Supplementary Materials

A Basis sampling strategies

As discussed in Sec. 3.1 in the main paper, we need weights bases for the sampling. There are different ways to sample the basis weights for weights generation. For instance, the number of group G can be chosen as large as N. We do not take every single weight vector, but a group of them, from A as the basis. This is because generating the weights group-wisely could better utilize the spatial continuity of the inputs. The number of basis weights for the generation would affect the flexibility of the generated weights. Using more basis would give greater freedom but may also hurt the efficiency of the method. Here, we use two bases for the best trade-off.

B Architecture and performance

MedBlock illustration. In classification, the input image resolution is often set to 224×224. Suppose the patch size is 16×16. The number of token, *i.e.*, the spatial size, should be $14 \times 14 = 196$. Such a large value will result in a large number of weight basis, consuming huge amount of memory. To mitigate this issue, we adopt a similar strategy to ViP [5] and encode the spatial information along the height and width dimension separately with the permutation strategy. Different from ViP, we further decompose the layer into two consecutive layers with a bottleneck structure. A diagrammatic illustration of our building block can be found in Fig. 2 in the main paper which contains two components for spatial information mixing and channel mixing, respectively. The channel mixing component is a normal MLP which consists of two fully connected layers with a non-linear activation. For spatial information mixing, we use two branches to encode the information along the height and width dimension, respectively, each of which has two AdaFCs. Suppose the input tensor has the shape of $C \times S'$. Without the bottleneck structure, to guarantee the spatial size of the output is still S', our basis sampling strategy should be applied to both the input and output dimensions of the weight matrix. This means the weight matrix would have a shape of $S' \times S'$ and the computation cost will be proportional to $C \cdot S' \cdot S'$, which is quadratic in S'.

Performance of MedMLP on natural image dataset. The results can be found in Tab. A.

C Dataset Description

PACD is a spectrum of disease that is characterized in common by an obstruction to aqueous humor outflow. It may culminate in developing a more visually debilitating form of glaucomatous optic neuropathy. We randomly split a subset of data from the Singapore Chinese Eye Study (SCES), the Singapore Indian Chinese Cohort (ICC), and the Iris Surface Features (ISF), in total 4715 eyes

Networks	Param.	FLOPs	ImageNet (%)	Real (%)
ViP (scaled) [5]	6.7M	1.5B	70.3	78.4
MedMLP-B0 (Ours)	4.9M	0.6B	74.3	81.6
gMLP-tiny16	6.0M	2.7B	76.4	–
MedMLP-B1 (Ours)	8.4M	1.0B	76.2	83.0
Mixer-B/16 [7]	59.0M	11.6B	76.4	82.0
ResMLP-S12 [8]	16.0M	0.8B	76.2	83.5
MedMLP-B2 (Ours)	12.7M	2.1B	78.3	84.8

Table A. Top-1 accuracy comparison of our MedMLP with the recent MLP-like models on ImageNet [4] and ImageNet Real [2] ('Real'). All the models are trained without external data.

Stage i	Operator A_i	Resolution $H_i \times W_i$	#Filters C_i	#Layers L_i
1	PatchEmbed 4x4	224 × 224	32	1
2	AdaFC, e4, h1	56 × 56	42	2
3	AdaFC, e2, h2	28 × 28	56	4
4	AdaFC, e3, h4	14 × 14	96	4
5	AdaFC, e6, h8	14 × 14	112	4
6	AdaFC, e6, h32	7 × 7	224	4
7	Head & LayerNorm	1 × 1	1000	1

Table B. Architecture definition of AdaMLP-B0 model. We use 'h' to denote the number of heads and 'e' the expansion ratio in channel mixing MLP.

into training, validation, and testing dataset following a ratio of 7:1:2. The other iris fundus photo dataset used for external validation is sub-set of the Singapore Indian Eye Study (SINDI) which contains 250 eyes. With MedMLP, the average accuracy over all datasets is improved from 66% to 79.2%, compared to ResNet50. This shows the superiority on generalization capability of MedMLP.

C.1 Implementation details

We use Pytorch for all model training. For the comparisons with MobileNetV2, we use AdamW [6] optimizer with initial learning rate $1e^{-3}$ and weight decay of 0.05. We train the model for 300 epochs without cutmix and auto-augmentation, which are adapted by previous All MLP networks [7,8] reproduced in the timm [9] library. The reported results of MobileNetV2 are reproduced with the same training settings. When comparing with other SOTA models, we report the results with advanced training recipes with CutMix [10] and RandAug [3] added using same settings as previous methods [5,7,8].

D MedMLP Architecture

The architecture of building blocks of the proposed MedMLP can be found in Fig. 3 in the main paper and Tab. B describes the complete architecture design of MedMLP. Our network takes an image of arbitrary size $n \times n$ as input and uniformly splits it into a sequence of image patches (4×4). All the patches are then mapped into linear embeddings (or called tokens) using a shared linear layer as [7] followed by a layer normalization [1]. We next feed all the tokens into a sequence of Adaptive MLP block to encode both spatial and channel information.

Table C. Top-1 accuracy on ImageNet with different model sizes of MedMLP. Our proposed MedMLP demonstrates outstanding scalability.

Model	Params. (M)	MAdds (B)	Top-1 Acc. (%)
MedMLP-B0	4.9	0.6	74.3
MedMLP-B1	8.4	1.0	76.2 (+1.9)
MedMLP-B2	12.7	2.1	78.3 (+2.1)
MedMLP-B3	25.7	4.1	81.1 (+2.8)

Table D. Impacts of natural image pre-training.

Networks	Param.	Pre-trained	ImageNet (%)	Real (%)
ViP (scaled) [5]	6.7M	1.5B	70.3	78.4
MedMLP-B0 (Ours)	4.9M	0.6B	74.3	81.6

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 2
2. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020) 2
3. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020) 2
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 2
5. Hou, Q., Jiang, Z., Yuan, L., Cheng, M.M., Yan, S., Feng, J.: Vision permutator: A permutable mlp-like architecture for visual recognition. arXiv preprint arXiv:2106.12368 (2021) 1, 2, 3
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 2
7. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al.: Mlp-mixer: An all-mlp architecture for vision. arXiv preprint arXiv:2105.01601 (2021) 2
8. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Joulin, A., Synnaeve, G., Verbeek, J., Jégou, H.: Resmlp: Feedforward networks for image classification with data-efficient training. arXiv preprint arXiv:2105.03404 (2021) 2
9. Wightman, R.: Pytorch image models. `https://github.com/rwightman/pytorch-image-models` (2019). `https://doi.org/10.5281/zenodo.4414861` 2
10. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032 (2019) 2