribution.

4 Results

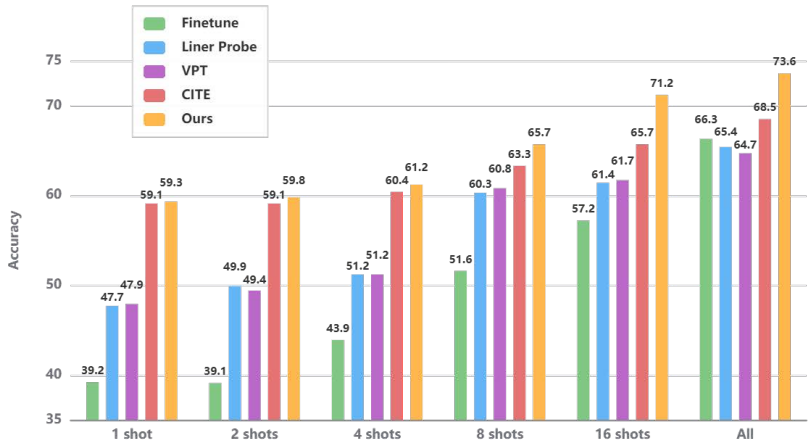

Fig. 3. Comparison of accuracy on the PatchGastric 3-category classification task.

Comparison with Baseline Models. We evaluate the performance of our proposed model, PATE, against several related baseline methods, including traditional fine-tuning [1], Linear Probe [15], VPT [14], and CITE [33], on the PatchGastric [29] dataset. The results in Figure 3 demonstrate that PATE outperforms all baselines across various data scales, particularly in few-shot settings.

In the 1-shot scenario, PATE achieves 59.3% accuracy, surpassing CITE (59.1%), Finetune (39.2%), Linear Probe (47.7%), and VPT (47.9%). As the number of training samples increases, PATE's advantage grows, reaching 71.2% accuracy in the 16-shot scenario, outperforming CITE (65.7%) and other baselines. This trend holds for 4-shot and 8-shot settings, where PATE improves more significantly. PATE's ability to leverage both visual and textual information helps it generalize better, especially with more data. When using all available data, PATE achieves 73.6%, outperforming CITE (68.5%), Finetune (66.3%), Linear Probe (65.4%), and VPT (64.7%), highlighting its superior performance.

Ablation Study. We conduct an ablation study to evaluate the contributions of various components in our proposed method. The factors considered include Learnable Contexts (LC), Visual Prompts (VP), Textual Prompts (TP), bridging Function (CF), and Gaussian-weighted Prompt Integration (GPI). Table 1 summarizes the performance under different few-shot learning settings (1, 2, 4, 8, 16 shots) on the PatchGastric [29] dataset. To simulate real-world limited labeled data, we evaluate the model using 1-shot and 2-shot settings. The baseline accuracy, without any additional components, is 39.1% in the 1-shot and 39.0% in the 2-shot settings. Introducing Learnable Contexts (LC) alone results in a noticeable improvement, with accuracy rising to 47.9% in the 1-shot setting and 49.6% in the 2-shot setting. The addition of Visual Prompts (VP) further enhances performance, pushing the accuracy to 50.3% in the 1-shot and 53.8% in the 2-shot settings. When both VP and Textual Prompts (TP) are incorporated, we observe a significant performance boost, with accuracy reaching 58.7% in the 1-shot and 59.3% in the 2-shot settings. This demonstrates the substantial synergy between visual and textual cues in improving the model's classification ability. Next, the bridging function (BF), which enables better interaction between the visual and textual prompts, refines the model even further, achieving 59.3% in the 1-shot and 59.4% in the 2-shot settings. Notably, the Gaussian-weighted Prompt Integration (GPI) strategy demonstrates consistent improvements across all few-shot settings, particularly showing a 2.3% accuracy gain in the 16-shot scenario, validating its effectiveness in mitigating overfitting through mid-epoch prompt emphasis. These results highlight the significant impact of VP and TP in improving model performance, with CF and GPI providing additional refinement, demonstrating the effectiveness of each component in enhancing generalization with limited labeled data.

Table 1. Ablation study of PATE.

LC	VP	TP	BF	GPI	1	2	4	8	16	All
					39.1	39.0	44.1	51.7	51.7	66.0
✓					47.9	49.6	52.3	56.4	59.2	66.7
✓	✓				50.3	53.8	57.9	59.2	63.1	68.4
✓	✓	✓			58.7	59.3	60.4	63.6	66.1	69.5
✓	✓	✓	✓		59.3	59.4	60.9	64.2	68.9	70.3
✓	✓	✓	✓	✓	**59.3**	**59.8**	**61.2**	**65.7**	**71.2**	**73.6**

5 Conclusion

We propose PATE, a novel framework for few-shot pathological image classification that leverages multimodal prompt learning. By incorporating deep learnable prompts into both vision and language encoders, PATE improves multimodal feature extraction and enhances model performance in data-scarce scenarios. The dynamic bridging function facilitates bidirectional information exchange between the modalities, while Gaussian-weighted Prompt Integration (GPI) boosts generalization and reduces overfitting. Experiments on the PatchGastric [29] dataset show that PATE outperforms state-of-the-art methods, significantly improving rare disease detection and digital pathology. Future work will extend PATE to other medical imaging tasks and explore its broader applicability [37].

Acknowledgements This work was supported by the Natural Science Foundation of China (Grant 62271465), the Suzhou Basic Research Program (Grant SYG202338), and IMI BigPicture project (IMI945358)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alt, C., Hübner, M., Hennig, L.: Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. arXiv preprint arXiv:1906.08646 (2019)
2. Branson, C.F., Williams, M., Chan, T.M., Graber, M.L., Lane, K.P., Grieser, S., Landis-Lewis, Z., Cooke, J., Upadhyay, D.K., Mondoux, S., et al.: Improving diagnostic performance through feedback: the diagnosis learning cycle. BMJ quality & safety **30**(12), 1002–1009 (2021)
3. Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S.: Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4109–4118 (2018)
4. Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep learning for whole slide image analysis: an overview. Frontiers in medicine **6**, 264 (2019)
5. Ding, C., Gao, X., Dong, S., He, Y., Wang, Q., Kot, A., Gong, Y.: Lobg: Less overfitting for better generalization in vision-language model. arXiv preprint arXiv:2410.10247 (2024)
6. Du, Y., Liu, Z., Li, J., Zhao, W.X.: A survey of vision-language pre-trained models. arXiv preprint arXiv:2202.10936 (2022)
7. Eloy, C., Marques, A., Pinto, J., Pinheiro, J., Campelos, S., Curado, M., Vale, J., Polónia, A.: Artificial intelligence–assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies. Virchows Archiv **482**(3), 595–604 (2023)
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
9. Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., Zavolan, M.: Clip and complementary methods. Nature Reviews Methods Primers **1**(1), 20 (2021)
10. Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S.Q.: Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608 (2024)

11. Harb, R., Pock, T., Müller, H.: Diffusion-based generation of histopathological whole slide images at a gigapixel scale. In: WACV. pp. 5131–5140 (2024)
12. Huang, Z., Li, H., Shao, S., Zhu, H., Hu, H., Cheng, Z., Wang, J., Kevin Zhou, S.: Pele scores: pelvic x-ray landmark detection with pelvis extraction and enhancement. IJCARS **19**(5), 939–950 (2024)
13. Huang, Z., Zhou, X., He, X., Wei, Y., Yang, W., Wang, S., Sun, X., Li, H.: Casemark: A hybrid model for robust anatomical landmark detection in multi-structure x-rays. Journal of King Saud University Computer and Information Sciences **37**(3), 1–18 (2025)
14. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
15. Knuth, D.E.: Linear probing and graphs. Algorithmica **22**, 561–568 (1998)
16. Kothari, S., Phan, J.H., Stokes, T.H., Wang, M.D.: Pathology imaging informatics for quantitative analysis of whole-slide images. Journal of the American Medical Informatics Association **20**(6), 1099–1108 (2013)
17. Laurer, M., Van Atteveldt, W., Casas, A., Welbers, K.: Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. Political Analysis **32**(1), 84–100 (2024)
18. Li, X., Zhao, L., Zhang, L., Wu, Z., Liu, Z., Jiang, H., Cao, C., Xu, S., Li, Y., Dai, H., et al.: Artificial general intelligence for medical imaging analysis. IEEE Reviews in Biomedical Engineering (2024)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Nouyed, M.I.: Efficient classification of very high resolution images (2024)
21. Prud'hommeaux, E., Roark, B.: Graph-based word alignment for clinical language evaluation. Computational Linguistics **41**(4), 549–578 (2015)
22. Qu, L., Fu, K., Wang, M., Song, Z., et al.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. NIPS **36**, 67551–67564 (2023)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
24. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine **34**(6), 96–108 (2017)
25. Rawhani, M., Karaboğa, D., Nalbantoglu, U., Başt ürk, A., Akay, B.: Efficient unsupervised domain adaptation with peft combinations. In: 2024 9th International Conference on Computer Science and Engineering (UBMK). pp. 169–174. IEEE (2024)
26. Ren, Z., Han, H., Cui, X., Lu, H., Luo, M.: Novel data-pulling-based strategy for chiller fault diagnosis in data-scarce scenarios. Energy **279**, 128019 (2023)
27. Shao, S., Yuan, X., Huang, Z., Qiu, Z., Wang, S., Zhou, K.: Diffuseexpand: Expanding dataset for 2d medical image segmentation using diffusion models. arXiv preprint arXiv:2304.13416 (2023)
28. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383 (2021)
29. Tsuneki, M., Kanavati, F.: Inference of captions from histopathological patches. In: International Conference on Medical Imaging with Deep Learning. pp. 1235–1250. PMLR (2022)

30. Ul Haq, M.U., Rigoni, D., Sperduti, A.: Prompt-based data augmentation using contrastive learning under scarcity of annotated data. In: ECAI 2024, pp. 2717–2724. IOS Press (2024)
31. Wang, R., Yao, Q., Lai, H., He, Z., Tao, X., Jiang, Z., Zhou, S.K.: Ecamp: Entity-centered context-aware medical vision language pre-training. arXiv preprint arXiv:2312.13316 (2023)
32. Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., Tan, T.: Improving zero-shot generalization for clip with synthesized prompts. arXiv preprint arXiv:2307.07397 (2023)
33. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: MICCAI. pp. 272–282. Springer (2023)
34. Zhao, A., Ding, M., Lu, Z., Xiang, T., Niu, Y., Guan, J., Wen, J.R.: Domain-adaptive few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1390–1399 (2021)
35. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
36. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)
37. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)
38. Zhou, X., Huang, Z., Zhu, H., Yao, Q., Zhou, S.K.: Hybrid attention network: An efficient approach for anatomy-free landmark detection. arXiv preprint arXiv:2412.06499 (2024)