Fine-grained Prompt Tuning: A Parameter and Memory Efficient Transfer Learning Method for High-resolution Medical Image Classification

Yijin Huang[1,2], Pujin Cheng[1,3,4], Roger Tam[2(✉)], and Xiaoying Tang[1,4(✉)]

[1] Department of Electronic and Electrical Engineering,
Southern University of Science and Technology, Shenzhen, China
tangxy@sustech.edu.cn
[2] School of Biomedical Engineering,
The University of British Columbia, Vancouver Canada
roger.tam@ubc.ca
[3] Department of Electrical and Electronic Engineering,
The University of Hong Kong, Hong Kong, China
[4] Jiaxing Research Institute,
Southern University of Science and Technology, Jiaxing, China

Abstract. Parameter-efficient transfer learning (PETL) is proposed as a cost-effective way to transfer pre-trained models to downstream tasks, avoiding the high cost of updating entire large-scale pre-trained models (LPMs). In this work, we present Fine-grained Prompt Tuning (FPT), a novel PETL method for medical image classification. FPT significantly reduces memory consumption compared to other PETL methods, especially in high-resolution input contexts. To achieve this, we first freeze the weights of the LPM and construct a learnable lightweight side network. The frozen LPM takes high-resolution images as input to extract fine-grained features, while the side network is fed low-resolution images to reduce memory usage. To allow the side network to access pre-trained knowledge, we introduce fine-grained prompts that summarize information from the LPM through a fusion module. Important tokens selection and preloading techniques are employed to further reduce training cost and memory requirements. We evaluate FPT on four medical datasets with varying sizes, modalities, and complexities. Experimental results demonstrate that FPT achieves comparable performance to fine-tuning the entire LPM while using only 1.8% of the learnable parameters and 13% of the memory costs of an encoder ViT-B model with a 512×512 input resolution.

Keywords: Parameter-efficient transfer learning · Memory-efficient transfer learning · High-resolution medical image classification.

1 Introduction

By utilizing the technique of fine-tuning [26], pre-trained models can be effectively adapted to specific downstream tasks by initializing the task-specific model
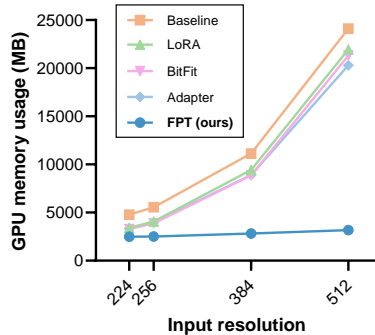
Fig. 1: High-resolution comes at the cost of heightened GPU memory consumption.
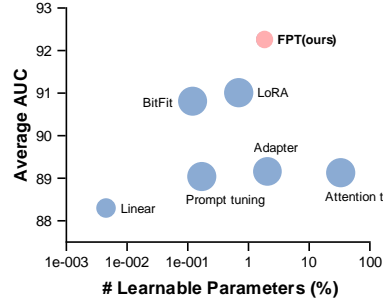

Fig. 2: Our proposed FPT shows the best trade-off between performance and efficiency. The size of the dots represents memory usage.

with the weights from the pre-trained model. Recently, the remarkable achievements of large-scale pre-trained models (LPMs) [5,14,19,20] further underscore the importance of this technique. However, as the model size grows rapidly, fine-tuning the parameters of an entire LPM has become very costly. To address this challenge, the concept of parameter-efficient transfer learning (PETL) [9,10,13,23,27,29] has emerged, offering a strategic approach for transferring pre-trained models. PETL involves selectively updating a small subset of pre-trained parameters or introducing a modest number of additional parameters specific to new tasks, while keeping the majority of pre-trained parameters frozen.

PETL has successfully established its effectiveness in computer vision, but the field of medical image has not fully benefited from such advances yet [7], mainly because of the domain gap between them. In natural images, the objects of interest typically occupy a large portion of the image and exhibit distinct characteristics. In contrast, diagnostic cues in medical images often occupy a small portion and are distributed throughout the entire image [12]. Providing such fine-grained information often requires the use of high-resolution input images [2,11,30]. However, as shown in Fig. 1, this preference for high-resolution images comes at the cost of increased GPU memory consumption and training expenses.

In this work, we propose a novel parameter and memory efficient transfer learning method, namely Fine-grained Prompt Tuning (FPT). FPT aims to enhance the effectiveness of PETL specifically for medical images in high-resolution input contexts by addressing two main challenges:

1) How to efficiently extract fine-grained information from high-resolution images? Existing PETL methods involve training a subset of parameters within the large-scale pre-trained model (LPM) [9,10]. In contrast, FPT utilizes a lightweight additive network inspired by the concept of a side network [23,29]. This learnable side network is introduced outside the LPM, eliminating

the need for back-propagation through the LPM. However, training the side network can still be computationally expensive with high-resolution input images due to the long input sequence. FPT addresses this concern by strategically introducing asymmetric input resolution and employing important token selection to significantly reduce the length of the input sequence for the learnable network.

2) How to effectively adapt pre-trained knowledge from LPMs? LPMs are primarily pre-trained on natural image datasets like ImageNet [21]. To effectively adapt pre-trained knowledge from LPMs of domains outside of medical images, FPT introduces the concept of fine-grained prompts and a Fine-grained Fusion Module (FFM) as bridging components. Fine-grained prompts are a small set of learnable embeddings that summarize pre-trained knowledge from the LPM through the FFM. These prompts are then prepended to the intermediate layers of the side network to convey fine-grained information by integrating them into the forward propagation.

Our main contributions are summarized as follows:

1. We present a novel PETL method, namely Fine-grained Prompt Tuning (FPT), for medical image classification in high-resolution contexts. Asymmetric input and important token selection are proposed to improve memory efficiency. Fine-grained prompts and fine-grained fusion module are introduced to adapt pre-trained knowledge effectively and efficiently. Our code is available online.
2. To the best of our knowledge, this is the first work to enhance the efficiency of PETL in high-resolution input settings, which is particularly significant in the field of medical image analysis.
3. We introduce a new metric to evaluate the trade-off between performance and memory efficiency for transfer learning methods.
4. We conduct extensive experiments on four medical image datasets with different modalities. As shown in Fig. 2, FPT achieves the best trade-off between performance and parameter/memory efficiency.

2 Method

2.1 Side Tuning

As illustrated in part (a) of Fig. 3, the FPT framework consists of two networks: a frozen LPM M and a learnable side network S. Unlike other PETL methods that introduce additional learnable parameters within the LPM, in our approach, the entire LPM remains frozen while the side network is kept learnable and separate. We adopt a lightweight architecture design for the side network, which is a scaled-down variant of the LPM. The hidden dimensions of the side network are $1/k$ times that of the LPM, where k represents a reduction factor. To leverage the pre-trained knowledge from the LPM, the side network reuses the intermediate features at each layer of the LPM. Specifically, given two models M and S with L layers, parameterized by θ_M and θ_S respectively, the intermediate activation
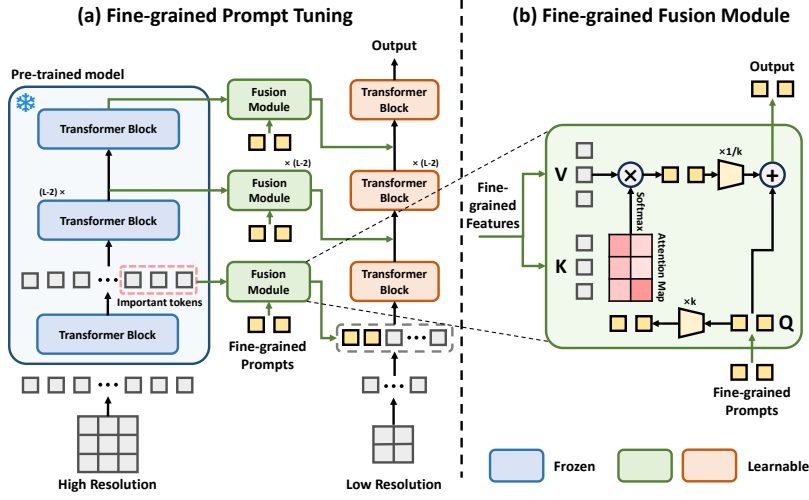
Fig. 3: The proposed FPT framework.

$z_{S,I}^l$ of layer l is obtained as follows:

$$z_M^l = \theta_M^l(z_M^{l-1}), \tag{1}$$

$$z_S^l = \theta_S^l(\mathcal{F}(z_S^{l-1}, z_M^l)), \tag{2}$$

where $\mathcal{F}$ denotes the module that fuses features of M and S, and z_S^L is considered the final output of the framework. The use of the side network not only reduces the number of trainable parameters due to its lightweight design but also helps mitigate memory expenses during the training phase. As shown in part (a) of Fig. 4, the side network eliminates the need for resource-intensive back-propagation from the pre-trained model by excluding any learnable parameters in the forward pass of the heavy LPM model.

2.2 Asymmetric Input

Fine-grained information holds significant importance in the context of medical image analysis and is typically acquired by using high input resolutions. Although training with high-resolution inputs may be infeasible due to high memory consumption, using a LPM solely for inference with high-resolution images remains practical. Therefore, we propose an asymmetric input strategy within the FPT framework. Specifically, given a high-resolution image I, we simply resize it to obtain a low-resolution image I'. Then, the frozen pre-trained model M is provided with the image I, while the learnable side network S is fed low-resolution image I'. Thus, we have the intermediate activation of the side network $z_{S,I'}^l = \theta_S^l(\mathcal{F}(z_{S,I'}^{l-1}, z_{M,I}^l))$.
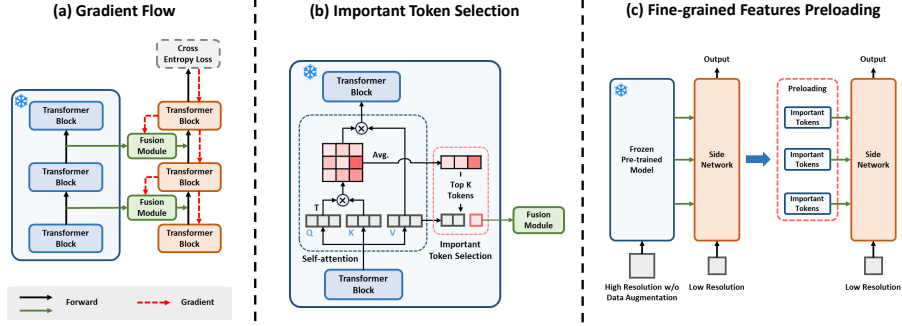
Fig. 4: Components of FPT: (a) The gradient flow of FPT. (b) Important tokens selection mechanism. (c) Fine-grained features preloading.

2.3   Fine-grained Prompts and Fusion Module

In this section, we introduce our proposed fusion module $\mathcal{F}$. As shown in the part (b) of Fig. 3, we utilize the cross-attention mechanism [1] inside the FFM to fuse features from the LPM to the side network. In the context of cross-attention, we reuse the key and value from the self-attention layer of the pre-trained model M. Regarding the query, one approach is to directly reuse the query from the side network S. However, the cross-attention map can be large if the input sequence is long, leading to increased memory consumption. Therefore, we introduce a small set of learnable embeddings z_p, namely fine-grained prompts, into each layer of the fusion modules as the query. Unlike prompt tuning [13,16], which directly uses prompts as part of the input sequence, fine-grained prompts serve as a bridge linking the frozen LPM and the side network. These prompts are concatenated with the intermediate sequence of the side network to join the forward propagation after fusing pre-trained features from the LPM. They are then removed after the layer's forward processing. Specifically, the fusion module is processed as follows:

$$\mathcal{F}(z_S, z_M) = [z_S, f_{out}(\mathrm{CA}(z_p, z_M)) + z_p], \tag{3}$$

$$\mathrm{CA}(z_p, z_M) = \mathrm{Attn}(Q, K, V) = \mathrm{Attn}(f_{in}(z_p), K_M(z_M), V_M(z_M)), \tag{4}$$

where the notations for the input source and the index of the layer are omitted in the formula for simplification. Here, $\mathrm{CA}(\cdot)$ and $\mathrm{Attn}(\cdot)$ denote the cross-attention module and attention function [25] respectively, $[\cdot, \cdot]$ denotes the concatenation operation, f_{in} and f_{out} denote linear layers that align the hidden dimension between the features. The terms K_{M} and V_{M} refer to the key and value mapping matrices within the corresponding self-attention layer of M respectively.

2.4   Important Token Selection

Building on knowledge in the medical image domain, where images of the same modality often display similar anatomical structures and the objects of interest

typically occupy a small proportion of the entire image, we propose a method to further reduce memory consumption, as shown in part (b) of Fig. 4. Specifically, we introduce important token selection, which selects the top $m\%$ tokens with the highest average scores on the self-attention map, considering them as important tokens. Only the features associated with these important tokens are passed to the FFM. This approach significantly reduces the overhead introduced by high-resolution inputs while preserving essential fine-grained information.

2.5   Fine-grained Features Preloading

To further accelerate training procedure, we opt not to use any data augmentation on the input of the frozen LPM. This choice ensures that the intermediate features from the LPM associated with an image remain consistent throughout training. This approach allows us to pre-store these features before training, leading to significant reductions in training costs. Note that while the preloaded features remain fixed, the data augmentation applied to the input of the side network maintains the diversity of training samples.

3   Experiments

3.1   Datasets

We evaluate FPT on four medical datasets with different modalities, including fundus images (messidor-2 [4]), dermoscopic images (ISIC 2018 [3]), mammography (DDSM [15]), and chest X-ray (COVID [22]). The dataset sizes range from 1,748 to 11,527 samples, with classification categories varying from 3 to 7. We use official dataset splits when available. Otherwise, we employ a random partition of 70%/10%/20% for training, validation, and testing, respectively.

3.2   Training and Evaluation Setup

Experiment setup In this study, we utilize a popular variant of Vision Transformers (ViT) [6], specifically ViT-B(ase), which is pre-trained on ImageNet-21K [21]. All methods are fine-tuned for 20 epochs with a mini-batch size of 16. We use the AdamW optimizer [18] with cross-entropy as the loss function for all datasets. To ensure fair comparisons, we conduct a grid search of hyper-parameters for all methods. All methods are run at a resolution of 512×512. In the context of FPT, the high input resolution for the LPM remains 512×512, while the low input resolution for the side network is set to 224×224. All methods were trained employing the same data augmentations as those for low-resolution inputs in FPT. We set the reduction factor k to 8 and use 16 fine-grained prompts with the same hidden dimension as S. For important token selection, we retain the top 20% of important tokens.

Table 1: Comparison results with state-of-the-art PETL methods across four evaluation datasets. 'Params.' denotes the ratio of learnable parameters to the total number of parameters. 'Mem.' denotes the memory usage (MB). The best and second-best results are bold and underlined, respectively.

Method	Computing cost		Fundus	Dermoscopy	Mammography	Chest X-ray	Performance		
	Params.	Mem.	(Messidor2)	(ISIC2018)	(DDSM)	(COVID)	Avg. AUC	PPE	PME
Full fine-tuning	100	24,116	86.87 ± 0.53	96.65 ± 0.29	92.49 ± 0.34	99.85 ± 0.05	93.96	69.54	69.54
Linear probing	0.01	4,364	79.73 ± 0.84	93.37 ± 0.31	80.89 ± 0.39	99.21 ± 0.07	88.30	88.30	82.15
Prompt tuning [13]	0.17	21,530	80.02 ± 2.34	94.20 ± 0.20	82.67 ± 0.34	99.27 ± 0.03	89.04	88.97	67.49
Attention tuning [24]	33.04	21,740	81.87 ± 1.42	94.40 ± 0.40	80.67 ± 2.33	99.58 ± 0.19	89.13	78.74	67.42
Adapter [9]	2.05	20,308	80.77 ± 1.48	$\mathbf{95.96 \pm 0.13}$	80.76 ± 3.38	99.17 ± 0.48	89.16	88.38	68.39
BitFit [28]	0.12	21,330	83.81 ± 1.11	94.84 ± 0.15	$\underline{84.77 \pm 0.54}$	$\mathbf{99.81 \pm 0.03}$	90.81	$\underline{90.76}$	$\underline{68.97}$
LoRA [10]	0.69	21,944	$\mathbf{86.08 \pm 0.95}$	$\underline{95.02 \pm 0.22}$	82.26 ± 4.04	99.69 ± 0.05	$\underline{91.01}$	90.74	68.72
FPT (Ours)	1.81	3,182	$\underline{84.95 \pm 2.01}$	93.88 ± 0.60	$\mathbf{90.52 \pm 0.59}$	$\underline{99.70 \pm 0.30}$	$\mathbf{92.26}$	$\mathbf{91.54}$	$\mathbf{87.42}$

Evaluation metric We use the Area Under the Receiver Operating Characteristic Curve (AUC) to evaluate the classification performance of each dataset. The performance-efficiency (PE) metric [8,17] is a metric to assess the performance and efficiency trade-off. However, PE only considers the impact of the number of learnable parameters. The memory requirement is another crucial factor that significantly influences training expenses. Therefore, we extend the PE metric to include performance-parameter-efficiency (PPE) and performance-memory-efficiency (PME). The PPE formula remains the same as PE, defined as $\text{PPE} = \text{score} * \exp(-\log_{10}(r+1))$, where the score represents the average performance across all datasets, and r is the ratio of learnable parameters to all parameters. Similar to PPE, PME is defined as $\text{PME} = \text{score} * \exp(-\log_{10}(m+1))$, where m is the ratio of the GPU memory requirement of the method to that of fine-tuning the entire LPM.

3.3 Comparisons with State-of-the-art

We compare FPT against full fine-tuning, linear probing, and state-of-the-art (SOTA) PETL approaches. In full fine-tuning, all parameters of the LPM are made learnable during training on the downstream tasks. Linear probing involves solely training the new task-specific head on top of the LPM. Generally, full fine-tuning often represents the upper performance bound for transfer learning, while linear probing represents the lower bound. We also compare FPT against five other popular SOTA PETL methods.

The performance and efficiency of PETL methods are tabulated in Table 1. It can be observed that fine-tuning the entire LPM (full fine-tuning), as the upper bound, achieves the best average AUC of 93.96%, requiring 24,116MB of training GPU memory with a batch size of 16. Although other PETL methods improve transfer learning efficiency by reducing the number of learnable parameters, they are unable to reduce the overhead brought by the high input resolution; all compared PETL methods require at least 20,000MB of memory, and even linear probing requires 4,364MB of memory. In contrast, FPT achieves the second-

Table 2: The performance and efficiency of FPT with different components.

Components	Avg. AUC	Mem.	PME
Sole side network	80.12	17,218	62.95
+ LPM w/ FFM	90.82 (+10.70)	21,070 (+3,852)	69.14 (+6.19)
+ Asymmetric input	92.14 (+1.32)	8,796 (−12,274)	80.50 (+11.36)
+ Token selection	92.26 (+0.12)	4,880 (−3,916)	84.82 (+4.32)
+ Features preloading	92.26 (+0.00)	3,182 (−1,698)	87.42 (+2.60)

Table 3: The performance and efficiency of FPT with different ratios for important token selection.

Ratio	Avg. AUC	Mem.	PME
10%	91.11	**2,760**	<u>86.92</u>
20%	**92.26**	<u>3,182</u>	**87.42**
30%	92.20	3,606	86.79
40%	91.99	4,020	86.03
50%	<u>92.21</u>	4,424	85.71
100%	92.14	8,796	80.50

best AUC of 92.26%, while significantly reducing the memory requirement to only 3,182MB (13% of full fine-tuning). Moreover, with a separate network, FPT utilizes only 1.81% learnable parameters compared to full fine-tuning. Therefore, in terms of efficiency, FPT achieves the best PPE and PME, presenting the best trade-off between performance and efficiency. This parameter and memory efficiency positions FPT as a practical and feasible choice for leveraging LPMs in high-resolution contexts.

3.4   Impact of Components

To assess the impact of the proposed components in FPT, we evaluate the performance and efficiency of our framework by incrementally incorporating components. As shown in Table 2, starting with a sole side network, although the side network is lightweight, the long input sequence still consumes a large amount of memory at 17,218MB for training. Employing fusion modules with fine-grained prompts to extract pre-trained knowledge from a frozen LPM notably enhances performance but further increases the memory burden. Then, the introduction of asymmetric input significantly lowers memory usage by 58% through decreasing the resolution of the input for the side network. It is worth noting that decreasing the input resolution for the side network enhances performance because the fine-grained information provided by the fine-grained prompts is sufficient, and smaller inputs for the side network reduce redundancy in features. Finally, by applying important token selection and preloading techniques, FPT further lowers the memory requirement by 64% without any loss of performance.

3.5   Impact of Important Token Selection Ratio

We evaluate the impact of different ratios of important tokens on FPT. As shown in Table 3, we observe that the classification performance remains similar when the ratio is between 20% and 50%. We then notice that 20% of the content of an image is sufficient for diagnosis in the tasks and modalities considered. Regarding efficiency, memory requirements increase as the ratio increases, leading to lower PME. Therefore, the ratio for important token selection is set to 20%.

4   Conclusion

In this paper, we introduce a novel PETL method, namely Fine-grained Prompt Tuning (FPT), for medical image classification. FPT significantly reduces memory requirements, particularly in the high-resolution context that commonly used in medical image analysis. To address the challenge of high memory requirement, we first adopt the design of side tuning and enhance it with an asymmetric input strategy. We then introduce fine-grained prompts and the fine-grained fusion module to allow effective adaptation of pre-trained knowledge from an out-of-domain LPM that takes images of a different scale as input. To further reduce memory requirement, important token selection and the preloading of pre-trained features are applied. By integrating these components, our PETL method achieves superior performance across four medical datasets while maintaining the best parameter and memory efficiency.

Acknowledgments. This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (62071210); the Shenzhen Science and Technology Program (RCYX202 10609103056042); the Shenzhen Science and Technology Innovation Committee (KXFZ 20C20122117340001); the Guangdong Basic and Applied Basic Research (2021A151522 0131).

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

References

1. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 357–366 (2021)
2. Chen, Z., Guo, X., Woo, P.Y., Yuan, Y.: Super-resolution enhanced medical image diagnosis with sample affinity interaction. IEEE Transactions on Medical Imaging **40**(5), 1377–1389 (2021)
3. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
4. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al.: Feedback on a publicly distributed image database: the messidor database. Image Analysis & Stereology **33**(3), 231–234 (2014)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. 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. arXiv preprint arXiv:2010.11929 (2020)

7. Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S.A., Hospedales, T.: Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. arXiv preprint arXiv:2305.08252 (2023)

8. He, X., Li, C., Zhang, P., Yang, J., Wang, X.E.: Parameter-efficient model adaptation for vision transformers. arXiv preprint arXiv:2203.16329 (2022)

9. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)

10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

11. Huang, Y., Lin, L., Cheng, P., Lyu, J., Tam, R., Tang, X.: Identifying the key components in resnet-50 for diabetic retinopathy grading from fundus images: a systematic investigation. Diagnostics **13**(10), 1664 (2023)

12. Huang, Y., Lyu, J., Cheng, P., Tam, R., Tang, X.: Ssit: Saliency-guided self-supervised image transformer for diabetic retinopathy grading. IEEE Journal of Biomedical and Health Informatics (2024)

13. 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)

14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

15. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data **4**(1), 1–9 (2017)

16. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)

17. Li, C., Liu, H., Li, L., Zhang, P., Aneja, J., Yang, J., Jin, P., Hu, H., Liu, Z., Lee, Y.J., et al.: Elevater: A benchmark and toolkit for evaluating language-augmented visual models. Advances in Neural Information Processing Systems **35**, 9287–9301 (2022)

18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

19. 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: International conference on machine learning. pp. 8748–8763. PMLR (2021)

20. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

21. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021)

22. Siddhartha, M.: Covid cxr image dataset (research) (2021), https://www.kaggle.com/datasets/sid321axn/covid-cxr-image-dataset-research

23. Sung, Y.L., Cho, J., Bansal, M.: Lst: Ladder side-tuning for parameter and memory efficient transfer learning. Advances in Neural Information Processing Systems **35**, 12991–13005 (2022)

24. Touvron, H., Cord, M., El-Nouby, A., Verbeek, J., Jégou, H.: Three things everyone should know about vision transformers. In: European Conference on Computer Vision. pp. 497–515. Springer (2022)

25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
26. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**(1), 1–40 (2016)
27. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
28. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021)
29. Zhang, J.O., Sax, A., Zamir, A., Guibas, L., Malik, J.: Side-tuning: a baseline for network adaptation via additive side networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 698–714. Springer (2020)
30. Zhang, J., Kapse, S., Ma, K., Prasanna, P., Saltz, J., Vakalopoulou, M., Samaras, D.: Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 624–634. Springer (2023)