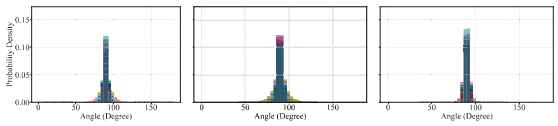
Efficient Adaptation of Pre-trained Vision Transformer underpinned by Approximately Orthogonal Fine-Tuning Strategy

Supplementary Material

A. Angle Distribution

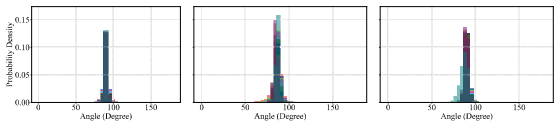
In the weights of the ViT-B/16 pre-trained model $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o, \mathbf{W}_{\text{FC1}}, \mathbf{W}_{\text{FC2}}$. The distribution of pairwise angles between the column vectors in these weight matrices is shown in the Fig. 8. It can be observed that in the ViT model, the matrices are approximately orthogonal.



(a) The degree distribution of pre-trained matrices $\mathbf{W}_q$.

(b) The degree distribution of pre-trained matrices $\mathbf{W}_k$.

(c) The degree distribution of pre-trained matrices $\mathbf{W}_v$.

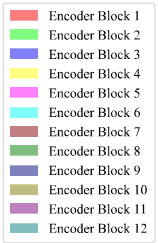


(d) The degree distribution of pre-trained matrices $\mathbf{W}_o$.

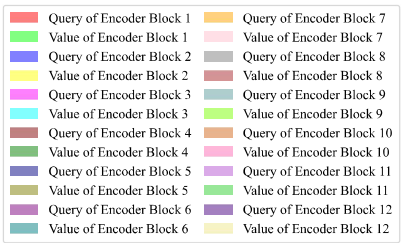
(e) The degree distribution of pre-trained matrices $\mathbf{W}_{\text{FC1}}$.

(f) The degree distribution of pre-trained matrices $\mathbf{W}_{\text{FC2}}$.

Figure 8. Illustration of approximate orthogonality among any two column vectors of weight matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o, \mathbf{W}_{\text{FC1}}, \mathbf{W}_{\text{FC2}}$ in the ViT-B model after training. The histogram represents the distribution of angles between any two column vectors within each weight matrix. Specifically, (a)-(f) represent approximate orthogonality in the pre-trained model.Their legends can be found in Fig. 9.



(a) Legends for Adapter.



(b) Legends for LoRA.

Figure 9. The legends for the bar chart.

Fig. 9 presents the legends for the statistical charts used throughout this paper. Fig. 9(a), specifically designed for charts that do not include the LoRA (Low-Rank Adaptation) method, clearly labels the line colors corresponding to different layers of the model, ensuring accurate differentiation among them. Meanwhile, Fig. 9(b) is tailored for charts that incorporate the LoRA method. It distinctly showcases the line colors for different layers and matrices within those layers, facilitating precise differentiation.

B. Detailed Dataset Statistics

In this section, we provide a comprehensive overview of the visual adaptation classification tasks utilized in our study, presenting the specifics for both the Fine-Grained Visual Classification (FGVC) datasets and the Visual Task Adaptation Benchmark-1k (VTAB-1k) datasets. The dataset splits employed in our experiments follow the protocol established by VPT [11].

B.1. FGVC Datasets

The details of the FGVC datasets used in our study are summarized in Tab. 1. For each dataset, we report the number of classes, as well as the sizes of the training, validation, and test sets. These datasets are characterized by their fine-grained nature, requiring models to distinguish between subtle differences in appearance, making them ideal for evaluating the efficacy of visual adaptation techniques.

B.2. VTAB-1k Datasets

Similarly, the VTAB-1k datasets used in our experiments are detailed in Tab. 2. Again, we provide the number of classes and the sizes of the training, validation, and test sets for each dataset. The VTAB-1k benchmark is designed to assess the generalization capabilities of visual models across a diverse range of tasks, making it a valuable tool for evaluating the robustness and adaptability of our proposed methods.

By adhering to the dataset splits established by VPT [11], we ensure a fair and consistent comparison with existing work, facilitating the reproducibility and validation of our findings.

C. Experimental Details on Large-Scale, Huge-Scale, and Hierarchical ViT Backbones

In this section, we present the comprehensive results of our comparison among ViT-Large, ViT-Huge, and Swin-Base models, as discussed in Section 4. The detailed findings for each model are displayed in Tab. 3, 4, and 5, respectively.

Tab. 3 showcases the performance of the ViT-Large backbone across various tasks within the VTAB-1k benchmark. This table provides a thorough breakdown of the

Table 1. Dataset statistics for FGVC. "*" denotes the train/val split of datasets following the dataset setting in VPT [11].

Dataset	Description	Classes	Train size	Val size	Test size
CUB-200-2011 [31]	Fine-grained bird species recognition	200	5,394*	600*	5,794
NABirds [29]	Fine-grained bird species recognition	555	21,536*	2,393*	24,633
Oxford Flowers [23]	Fine-grained flower species recognition	102	1,020	1,020	6,149
Stanford Dogs [13]	Fine-grained dog species recognition	120	10,800*	1,200*	8,580
Stanford Cars [5]	Fine-grained car classificatio	196	7,329*	815*	8,041

Table 2. Dataset statistics for VTAB-1k [33].

Dataset	Description	Classes	Train size	Val size	Test size
CIFAR-100	Natural	100	800/1,000	200	10,000
Caltech101		102			6,084
DTD		47			1,880
Flowers102		102			6,149
Pets		37			3,669
SVHN		10			26,032
Sun397		397			21,750
Patch Camelyon	Specialized	2	800/1,000	200	32,768
EuroSAT		10			5,400
Resisc45		45			6,300
Retinopathy		5			42,670
Clevr/count	Structured	8	800/1,000	200	15,000
Clevr/distance		6			15,000
DMLab		6			22,735
KITTI/distance		4			711
dSprites/location		16			73,728
dSprites/orientation		16			73,728
SmallNORB/azimuth		18			12,150
SmallNORB/elevation		9			12,150

results, allowing for a detailed analysis of the model's strengths and weaknesses in different contexts.

Similarly, Tab. 4 presents the results for the ViT-Huge backbone. By comparing these results to those of the ViT-Large model, we can gain insights into the benefits and trade-offs associated with scaling up the model size.

Lastly, Tab. 5 outlines the performance of the Swin-Base model, which adopts a hierarchical architecture. This table enables us to assess the effectiveness of hierarchical designs in comparison to the standard ViT architectures.

By examining these detailed results, we can draw meaningful conclusions regarding the performance characteristics of large-scale, huge-scale, and hierarchical ViT backbones. These insights contribute to a deeper understanding of the capabilities and limitations of these models in various visual adaptation tasks.

D. Orthogonality of Operators

In this work, a method is applied to generate multiple orthogonal basis from a vector to form a column of orthogonal matrices. The matrix $\mathbf{Q}$ is derived from the matrix:

$$\begin{bmatrix} \cos\varphi & -x_1\sin\varphi & \cdots & -x_i\sin\varphi \\ x_1\sin\varphi & 1+x_1^2(\cos\varphi-1) & \cdots & x_ix_1(\cos\varphi-1) \\ x_2\sin\varphi & x_1x_2(\cos\varphi-1) & \cdots & x_ix_2(\cos\varphi-1) \\ \vdots & \vdots & \vdots & \vdots \\ x_i\sin\varphi & x_1x_i(\cos\varphi-1) & \cdots & 1+x_i^2(\cos\varphi-1) \\ \vdots & \vdots & \vdots & \vdots \\ x_N\sin\varphi & x_1x_N(\cos\varphi-1) & \cdots & x_ix_N(\cos\varphi-1) \end{bmatrix}, \tag{10}$$

let $\cos\varphi = q_0$ and $\sin\varphi = (\sum_{i=1}^{N}|q_i|^2)^{1/2}$, where $\sum_{i=1}^{N}|q_i|^2 = 1$, that means $x_i = \frac{q_i}{\sin\varphi}(i = 1, 2, ..., N)$,

Table 3. This table is extended from Tab. 3 in Section 4 and describes the detailed experimental results of the performance comparison on VTAB-1k using ViT-Large pre-trained on ImageNet-21k as the backbone. Where Full fine-tuning and Linear probing are only used as controls and are not included in the bold comparison. The best results are shown in **bold**.

Methods \ Datasets	Natural								Specialized					Structed										
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVNH	Sun397	Mean	Camelyon	EuroSAT	Resisc45	Retinopathy	Mean	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Mean	Mean Total	Params.(M)
Full fine-tuning	68.6	84.3	58.6	96.3	86.5	87.5	41.4	74.7	82.6	95.9	82.4	74.2	83.8	55.4	55.0	42.2	74.2	56.8	43.0	28.5	29.7	48.1	65.4	303.4
Linear probing	72.2	86.4	63.6	97.4	85.8	38.1	52.5	70.9	76.9	87.3	66.6	45.4	69.1	28.2	28.0	34.7	54.0	10.6	14.2	14.6	21.9	25.8	51.5	0.05
Adapter [9]	75.3	84.2	54.5	**97.4**	**84.3**	31.3	**52.9**	68.6	75.8	85.1	63.4	69.5	73.5	35.4	**34.1**	30.8	47.1	30.4	23.4	10.8	19.8	29.0	52.9	2.38
Adapter+AOFT*	**79.6**	**89.6**	**63.0**	84.3	73.7	**72.2**	22.2	**70.7**	**77.1**	**86.2**	**71.2**	**73.6**	**77.0**	**68.2**	25.4	**39.3**	**66.8**	**61.3**	**42.4**	**29.9**	**21.8**	**44.4**	**60.9**	0.10
LoRA [10]	75.8	89.9	73.6	99.1	90.8	**83.2**	57.5	81.4	86.0	95.0	83.4	**75.5**	85.0	78.1	60.5	46.7	**81.6**	76.7	51.3	28.0	35.4	57.3	72.0	0.74
LoRA+AOFT*	**78.2**	**95.0**	**74.7**	**99.5**	**92.0**	82.4	**59.2**	**83.3**	**86.7**	**95.1**	**86.0**	75.2	**85.9**	**81.5**	**63.2**	**50.7**	81.0	**86.7**	**53.0**	**28.8**	**43.3**	**60.2**	**74.3**	0.15

Table 4. This table is extended from Tab. 3 in Section 4 and describes the detailed experimental results of the performance comparison on VTAB-1k using ViT-Huge pre-trained on ImageNet-21k as the backbone. Where Full fine-tuning and Linear probing are only used as controls and are not included in the bold comparison. The best results are shown in **bold**.

Methods \ Datasets	Natural								Specialized					Structed										
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVNH	Sun397	Mean	Camelyon	EuroSAT	Resisc45	Retinopathy	Mean	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Mean	Mean Total	Params.(M)
Full fine-tuning	58.7	86.5	55.0	96.5	79.7	87.5	32.5	70.9	83.1	95.5	81.9	73.8	83.6	47.6	53.9	37.8	69.9	53.8	48.6	30.2	25.8	46.0	63.1	630.90
Linear probing	64.3	83.6	65.2	96.2	83.5	39.8	43.0	67.9	78.0	90.5	73.9	73.4	79.0	25.6	24.5	34.8	59.0	9.5	15.6	17.4	22.8	26.1	52.7	0.06
Adapter [9]	**69.4**	84.4	62.7	97.2	84.2	33.6	45.3	68.1	77.3	86.6	70.8	71.1	76.4	28.6	27.5	29.2	55.2	10.0	15.2	11.9	18.6	24.5	51.5	5.78
Adapter+AOFT*	67.9	**91.3**	**69.6**	**98.6**	**88.1**	**79.4**	**49.0**	**77.7**	**80.0**	**95.3**	**78.3**	**73.6**	**81.8**	**31.5**	**31.7**	**39.0**	**71.6**	**40.2**	**22.4**	**23.8**	**36.9**	**37.1**	**61.5**	0.17
LoRA [10]	63.0	89.4	68.1	98.0	87.0	**85.2**	48.7	77.1	82.2	94.3	**83.1**	74.2	83.5	68.6	**65.0**	44.8	76.4	70.8	**48.8**	30.4	38.3	55.4	69.3	1.21
LoRA+AOFT*	**68.9**	**93.0**	**69.9**	**98.7**	**89.1**	80.9	**51.5**	**78.8**	**84.2**	**94.5**	82.1	**74.6**	**83.8**	**74.1**	63.5	**46.5**	**79.3**	**79.9**	48.7	**31.5**	**43.2**	**58.3**	**71.3**	0.20

Table 5. This table is extended from Tab. 4 in Section 4 and describes the detailed experimental results of the performance comparison on VTAB-1k using Swin-Base pre-trained on ImageNet-21k as the backbone. The best results are shown in **bold**.

Methods \ Datasets	Natural								Specialized					Structed										
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVNH	Sun397	Mean	Camelyon	EuroSAT	Resisc45	Retinopathy	Mean	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Mean	Mean Total	Params.(M)
Full fine-tuning	72.2	88.0	71.4	98.3	89.5	89.4	45.1	79.1	86.6	**96.9**	**87.7**	73.6	86.2	75.7	**59.8**	54.6	78.6	79.4	53.6	**34.6**	**40.9**	59.7	72.4	86.9
Linear probing	61.4	90.2	74.8	95.5	90.2	46.9	**55.8**	73.5	81.5	90.1	82.1	69.4	80.8	39.1	35.9	40.1	65.0	20.3	26.0	14.3	27.6	33.5	58.2	0.05
MLP-4 [11]	54.9	87.4	71.4	99.5	89.1	39.7	52.5	70.6	80.5	90.9	76.8	74.4	80.7	60.9	38.8	40.2	66.5	9.4	21.1	14.5	28.8	31.2	57.7	4.04
Partial [11]	60.3	88.9	72.6	98.7	89.3	50.5	51.5	73.1	82.8	91.7	80.1	72.3	81.7	34.3	35.5	43.2	77.1	15.8	26.2	19.1	28.4	35.0	58.9	12.65
Bias [32]	73.1	86.8	65.7	97.7	87.5	56.4	52.3	74.2	80.4	91.6	76.1	72.5	80.1	47.3	48.5	34.7	66.3	57.6	36.2	17.2	31.6	42.4	62.1	0.25
VPT-Shallow [11]	78.0	91.3	77.2	99.4	90.4	68.4	54.3	79.9	80.1	93.9	83.0	72.7	82.5	40.8	43.9	34.1	63.2	28.4	44.5	21.5	26.3	37.8	62.9	0.05
VPT-Deep [11]	**79.6**	90.8	**78.0**	99.5	91.4	46.5	51.7	76.8	84.9	96.2	85.0	72.0	84.5	67.6	59.4	50.1	74.1	74.4	50.6	25.7	25.7	53.4	67.7	0.22
ARC [2]	62.5	90.0	71.9	99.2	87.8	90.7	51.1	79.0	**89.1**	95.8	84.5	77.0	86.6	75.4	57.4	53.4	83.1	**91.7**	**55.2**	31.6	31.8	59.9	72.6	0.27
RLRR [3]	66.1	90.6	75.5	99.3	92.1	**90.9**	54.7	81.3	87.1	95.9	87.1	76.5	86.7	66.0	57.8	**55.3**	84.1	91.1	**55.2**	28.6	34.0	59.0	73.0	0.41
LoRA+AOFT*	71.8	**92.3**	77.1	**99.5**	**92.6**	86.4	**55.8**	**82.8**	86.9	96.4	87.3	**77.6**	**87.1**	**84.5**	59.3	53.6	**84.7**	86.8	52.3	28.1	35.5	**60.6**	**73.3**	0.14

$x_i x_j(\cos\varphi - 1)$ can be calculated by this way as follows:

$$\begin{aligned} x_j x_i(\cos\varphi - 1) &= \frac{q_j}{\sin\varphi}\frac{q_i}{\sin\varphi}(\cos\varphi - 1) \\ &= \frac{q_j q_i}{\sin^2\varphi}(\cos\varphi - 1) \\ &= -\frac{q_j q_i}{1+\cos\varphi} = -\frac{q_j q_i}{1+q_0}. \end{aligned} \tag{11}$$

Based on Eq. (10) and Eq. (11), we obtain the simplified result as Eq. (5).

E. Detailed Configuration

Tab. 6 summarizes the detailed configurations we used for experiments. As mentioned in Section 4, we utilize grid search to select hyper-parameters such as learning rate, weight decay, batch size, and dropout rate, using the validation set of each task.

F. Experimental details on ablation study

In addition to incorporating our method into the MHA and FFN layers, we also added AOFT solely to the MHA layer. In order to better compare the effects of AOFT in

Table 6. The implementation details of configurations such as optimizer and hyper-parameters. We select the best hyper-parameters for each download task via using grid search.

Optimizer	AdamW
Learning Rate	{0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0001}
Weight Decay	{0.05, 0.01, 0.005, 0.001, 0}
Dropout Rate	{0, 0.1, 0.3, 0.5, 0.7}
Batch Size	{256, 128, 32}
Learning Rate Schedule	Cosine Decay
Training Epochs	100
Warmup Epochs	10

Table 7. Based on the ViT-B backbone, a comparison of the results is presented for applying AOFT to both the FFN and MHA layers, as well as with or without incorporating eigenvalues, on the VTAB dataset.

Methods \ Datasets	Natural								Specialized					Structed										
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVNH	Sun397	Mean	Camelyon	EuroSAT	Resisc45	Retinopathy	Mean	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Mean	Mean Total	Params.(M)
LoRA($\mathbf{W}_q, \mathbf{W}_v$)	67.1	91.4	69.4	98.9	90.4	85.3	54.0	79.5	84.9	95.3	84.4	73.6	84.6	82.9	69.2	49.8	78.5	75.7	47.1	31.0	44.0	59.8	72.3	0.29
LoRA+AOFT($\mathbf{W}_q, \mathbf{W}_v$)	74.0	91.0	72.7	99.3	89.3	80.6	56.8	80.52	84.9	94.6	82.7	75.6	84.4	71.4	57.5	42.7	82.0	83.4	53.9	22.6	44.5	57.3	71.5	0.08
LoRA+AOFT($\mathbf{W}_q, \mathbf{W}_v, \mathbf{W}_{\mathrm{FC1}}, \mathbf{W}_{\mathrm{FC2}}$)	75.0	93.2	70.2	99.2	91.1	84.0	57.4	81.7	85.1	95.6	84.7	75.6	85.3	79.8	60.5	49.1	82.0	79.64	54.92	32.7	45.8	60.6	73.4	0.16
Adapter($\mathbf{W}_{\mathrm{FFN}}$)	69.2	90.1	68.0	98.8	89.9	82.8	54.3	79.0	84.0	94.9	81.9	75.5	84.1	80.9	65.3	48.6	78.3	74.8	48.5	29.9	41.6	58.5	71.4	0.16
Adapter+AOFT($\mathbf{W}_{\mathrm{FFN}}$)	74.1	93.9	72.6	99.4	91.0	82.9	57.6	79.0	85.8	95.1	83.4	76.3	84.1	77.7	61.3	49.1	80.0	80.8	53.8	30.4	42.8	58.5	71.4	0.16
Adapter+AOFT($\mathbf{W}_{\mathrm{FFN}}, \mathbf{W}_{\mathrm{MHA}}$)	67.6	93.9	72.1	99.3	91.6	86.4	53.8	80.7	86.8	95.5	86.0	76.5	86.2	79.0	62.6	51.5	82.7	87.0	54.8	27.7	42.5	60.9	73.5	0.08

the adapter method across different layers, we conducted the following experiments, Tab. 7 shows the experimental results.