Efficient Adaptation of Pre-trained Vision Transformer via Householder Transformation

Supplementary Materials

In the supplementary materials involving our work, we demonstrate Detailed dataset statistic, Hyper-parameters in our work, Experimental details, and broader impacts, including:

- A **Detailed dataset statistic**
- B **Hyper-parameters in our work**
- C **Experimental details on larger-scale and hierarchical ViT backbones**
- D **Experimental details on ablation study**
- E **Broader impacts**

Due to the limitation that supplementary materials larger than 100MB cannot be uploaded to the OpenReview website, only the project code as the concise supplementary materials is uploaded to this website. Please reler to the anonymous link`https://drive.google.com/file/d/18sXhtqMlKZd4_LRICk2NvSlKiFiHrG2d/view?`to obtain the complete code, datasets, and models.

A Detailed dataset statistic

We provide detailed information about the datasets used in this paper, including the number of classes and the sizes of the training, validation, and test sets, in Table 1 (FGVC) and Table 2 (VTAB-1k).The FGVC datasets include CUB-200-2011, NABirds, Oxford Flowers, Stanford Dogs, and Stanford Cars, which are used for fine-grained classification tasks of birds, flowers, dogs, and cars, respectively. The VTAB-1k datasets cover natural, specialized, and structured tasks, including natural image datasets such as CIFAR-100, Caltech101, DTD, Flowers102, Pets, SVHN, and Sun397; specialized image datasets such as Patch Camelyon, EuroSAT, Resisc45, and Retinopathy; and structured image datasets such as Clevr/count, Clevr/distance, DMLab, KITTI/distance, dSprites/location, dSprites/orientation, SmallNORB/azimuth, and SmallNORB/elevation. Detailed information about these datasets is presented in the tables.

Table 1: Dataset statistics for FGVC. "*" denotes the train/val split of datasets following the dataset setting in VPT [21].

Dataset	Description	Classes	Train size	Val size	Test size
CUB-200-2011 [34]	Fine-grained bird species recognition	200	5,394*	600*	5,794
NABirds [35]	Fine-grained bird species recognition	555	21,536*	2,393*	24,633
Oxford Flowers [36]	Fine-grained flower species recognition	102	1,020	1,020	6,149
Stanford Dogs [37]	Fine-grained dog species recognition	120	10,800*	1,200*	8,580
Stanford Cars [38]	Fine-grained car classificatio	196	7,329*	815*	8,041

B Hyper-parameters in our work

Table 3 provides a summary of the configurations used in our experiments. As discussed in Section 4, we performed a grid search on the validation set of each task to determine the optimal hyperparameters, including learning rate, weight decay, batch size, and dropout rate.

C Experimental details on larger-scale and hierarchical ViT backbones

Table 4 and 5 respectively display the comprehensive results of the comparison conducted in Section 4 among ViT-Large and Swin-Base models.

D Experimental details on ablation study

We provide further explanation of the ablation experiments in Section 4. In the study on the transferability of HTA, we replaced the low-rank adaptation matrices in LoRA, we used an HTA module to replace the bottleneck part of LoRA, while in Adapter, we directly replaced the Adapter with an HTA module. The detailed experimental results are presented in the table 6. In the study of the low-rank adaptation part of HTA, we set its dimensions to 0, 1, 2, and 4, respectively. The results are shown in Table 7.

E Broader impacts

Practicality: Our approach differs from traditional methods by employing Householder transformations rather than standard unitary matrices, which can be efficiently derived. This approach boosts the efficiency of parameter usage and significantly cuts down on the number of parameters requiring fine-tuning. With this technique, we manage to achieve high performance while optimizing parameter use. Leveraging large-scale pre-trained models, our HTA method proves to be both highly efficient and practical across diverse applications.

Low Energy Consumption: Our approach enhances the model's computational efficiency by decreasing the necessary computational parameters, thus reducing energy usage during training. This reduction aids in conserving energy and lowering emissions, aligning with global sustainability goals and the push for eco-friendly practices. Moreover, our method not only improves the model's

Table 2: Dataset statistics for VTAB-1k [39].

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

Table 3: 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}
Batch Size	{64, 32, 16}
Adapter Dropout	{0.5, 0.3, 0.2, 0.1, 0}
Learning Rate Schedule	Cosine Decay
Training Epochs	100
Warmup Epochs	10

Table 4: This table is extended from Table 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.

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 [19]	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
Bias [20]	71.0	82.4	51.3	96.3	83.2	59.5	49.9	70.5	72.9	87.9	63.1	71.3	73.8	51.2	50.7	33.5	54.8	65.9	37.3	13.7	22.2	41.2	58.9	0.32
VPT-Shallow [21]	80.6	88.2	67.1	98.0	85.9	78.4	53.0	78.7	79.7	93.5	73.4	73.1	79.9	41.5	52.5	32.3	64.2	48.3	35.3	21.6	28.8	40.6	62.9	0.15
VPT-Deep [21]	**84.1**	88.9	70.8	98.8	90.0	89.0	55.9	82.5	82.5	**96.6**	82.6	73.9	83.9	63.7	60.7	46.1	75.7	83.7	47.4	18.9	36.9	54.1	70.8	0.49
LoRA [1]	75.8	89.8	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
ARC [3]	76.2	89.6	73.4	99.1	90.3	**90.9**	56.5	82.3	85.0	95.7	85.9	75.8	85.6	78.6	62.1	46.7	76.7	75.9	53.0	30.2	35.2	57.3	72.5	0.18
SSF [22]	73.5	91.3	70.0	99.3	91.3	90.6	57.5	81.9	85.9	94.9	85.5	74.4	85.2	80.6	60.0	53.3	80.0	77.6	54.0	31.8	35.0	59.0	73.0	0.60
RLRR [26]	79.3	92.0	74.6	**99.5**	92.1	89.6	**60.1**	83.9	87.3	95.3	**87.3**	75.7	86.4	**82.7**	62.1	**54.6**	80.6	**87.1**	**54.7**	31.3	41.9	61.9	75.2	0.82
HTA	80.8	**92.4**	**76.1**	**99.5**	**92.8**	87.2	59.9	**84.1**	**87.7**	95.5	86.8	**76.5**	**86.6**	82.6	**62.4**	53.4	80.0	**87.1**	53.7	**33.4**	**45.6**	**62.3**	**75.4**	0.54

performance and efficiency but also promotes environmental sustainability by embracing sustainable development principles.

Ethical Aspects: Our model utilizes the vast capabilities of large-scale pre-trained models for representation and generalization. However, it is trained on datasets that might contain problematic data, such as illegal content or inherent biases, which our model could inadvertently learn. To tackle this challenge, addressing model toxicity becomes critical. Consequently, it's imperative to develop

Table 5: This table is extended from Table 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.

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 [21]	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 [21]	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 [20]	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 [21]	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 [21]	**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 [3]	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 [26]	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
HTA	72.0	89.6	76.4	**99.5**	**92.1**	87.8	55.5	**81.8**	86.7	96.3	87.5	76.3	**86.7**	**85.0**	**62.2**	53.7	**84.3**	89.1	52.4	27.6	36.4	**61.3**	**74.2**	0.23

Table 6: This table is extended from Table 5 in Section 4. LoRA and Adapter both follow the configurations from the A paper. In our implementation, the fully connected layers within the bottlenecks are replaced with Householder transformations. "$(\cdot)$" indicates specific configuration information

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	98.8	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
HTA($\mathbf{W}_q, \mathbf{W}_v$)	71.2	93.3	72.5	99.3	91.2	82.7	56.6	81.0	85.3	94.9	82.5	75.7	84.6	80.8	62.9	50.2	78.9	77.4	51.1	29.7	45.4	59.6	72.7	0.09
HTA($\mathbf{W}_q, \mathbf{W}_v, \mathbf{W}_{\text{FC1}}, \mathbf{W}_{\text{FC2}}$)	71.7	93.1	70.9	99.2	90.5	86.6	55.8	81.1	87.6	96.3	85.1	76.2	86.3	80.8	60.1	51.0	82.0	86.9	52.2	32.9	46.0	61.48	73.9	0.28
Adapter($\mathbf{W}_{\text{FC1}}, \mathbf{W}_{\text{FC2}}$)	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
HTA($\mathbf{W}_{\text{FC1}}, \mathbf{W}_{\text{FC2}}$)	72.6	93.0	71.1	99.3	91.4	82.1	57.2	81.0	85.3	95.0	82.9	76.3	84.9	81.6	63.9	49.5	81.2	79.2	51.4	28.1	45.4	60.0	73.0	0.05

Table 7: This table is extended from Fig. 2 in Section 4.

Methods \ Datasets	Natural								Specialized					Structured										
	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)
D'=0	73.0	90.1	71.8	99.3	91.1	83.4	53.7	80.3	82.3	94.2	82.7	73.7	83.2	77.3	61.6	49.2	80.0	81.7	53.3	28.1	41.2	59.1	72.0	0.15
D'=1	76.6	94.3	72.5	99.3	91.3	86.2	56.5	82.4	87.6	95.7	85.0	75.7	86.0	82.6	63.3	52.5	81.0	84.5	52.6	34.5	47.3	62.3	74.7	0.22
D'=2	75.5	94.2	73.0	99.3	91.2	85.7	56.3	82.2	88.0	95.2	84.6	76.1	86.0	82.2	63.1	52.1	80.2	85.5	52.3	33.9	47.5	62.1	74.5	0.30
D'=4	74.6	93.6	72.1	99.3	91.4	88.3	56.1	82.2	88.0	96.3	85.6	76.4	86.6	81.8	64.9	53.6	82.3	84.8	53.6	34.4	47.2	62.8	75.0	0.44

enhanced mechanisms that can both identify and reduce such biases and unlawful information in the datasets.