LKA: Large Kernel Adapter for Enhanced Medical Image Classification

Ziquan Zhu[1*], Si-Yuan Lu[2*], Tianjin Huang[3†], Lu Liu[3†], and Zhe Liu[4]

[1] University of Leicester, Leicester, UK
[2] Nanjing University of Posts and Telecommunications, Nnajing, China
[3] University of Exeter, Exeter, UK
[4] Jiangsu University, Zhenjiang, China

Abstract. Despite the notable success of current Parameter-Efficient Fine-Tuning (PEFT) methods across various domains, their effectiveness on medical datasets falls short of expectations. This limitation arises from two key factors: (1) medical images exhibit extensive anatomical variation and low contrast, necessitating a large receptive field to capture critical features, and (2) existing PEFT methods do not explicitly address the enhancement of receptive fields. To overcome these challenges, we propose the Large Kernel Adapter (`LKA`), designed to expand the receptive field while maintaining parameter efficiency. The proposed `LKA` consists of three key components: down-projection, channel-wise large kernel convolution, and up-projection. Through extensive experiments on various datasets and pre-trained models, we demonstrate that the incorporation of a larger kernel size is pivotal in enhancing the adaptation of pre-trained models for medical image analysis. Our proposed `LKA` outperforms 11 commonly used PEFT methods, surpassing the state-of-the-art by 3.5% in top-1 accuracy across five medical datasets. The code is available at: `https://github.com/misswayguy/LKA`.

Keywords: Medical Image Classification · Parameter-Efficient Fine-Tuning · Large Kernel Convolution.

1 Introduction

The *pre-train fine-tune* paradigm [1] has emerged as a powerful and effective strategy, demonstrating significant success in the domain of medical image [2]. Under this strategy, pre-training is usually conducted on out-of-domain non-medical images, followed by fine-tuning on in-domain medical images tailored to the specific task. Significant progress has been achieved in pre-trained models, with efforts focused on scaling these models to billions and even trillions of parameters [3]. The substantial size of large pre-trained models poses a considerable computational burden for fine-tuning in downstream tasks.

* Equal contribution.
† Corresponding author: Tianjin Huang and Lu Liu (`t.huang2, l.liu3@exeter.ac.uk`).

Various parameter-efficient fine-tuning methods (PEFT), such as prompt tuning [4],adapters [5], and low-rank adaptation (LoRA) [6], have been developed to address the fine-tuning challenge. LoRA reduces the number of trainable parameters by learning low-rank updates to the model's weights, while adapters introduce small, additional modules into the model that can be fine-tuned with minimal computational overhead. Prompt tuning optimizes input prompts to better guide the model predictions without altering the core model parameters. These PEFT methods mitigate computational burden associated with fine-tuning, enabling more efficient adaptation of large models to specific tasks. However, the direct application of these PEFT methods for adapting large pre-trained models to medical imaging tasks may fall short of expectations due to two critical factors: (1) *The characteristics of medical images differ significantly from the training corpora of large pre-trained models*, which are typically based on non-medical images. Medical images often exhibit extensive anatomical variation and low contrast, making them challenging to analyze. Capturing these intricate details effectively requires models with a large receptive field, as smaller receptive fields may fail to fully capture the critical features necessary for accurate interpretation [7]. (2) *Existing PEFT methods including LoRA, adapters, and prompt tuning, are not explicitly designed to capture information from large receptive fields.* While these techniques are effective in reducing the computational burden, they do not explicitly account for the need to capture extensive spatial context, which is crucial for accurately interpreting the complex anatomical structures and subtle variations inherent in medical images.

To address the aforementioned limitations, we begin by investigating the effectiveness of expanding a large receptive field within the adapter framework via the use of large kernel convolutions . Specifically, we employ a channel-wise large kernel convolution [8] after the down-projection layer. Through a comprehensive experimental analysis with varying kernel sizes with adapters, we find that larger kernel convolutions are essential for effectively adapting large pre-trained models to medical imaging domain. Building on this insight, we introduce the Large Kernel Adapter (`LKA`), which comprises a down-projection layer, an activation layer, a channel-wise large kernel convolution layer, and an up-projection layer. The proposed `LKA` can be easily integrated into popular architectures such as Swin [9], ConvNeXt [10], ViT [11], and so on.

In summary, our main contributions are as follows:

- ⋆ We demonstrate that integrating large kernel convolution within adapters significantly expands the effective receptive field (ERF), which is crucial for adapting large pre-trained models to the medical imaging domain. Moreover, our findings indicate that among various methods for expanding the ERF, integrating large kernel convolutions is the most effective.
- ⋆ Leveraging large kernel convolution, we propose the `LKA`, which consists of a down-projection, a channel-wise large kernel convolution, and an up-

We focus on adapters because prompt tuning and LoRA are not inherently compatible with large kernel convolutions.

projection. The channel-wise large kernel convolution is designed to expand ERF, enhancing the ability to capture complex long-range spatial context.

⋆ Extensive experiments demonstrate that the proposed LKA achieves superior performance across various pre-trained model backbones and sizes. Remarkably, compared to 11 other state-of-the-art PEFT methods, our LKA improves the top-1 accuracy by 3.5% on five medical datasets.

⋆ Ablation studies find that an oversized kernel convolution may degrade performance and that integrating the LKA adapter in parallel with the transformation block is more effective than sequential integrations.

2 LKA: Large Kernel Adapter

The current PEFT methods exhibit certain limitations when applied to adapting large pre-trained models for medical imaging tasks, primarily due to their insufficient consideration of large receptive fields, which are crucial for accurately capturing the complex anatomical structures and subtle variations in medical images [12,10]. To address this challenge, we propose the Large Kernel Adapter (LKA). Structurally similar to the adapter [5], our LKA further introduces channel-wise convolution with large kernel size to significantly enhance the model's receptive field, as illustrated in Figure 1. In the LKA, the down-projection with parameters $W_{\text{down}} \in \mathbb{R}^{d \times \hat{d}}$ and the up-projection with parameters $W_{\text{up}} \in \mathbb{R}^{\hat{d} \times d}$ to limit the number of trainable parameters, where $\hat{d}$ is the bottleneck with and satisfies $\hat{d} \ll d$. We introduced a large kernel convolution to expand the receptive field and employed channel-wise convolution to ensure the efficiency. The formula for our LKA can be expressed as:

$$\text{LKA-Conv}(x) = \text{DWConv}_{k \times k}(x) \tag{1}$$

$$x_{\texttt{LKA}} = W_{\text{up}} \cdot (\text{GeLU}(\text{LKA-Conv}(W_{\text{down}} \cdot x)) + x, \tag{2}$$

where LKA-Conv$(\cdot)$ denotes the channel-wise large kernel convolution for enhancing receptive field with kernel size k and GeLU$(\cdot)$ is the activation function used in the LKA.

A key question surrounding LKA is its optimal placement within pretrained models. Our empirical analysis, presented in Table 4, reveals that positioning LKA in parallel with both the MSA and FFN modules (as illustrated in Figure 1) yields the best performance. Formally, the integration of LKA within the Transformer block can be expressed as follows:

$$x_l^* = \text{MSA}(\text{LN}(x_{l-1})) + \texttt{LKA}(\text{LN}(x_{l-1})), \tag{3}$$

$$x_l = \text{MLP}(\text{LN}(x_l^*)) + \texttt{LKA}(\text{LN}(x_l^*)), \tag{4}$$

where LKA$(\cdot)$ has been explained in the previous subsection.

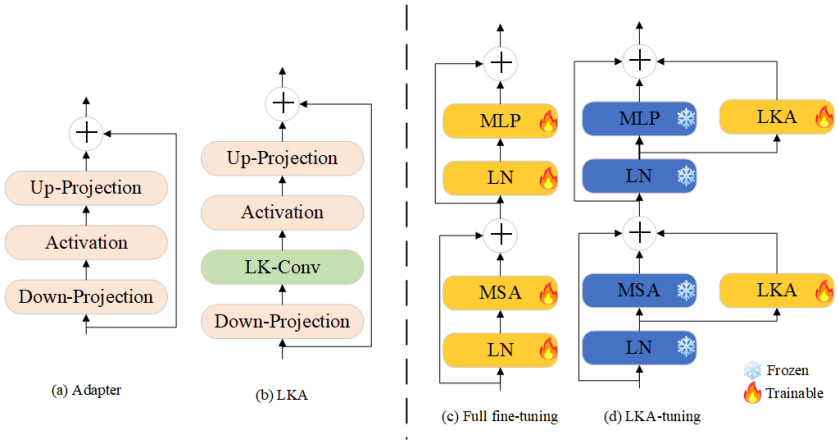


Fig. 1. The structures of Adapter(a) and `LKA`(b), full fine-tuning (c) and `LKA`-tuning (d). During training, only `LKA`s are trainable, while all other layers remain frozen.

3 Experiments and Results

3.1 Experiment Settings

To demonstrate the effectiveness of our proposed `LKA` method, we conduct experiments using 3 popular pre-trained models across 5 medical datasets and report the the top-1 accuracy in all experiments. Additionally, we benchmark our results against 11 PEFT methods to provide a comprehensive comparison. **Datasets.** We evaluate the `LKA` on five available medical datasets, which are Blood cell dataset [17] which contains microscopic images of blood cells, BUSI dataset [18] with 780 breast ultrasound images covering benign, malignant, and normal cases, Brain tumor dataset [19] which is brain MRI images across three tumor categories: glioma, meningioma, and pituitary tumor, Tuberculosis (TB) dataset [20] which features chest X-ray images labeled as TB-positive or normal, and Covid-19 dataset[21]which comprises chest X-ray images from patients with Covid-19, pneumonia, and normal. For all datasets, we follow an 80% training and 20% testing split. **Pre-trained Models.** The proposed `LKA` is tested across various popular architectures, primarily Swin [9], ConvNeXt [10], and ViT [11]. **Implementation Details.** During training, we directly load the pre-trained weights from upstream tasks for the original networks and keep them frozen during the fine-tuning process. In contrast, for our `LKA`, the weights are updated throughout the training. The epoch is set to 100. The optimizer was AdamW with a cosine learning rate scheduler. **Baselines.** We compare our proposed method with other state-of-the-art methods. (**I**) Traditional Fine-tuning: Linear probing [22], Full fine-tuning [23]; (**II**) Adapter-tuning: Adapter [5], ST-Adapter [13], Convpass [16], CIAT [24], AIM [14], LOSSLESS ADAPTATION [25], RepAdapter [26], Adapterformer [27]; (**III**) Other PEFT Methods: Bitfit [28], VPT [4], LoRA [6].

Table 1. Test accuracy of LKA with varying kernel size among different pre-trained models.

Pretrained Models	Kernel Size	Cell	BUSI	Brain	TB	Covid	Average
Swin-L [9]	None	0.914	0.843	0.920	0.987	0.937	0.920
	3×3	0.926	0.874	0.952	0.992	0.956	0.940
	5×5	0.940	0.890	0.960	0.995	0.960	0.949
	7×7	**0.945**	**0.902**	**0.968**	**0.997**	**0.974**	**0.957**
ConvNeXt-L [10]	None	0.882	0.865	0.912	0.980	0.911	0.910
	3×3	0.918	0.896	0.937	0.990	0.935	0.935
	5×5	0.928	0.902	0.945	0.994	0.952	0.944
	7×7	**0.941**	**0.911**	**0.952**	**0.996**	**0.968**	**0.954**
ViT-L [11]	None	0.779	0.803	0.896	0.944	0.894	0.863
	3×3	0.869	0.824	0.918	0.972	0.940	0.905
	5×5	0.880	0.856	0.927	0.980	0.949	0.918
	7×7	**0.884**	**0.875**	**0.938**	**0.987**	**0.958**	**0.928**

3.2 Results and Discussion

Large Kernel is Crucial for effectively Applying the Vanilla Adapter in Medical Imaging. To demonstrate the critical role of large kernels in adapting large pre-trained models for the medical imaging domain, We conduct a comparative analysis by testing LKA with varying kernel sizes, ranging from 3×3 to 7×7. This evaluation is performed across three pre-trained architectures and five medical datasets. Additionally, we provide a visualization of the effective receptive fields (ERFs) corresponding to each kernel configuration.

❶ **Integrating Large Kernel Convolution within Adapters Achieves Better Performance.** Table 1 presents the performance of various pre-trained models—Swin-L [9], ConvNeXt-L [10], and ViT-L [11] on five medical datasets fine-tuned by LKA with different kernel sizes as well as the vanilla adapter (indicated as "None"). The results demonstrate that incorporating a convolution within the adapter consistently enhances performance, with larger kernel sizes yielding significantly better results than smaller ones. This trend holds across different pre-trained models and various medical datasets. Notably, When using a kernel size of 7×7, the LKA consistently outperforms the vanilla adapter by an average margin of 4.9% across all datasets.

❷ **Integrating Large Kernel Convolution within Adapters Expands ERF.** The concept of ERF is crucial in computer vision, particularly in understanding how neural networks process visual information [29]. Following [30,12], we sample and resize 50 images from the validation set to 1024×1024, and measure the contribution of each pixel on input images to the central point of the feature map generated in the last layer. The contribution scores are further accumulated and projected to a 1024×1024 matrix. The visualization is shown in Figure 2. We find that integrating an adapter with larger kernel convolutions (i.e. LKA with larger kernel size) leads to an expanded ERF. This finding indicates that the combination of an adapter and larger kernel convolutions enhances the model's ability to capture long-range spatial contextual information.

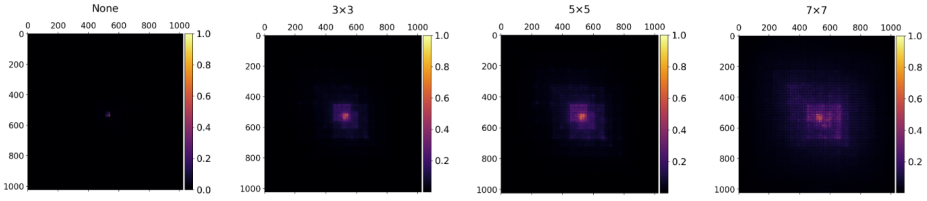
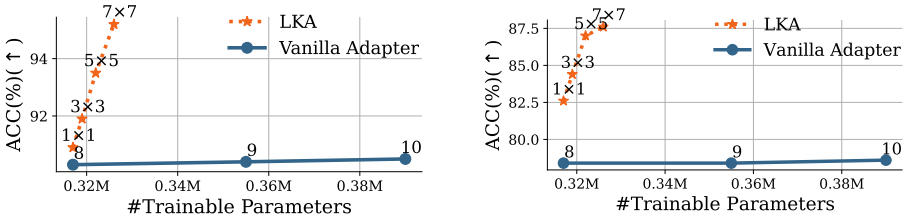


Fig. 2. Effective receptive fields (ERFs) for `LKA` with various kernel sizes based on pre-trained Swin-T.



(a) Covid (b) BUSI

Fig. 3. Comparison of `LKA` and Vanilla Adapter with varying #Trainable Parameters based on pre-trained Swin-T and two datasets: Covid (a) and BUSI (b).

❸ Large Kernel Matters Instead of #Trainable Parameters. When increasing the kernel size in convolutional layers, it inevitably leads to an increase in the number of trainable parameters. To substantiate our claim that large kernel convolutions are critical for effectively adapting pre-trained models in medical imaging, we conduct experiments using a vanilla adapter with an enlarged bottleneck width and compare its performance with the `LKA` with varying kernel sizes. The results in Figure 3 demonstrate that increasing the kernel size from 1×1 to 7×7 significantly improves accuracy across different datasets. Specifically, in the Covid dataset, accuracy improves from 90.3% to 95.2%, while in the BUSI dataset, the accuracy increases from 80.3% to 87.5%, despite only a marginal increase in trainable parameters. These results highlight the greater impact of kernel size over simply increasing trainable parameters, reinforcing the effectiveness of large-kernel convolutions in medical image adaptation.

Achieving State-of-the-Art. To demonstrate the superiority of the `LKA` over other PEFT methods, we compared `LKA` with the commonly used 11 PEFT methods based on pre-trained Swin-L model on the five medical datasets. The results in Table 2 present that the `LKA` outperforms all 11 PEFT methods across all five medical datasets. On average, our `LKA` approach achieves the 3.5% higher top-1 accuracy compared to these advanced PEFT methods. Notably, `LKA` even surpasses the results of full fine-tuning in most cases. The consistent superior performance of `LKA` across multiple datasets underscores the effectiveness of leveraging larger receptive fields to enhance top-1 accuracy in medical tasks.

Table 2. Comparison of LKA with other baselines on five medical datasets based on pre-trained Swin-L. Bold indicates the highest value achieved among the PEFT methods.

Kernel Size	#Trainable Parameters	Cell	BUSI	Brain	TB	Covid	Average
Full fine-tuning [23]	195M	0.953	0.894	0.967	0.995	0.974	0.957
Linear probing [22]	0.006M	0.857	0.871	0.822	0.939	0.934	0.885
Bitfit [28]	0.313M	0.924	0.881	0.920	0.993	0.946	0.933
Adapterformer [27]	0.320M	0.939	0.884	0.948	0.991	0.945	0.941
LOSSLESS ADAPTATION [25]	0.320M	0.924	0.887	0.952	0.991	0.948	0.940
ST-Adapter [13]	0.334M	0.777	0.713	0.867	0.933	0.833	0.825
RepAdapter [26]	0.486M	0.924	0.881	0.952	0.988	0.944	0.938
LoRA [6]	0.578M	0.940	0.887	0.952	0.993	0.943	0.943
Adapter [5]	0.633M	0.914	0.843	0.920	0.987	0.937	0.92
Convpass [16]	0.661M	0.921	0.884	0.933	0.993	0.955	0.937
AIM [14]	0.947M	0.849	0.845	0.877	0.972	0.937	0.896
CIAT [24]	0.966M	0.936	0.884	0.952	0.990	0.952	0.943
VPT [4]	1.052M	0.917	0.852	0.933	0.987	0.926	0.923
LKA(Ours)	0.652M	**0.945**	**0.902**	**0.968**	**0.997**	**0.974**	**0.957**

Table 3. Test accuracy of different kernel sizes of the LKA based on pre-trained Swin-T.

Kernel Size	#Trainable Parameters	Cell	BUSI	Brain	TB	Covid	Average
3×3	0.319M	0.875	0.844	0.900	0.987	0.919	0.905
5×5	0.322M	0.880	0.870	0.936	0.990	0.935	0.922
7×7	0.326M	**0.885**	**0.876**	**0.944**	**0.992**	**0.952**	**0.930**
9×9	0.340M	0.870	0.863	0.933	0.951	0.936	0.918
11×11	0.345M	0.865	0.860	0.933	0.989	0.936	0.918
31×31	0.502M	0.870	0.861	0.933	0.985	0.929	0.916

Table 4. Test accuracy of various integration positions for LKAs based on pre-trained Swin-T.

Position	Cell	BUSI	Brain	TB	Covid	Average
(a)	**0.885**	**0.876**	**0.944**	**0.992**	**0.952**	**0.930**
(b)	0.877	0.861	0.933	0.986	0.932	0.919
(c)	0.882	0.870	0.933	0.989	0.937	0.922

3.3 Ablation Studies and Extra Analysis

Oversized Kernels May Hurt. The observation above indicates that integrating large kernel convolution within adapter expands the Effective Receptive Field (ERF), leading to improved performance. However, it remains an intriguing question whether using even larger kernel convolution could offer additional benefits. Therefore, we conduct experiments with increasing kernel size to 31×31 based on the pre-trained Swin-T. The results in Table 3 indicate a kernel size is 7×7 achieve the best performance. The kernel size significantly affects model performance, as small kernels fail to capture global information, while excessively large kernels may miss key regions, both leading to suboptimal results.

Position Matters. Existing adapters [5,13,14] are commonly integrated into Transformer blocks using a sequential approach, a method originally developed for natural language processing. However, given the fundamental differences between vision and language domains, recent research has explored alternative placement strategies. For example,[15,16] demonstrated that inserting adapters solely after the feed-forward network (FFN) module can also be effective. Al-

Table 5. Test accuracy of different enlarging receptive field recipes into the LKA based on pre-trained Swin-T. "n" and "d" represent the number of convolutional layers and the dilation rate in dilated convolution.

Method	Cell	BUSI	Brain	TB	Covid	Average
$3 \times 3_{(n=3)}$/CW-Conv	0.880	0.865	0.933	0.990	0.950	0.924
$3 \times 3_{(d=3)}$/D-Conv	0.883	0.873	0.933	0.991	0.950	0.926
7×7/CW-Conv	**0.885**	**0.876**	**0.944**	**0.992**	**0.952**	**0.930**

Table 6. Test accuracy of different bottleneck widths based on pre-trained Swin-T.

Bottleneck Width	#Trainable Parameters	Cell	BUSI	Brain	TB	Covid	Average
4	0.174M	0.868	0.855	0.936	0.982	0.940	0.916
8	0.326M	**0.885**	**0.876**	**0.944**	**0.992**	**0.952**	**0.930**
12	0.479M	0.884	0.873	0.944	0.993	0.948	0.928
16	0.631M	0.880	0.868	0.933	0.946	0.925	0.910
32	1.240M	0.880	0.868	0.936	0.988	0.945	0.923

though these strategies have shown promise, the unique requirements of medical image analysis warrant a more nuanced investigation into the optimal positioning of LKAs within Transformer blocks.

To this end, we evaluate three placement strategies: (a) positioning LKAs in parallel with the entire Transformer block,alongside the MLP and MSA modules; (b) inserting LKAs sequentially before the residual connections in the MSA and FFN modules; and (c) placing LKAs sequentially after the residual connections in these modules. As shown in Table 4, the parallel configuration (a) consistently delivers superior performance compared to the other two strategies.

Comparing Various Enlarging Receptive Field Recipes. We explores the effectiveness of integrating different convolutions into the LKA to expand the receptive field: channel-wise convolution (CW-Conv) and dilated convolution (D-Conv) [31]. Given that the LKA performs best when the kernel size is 7, we specifically examine three recipes: a single 7×7 CW-Conv, a single 3×3 D-Conv with the dilation rate of 3, and stacking three 3×3 CW-Convs. Results in Table 5 show that the 7×7 CW-Conv performed best by effectively integrating global information with a single large kernel.

Effects of Bottleneck Widths. To determine the optimal bottleneck width $\hat{d}$, we conduct a series of experiments comparing different bottleneck widths. As shown in Table 6, the bottleneck width of 8 achieves the best top-1 accuracy across five medical datasets, marking the point of performance saturation. Interestingly, further increasing the bottleneck width results in a decline in accuracy.

Parameter Efficiency Analysis. The proposed LKA not only significantly outperforms the vanilla adapter across five medical datasets, but also does so with minimal increase in trainable parameters. The total number of trainable parameters of LKA is $2 \times d \times \hat{d} + (k^2 + 2) \times \hat{d} + d$, where k is the kernel size, while the vanilla adapter has $2 \times d \times \hat{d} + \hat{d} + d$. Compared to the vanilla adapter, LKA only

introduces the additional $(k^2 + 1) \times \hat{d}$ parameters. Since $\hat{d}$ and k are relatively small values, far less than d, the added parameters can be considered negligible.

4 Conclusion

In this paper, we primarily focus on the challenges of suboptimal performance in current PEFT methods when adapting pre-trained models to medical imaging tasks, which is primarily due to the inability to provide large receptive fields. To overcome this limitation, we propose the Large Kernel Adapter (LKA), which enhances the receptive field by integrating channel-wise large kernel convolution into the adapter while maintaining parameter efficiency. Through extensive experiments, we demonstrate that the LKA consistently delivers significant performance improvements across various pre-trained models and medical imaging datasets, highlighting its broad applicability and effectiveness. Compared to 11 other advanced PEFT methods, the LKA achieves superior performance across five medical datasets, further validating its effectiveness in medical imaging tasks. We believe our findings hold potential clinical value by enabling more effective adaptation of foundation models to diverse medical imaging tasks.

Acknowledgments. This work is partially supported by the SLAIDER project.

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

References

1. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), pp. 1–35 (2023). ACM New York, NY.
2. Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T.: Transfer learning for medical image classification: A literature review. BMC Medical Imaging **22**(1), pp. 69 (2022). Springer.
3. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020).
4. Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.-N.: Visual prompt tuning. European Conference on Computer Vision, pp. 709–727 (2022). Springer.
5. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. International Conference on Machine Learning, pp. 2790–2799 (2019). PMLR.
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (202
7. Li, H., Nan, Y., Del Ser, J., Yang, G.: Large-kernel attention for 3D medical image segmentation. Cognitive Computation, pp. 1–15 (2023). Springer.

8. Gao, H., Wang, Z., Ji, S.: ChannelNets: Compact and efficient convolutional neural networks via channel-wise convolutions. Advances in Neural Information Processing Systems **31** (2018).
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021).
10. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022).
11. 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).
12. Liu, S., Chen, T., Chen, X., Chen, X., Xiao, Q., Wu, B., Kärkkäinen, T., Pechenizkiy, M., Mocanu, D., Wang, Z.: More ConvNets in the 2020s: Scaling up kernels beyond 51×51 using sparsity. arXiv preprint arXiv:2207.03620 (2022).
13. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: ST-Adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems **35**, pp. 26462–26477 (2022).
14. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: AIM: Adapting image models for efficient video action recognition. arXiv preprint arXiv:2302.03024 (2023).
15. Bapna, A., Arivazhagan, N., Firat, O.: Simple, scalable adaptation for neural machine translation. arXiv preprint arXiv:1909.08478 (2019).
16. Jie, S., Deng, Z.-H.: Convolutional bypasses are better vision transformer adapters. arXiv preprint arXiv:2207.07039 (2022).
17. Girdhar, A., Kapur, H., Kumar, V.: Classification of white blood cell using convolution neural network. Biomedical Signal Processing and Control **71**, pp. 103156 (2022). Elsevier.
18. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, pp. 104863 (2020). Elsevier.
19. Summers, D.: Harvard Whole Brain Atlas: www.med.harvard.edu/AANLIB/home.html. Journal of Neurology, Neurosurgery & Psychiatry **74**(3), pp. 288–288 (2003). BMJ Publishing Group Ltd.
20. Rahman, T., Khandakar, A., Kadir, M.A., Islam, K.R., Islam, K.F., Mazhar, R., Hamid, T., Islam, M.T., Kashem, S., Mahbub, Z.B., et al.: Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access **8**, pp. 191586–191601 (2020). IEEE.
21. Choy, S.P., et al.: Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. NPJ Digital Medicine **6**(1) (2023).
22. Liang, W., Yuan, Y., Ding, H., Luo, X., Lin, W., Jia, D., Zhang, Z., Zhang, C., Hu, H.: Expediting large-scale vision transformer for dense prediction without fine-tuning. Advances in Neural Information Processing Systems **35**, pp. 35462–35477 (2022).
23. Sandler, M., Zhmoginov, A., Vladymyrov, M., Jackson, A.: Fine-tuning image transformers using learnable memory. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12155–12164 (2022).
24. Zhu, Y., Feng, J., Zhao, C., Wang, M., Li, L.: Counter-interference adapter for multilingual machine translation. arXiv preprint arXiv:2104.08154 (2021).
25. Sharma, M., Fantacci, C., Zhou, Y., Koppula, S., Heess, N., Scholz, J., Aytar, Y.: Lossless adaptation of pretrained vision models for robotic manipulation. arXiv preprint arXiv:2304.06600 (2023).

26. Luo, G., Huang, M., Zhou, Y., Sun, X., Jiang, G., Wang, Z., Ji, R.: Towards efficient visual adaptation via structural re-parameterization. arXiv preprint arXiv:2302.08106 (2023).
27. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: AdaptFormer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, pp. 16664–16678 (2022).
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. Araujo, A., Norris, W., Sim, J.: Computing receptive fields of convolutional neural networks. Distill **4**(11), pp. e21 (2019).
30. Ding, X., Chen, H., Zhang, X., Han, J., Ding, G.: RepMLPNet: Hierarchical vision MLP with re-parameterized locality. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 578–587 (2022).
31. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015).