

# Supplementary Materials: GIST: Improving Parameter Efficient Fine-Tuning via Knowledge Interaction

Anonymous Authors

## 1 MORE DETAILED MOTIVATIONS

In this paper, our primary motivation stems from the disparities between different types of knowledge. Initially, during the pre-training phase, the datasets employed are often large-scale and diverse, yet lacking specialized information. As a result, the pre-training phase mainly endows the model with task-agnostic knowledge. In contrast, the datasets used during the fine-tuning phase tend to be small-scale and specialized, embodying primarily task-specific knowledge. At this juncture, the pre-trained model is tasked with adjusting its intrinsic task-agnostic knowledge to bridge the gap with the task-specific requirements, thereby adapting to downstream tasks. However, the volume of samples in downstream datasets is significantly smaller than that of the pre-training datasets. This means that the information content of task-specific knowledge is also less than that of task-agnostic knowledge. Consequently, the Parameter-Efficient Fine-Tuning (PEFT) approach posits that it isn't necessary to update all model parameters. Instead, making adjustments to or introducing a small number of trainable parameters can suffice for acquiring task-specific knowledge during the fine-tuning phase.

However, during the fine-tuning phase, the PEFT method under the traditional framework introduces learnable parameters that lack an explicit connection with downstream targets, leading to an inadequate acquisition of downstream knowledge. To address this, we have introduced the 'Gist token', creating a bridge between the learnable parameters and downstream objectives for more effective learning of task-specific knowledge. Further enhancing this approach, we utilize knowledge interaction through a Bidirectional Kullback-Leibler Divergence loss. This method calculates the KL divergence between Class logits, representing task-agnostic knowledge, and Gist logits, which embody task-specific knowledge. Such an interaction allows for mutual guidance between these knowledge types, significantly improving the model's adaptability to downstream tasks.

## 2 DATASETS

In this section, we present detailed information about the VTAB-1K benchmark [48], FGVC datasets, GLUE benchmark [44] used in this paper, as shown in Tables 1, 2, and 3. Notably, following the previous work [1], we utilize 8 tasks on the GLUE benchmark, including MNLI [46], QQP<sup>1</sup>, QNLI [39], SST-2 [40], STS-B [5], MRPC [9], RTE [3, 8, 12, 14], and CoLA [45].

## 3 IMPLEMENTATION DETAILS ON VATB-1K

Our GIST framework is compared against the traditional fine-tuning framework. Therefore, for different PEFT methods, we keep the implementation settings the same with the traditional framework.

<sup>1</sup>[data.quora.com/First-Quora-Dataset-Release-Question-Pairs](https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs)

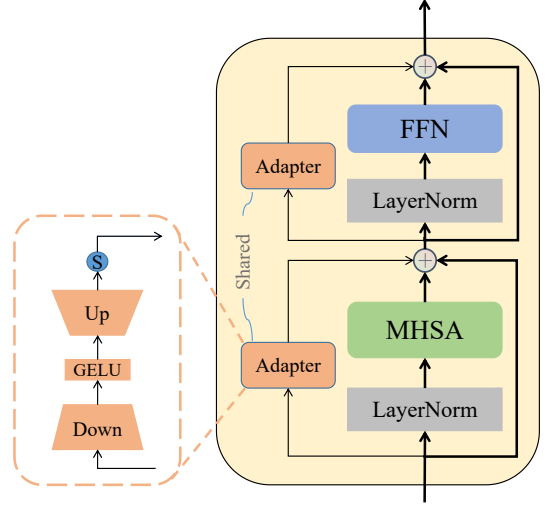


Figure 1: The Adapter method utilized on the VTAB-1K benchmark in this paper.

### 3.1 Adapter, VPT, LoRA and SSF

**3.1.1 Training settings.** For Adapter, VPT, LoRA and SSF methods, we follow the training settings of SSF [27]. Namely, we directly resize the image to  $224 \times 224$ . We employ AdamW [30] as the optimizer, set the batch size to 32, and designate 100 epochs with a provision for a 10-epoch warm-up at a warmup learning rate of  $1e-7$ . Regarding the initial learning rate (lr), previous study [27] have established different values for various datasets, as detailed in Table 4.

**3.1.2 Adapter method settings.** In this study, we opt for the parallel model Adapter instead of the sequential one, offering a stronger baseline (with fewer parameters and improved performance) when compared under the traditional fine-tuning framework, as illustrated in Table 5. Specifically, as depicted in Figure 1, within the Adapter, we set the dimension of the intermediate layer to 4 and designate a scaling factor  $s = 0.1$  for all experiments.

**3.1.3 VPT method settings.** In the original paper of the VPT method [18], the authors conducted an extensive search on the VTAB-1K benchmark for various tasks, experimenting with the lengths of the newly introduced prompt token in the set  $\{1, 5, 10, 50, 100, 200\}$ . This search process is both tedious and intricate. However, the primary aim of our study is to contrast the advantages of the GIST framework against the traditional fine-tuning framework. Consequently, for the VPT method, we consistently set the prompt token length to 20 across all experiments.

**3.1.4 LoRA method settings.** In this study, we adhere to the configurations detailed in the original LoRA paper [17], wherein LoRA

Group	Dataset	Train	Val	Test	# Class
Natural	CIFAR100 [25]	800/1,000	200	10,000	100
	Caltech101 [10]			6,084	102
	DTD [7]			1,880	47
	Oxford-Flowers102 [36]			6,149	102
	Oxford-Pets [37]			3,669	37
	SVHN [34]			26,032	10
	Sun397 [47]			21,750	397
Specialized	Patch Camelyon [43]	800/1,000	200	32,768	2
	EuroSAT [15]			5,400	10
	Resisc45 [6]			6,300	45
	Retinopathy [13]			42,670	5
Structured	Clevr/count [21]	800/1,000	200	15,000	8
	Clevr/distance [21]			15,000	6
	DMLab [2]			22,735	6
	KITTI-Dist [11]			711	4
	dSprites/location [33]			73,728	16
	dSprites/orientation [33]			73,728	16
	SmallNORB/azimuth [26]			12,150	18
	SmallNORB/elevation [26]			12,150	18

Table 1: The details of the VTAB-1K benchmark.

Dataset	Train	Val	Test	# Class
Food-101 [4]	(1/2/4/8/16)*(#Class)	20,200	30,300	101
Oxford-Pets [38]		736	3,669	37
Stanford Cars [24]		1,635	8,041	196
Oxford-Flowers102 [35]		1,633	2,463	102
FGVC-Aircraft [32]		3,333	3,333	100

Table 2: The details of the FGVC datasets.

Dataset	Task	Domain	Metric	Train	Test
MNLI	natural language inference	various	accuracy	393k	20k
QQP	paraphrase detection	social QA questions (Quora)	accuracy & F1	364k	391k
QNLI	natural language inference	Wikipedia	accuracy	105k	5.4k
SST-2	sentiment analysis	Movie Reviews	accuracy	67k	1.8k
STS-B	sentence similarity	various	Pearson & Spearman corr.	7k	1.4k
MRPC	paraphrase detection	news	accuracy & F1	3.7k	1.7k
RTE	natural language inference	News, Wikipedia	accuracy	2.5k	3k
CoLA	acceptability	various	Matthews corr.	8.5k	1k

Table 3: The details of 8 tasks we utilized on the GLUE benchmark.

Dataset	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele
Initial lr	5e-3	1e-3	5e-3	5e-3	5e-3	1e-2	5e-3	5e-3	3e-3	2e-3	5e-3	2e-3	5e-2	5e-3	1e-2	1e-2	5e-3	2e-2	5e-3

Table 4: The detailed initial learning rate of SSF [27] method on the VTAB-1K benchmark.

	Natural							Specialized				Structured									
Method	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean	Params. (M)
sequential Adapter	74.1	86.1	63.2	97.7	87.0	34.6	50.8	76.3	88.0	73.1	70.5	45.7	37.4	31.2	53.2	30.3	25.4	13.8	22.1	55.82	0.27
parallel Adapter	70.2	92.6	74.6	99.4	91.2	80.4	51.4	84.1	96.3	88.0	75.6	84.2	59.6	53.2	76.3	60.7	51.9	27.8	40.2	71.46	0.13

**Table 5: The comparative results on VTAB-1K, using the backbone of ViT-B/16 pre-trained on ImageNet-21K. The results of the sequential adapter are from [27].**

modules with  $r = 4$  are assigned to the Query and Value projection layers in Transformer blocks.

**3.1.5 SSF method settings.** SSF [27] eliminates the need for a tedious hyperparameter search process. Therefore, we conduct experiments under our GIST framework using the default settings from the SSF source code directly.

## 3.2 FacT, ReAdapter, Bi-Adapter

**3.2.1 Training settings.** Following [19, 20, 31], we resize the images to  $224 \times 224$  and then normalize them using ImageNet’s mean and standard deviation. We employ AdamW as our optimizer with a batch size set to 64. The initial learning rate is set at  $1e-3$ , with a weight decay of  $1e-4$ . Training is conducted over 100 epochs, inclusive of 10 warm-up epochs, and we utilize the CosineAnnealingLR [29] for the learning rate scheduler.

**3.2.2 FacT method settings.** In the original paper for FacT [19], the authors undertook an extensive hyperparameter search. Specifically, for different tasks within the VTAB-1K benchmark, the authors searched for the scaling factor  $s$  across  $\{0.01, 0.1, 1, 10, 100\}$ . Moreover, for the rank  $r$ , they searched within the set  $\{2, 4, 8, 16, 32\}$ . In our study, we directly utilized the default settings from the official FacT code to conduct experiments under the GIST framework. However, it’s important to note that for the FacT method under the traditional fine-tuning framework, we directly report the results that the authors provided in the original paper, which came after their elaborate hyperparameter search. Therefore, our GIST framework operates from a potentially disadvantaged baseline. Still, the results of Table 1 demonstrate that our GIST framework manages to surpass the traditional framework by 0.32%.

**3.2.3 ReAdapter method settings.** In the original ReAdapter [31] paper, the authors conducted a relatively straightforward hyperparameter search. Specifically, they only searched for the scaling factor  $s$  within the set  $\{0.1, 0.5, 1, 5, 10\}$ . Consequently, for ReAdapter under our GIST framework, we carried out the same hyperparameter search. The results indicate that using ReAdapter within our GIST framework outperforms its utilization within the traditional fine-tuning framework by 0.43%.

**3.2.4 Bi-Adapter method settings.** In [20], the optimal configuration for Bi-Adapter was identified as  $h = 32$ , and the authors also conducted a search for the scaling factor  $s$  within the set  $\{0.01, 0.1, 1, 10, 100\}$ . Following this setup, we conducted experiments based

on the GIST framework. The results demonstrate that, compared to traditional frameworks, our framework is able to further enhance the performance of Bi-Adapter by 0.42%, with virtually no addition of extra parameters.

## 4 IMPLEMENTATION DETAILS ON FGVC DATASETS

For the FGVC datasets, we use Adapter, VPT [18], and SSF [27] as representatives of three different PEFT methods, and have verified them in  $\{1, 2, 4, 8, 16\}$ -shot scenarios respectively.

### 4.1 Training settings

For the three different PEFT methods, we consistently use AdamW [30] as the optimizer, with a batch size set to 64, a learning rate of  $1e-3$ , weight decay of  $1e-3$ , training for 100 epochs with 10 warmup-epochs.

### 4.2 Methods settings

For the three PEFT methods, the setup is the same as in Sec. 3.1.2, 3.1.3, 3.1.5.

## 5 IMPLEMENTATION DETAILS ON GLUE BENCHMARK

### 5.1 Training settings

For the eight tasks within the GLUE benchmark, we employ consistent training configurations [1]. Specifically, we set the batch size to 32, the max token length to 256, and the learning rate to  $3e-4$ . Training was conducted over 20 epochs, incorporating 500 warm-up iterations.

### 5.2 Method settings

For NLP tasks on the GLUE benchmark, we carry out relatively straightforward experiments to further demonstrate the universality of our GIST framework. We employ the default parameters from Adapter[1, 22] for our experiments. Specifically, we use GELU [16] as the activation function and set the reduction factor to 32.

## 6 IMPLEMENTATION DETAILS ON VISION&LANGUAGE TASKS

For Vision&Language tasks (Flowers102 [36], DTD [7], and UCF101 [41] datasets), our configuration is similar to the experimental setup

$\lambda$	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean
-	70.2	92.6	74.6	99.4	91.2	80.4	51.4	84.1	96.3	88.0	75.6	84.2	59.6	53.2	76.3	60.7	51.9	27.8	40.2	71.46
0.25	74.3	92.3	76.9	99.5	92.0	85.7	54.5	87.1	96.5	87.5	77.2	83.3	61.2	53.4	80.2	70.8	52.0	29.3	39.3	73.31
0.5	74.5	92.2	76.3	99.5	92.1	85.0	54.4	88.1	96.3	87.6	77.4	83.1	59.8	54.0	78.7	69.5	51.9	29.1	41.0	73.18
0.75	74.4	92.3	76.6	99.5	92.3	85.3	54.6	88.2	96.4	87.9	76.5	83.6	60.1	53.0	81.2	72.3	52.1	29.2	39.8	73.44

Table 6: Detailed results of the ablation studies for different  $\lambda$  on the VTAB-1K benchmark.

Token length	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean	Params. (M)
1	74.5	92.3	76.9	99.5	92.3	85.7	54.6	88.2	96.5	87.9	77.4	83.6	61.2	54.0	81.2	72.3	52.1	29.3	41.0	73.71	0.13
10	74.0	92.3	76.9	99.6	92.3	86.3	54.2	87.2	96.1	87.4	75.8	83.6	61.5	53.3	79.5	71.3	52.9	28.2	42.6	73.42	0.14
50	73.0	92.9	74.9	99.5	91.9	87.8	52.9	85.8	96.3	87.7	75.6	83.3	54.8	52.6	79.4	65.3	53.4	19.2	40.8	71.96	0.16
100	72.9	92.9	73.1	99.4	91.1	88.1	52.2	85.3	96.3	86.9	76.8	81.7	37.2	52.0	79.9	66.1	52.2	28.8	40.2	71.21	0.21

Table 7: Detailed results of the ablation studies for different Gist token length on the VTAB-1K benchmark.

$\mathcal{L}_{cls}$	$\mathcal{L}_{gist}$	$\mathcal{L}_{fkl}$	$\mathcal{L}_{rkl}$	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean
✓				70.2	92.6	74.6	99.4	91.2	80.4	51.4	84.1	96.3	88.0	75.6	84.2	59.6	53.2	76.3	60.7	51.9	27.8	40.2	71.46
✓	✓	✓	✓	74.5	92.3	76.9	99.5	92.3	85.7	54.6	88.2	96.5	87.9	77.4	83.6	61.2	54.0	81.2	72.3	52.1	29.3	41.0	73.71
✓		✓	✓	73.9	91.9	76.6	99.5	92.2	85.4	54.2	87.6	96.4	88.2	76.6	83.7	60.2	53.6	79.9	70.4	53.2	28.4	40.3	73.29
✓	✓		✓	74.3	92.3	76.6	99.5	92.1	85.0	54.6	87.8	96.7	87.7	77.1	83.5	61.5	53.5	80.5	70.8	52.9	29.6	40.2	73.48
✓	✓	✓		74.3	92.2	76.8	99.5	92.1	85.4	54.6	87.5	96.6	87.8	76.2	83.5	60.5	53.7	79.8	68.2	51.8	27.0	41.0	73.07
✓			✓	73.4	91.9	76.5	99.5	92.1	84.8	54.0	87.4	96.7	87.5	76.8	83.7	60.5	53.2	79.2	66.4	52.7	26.5	39.7	72.76
✓		✓		72.5	92.0	75.9	99.5	92.2	83.0	53.8	84.2	96.4	87.6	76.6	84.0	59.8	53.6	78.2	69.0	52.4	27.0	39.3	72.46
✓	✓			72.2	92.3	75.3	99.4	91.8	83.0	53.7	86.4	96.5	87.6	76.7	83.6	61.0	53.1	79.1	67.0	52.1	28.9	41.9	72.71

Table 8: Detailed results of the ablation studies of our loss function on the VTAB-1K benchmark.

Loss function	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean
$\mathcal{L}_{cls}$	70.2	92.6	74.6	99.4	91.2	80.4	51.4	84.1	96.3	88.0	75.6	84.2	59.6	53.2	76.3	60.7	51.9	27.8	40.2	71.46
$\mathcal{L}_{cls} + \mathcal{L}_{gist} + \mathcal{L}_{mse}$	74.6	92.6	76.5	99.6	92.2	87.1	53.2	88.1	96.6	87.6	75.9	81.7	61.3	52.0	82.6	65.1	52.9	28.8	39.5	73.03
$\mathcal{L}_{cls} + \mathcal{L}_{gist} + \mathcal{L}_{cos}$	73.0	92.0	75.8	99.5	92.2	82.9	53.6	86.2	96.4	87.5	76.5	83.6	61.3	53.7	80.1	69.1	52.6	28.0	40.7	72.88
$\mathcal{L}_{cls} + \mathcal{L}_{gist} + \mathcal{L}_{bkl}$	74.5	92.3	76.9	99.5	92.3	85.7	54.6	88.2	96.5	87.9	77.4	83.6	61.2	54.0	81.2	72.3	52.1	29.3	41.0	73.71

Table 9: Detailed results on different loss functions for knowledge interaction on the VTAB-1K benchmark.

Method	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean	Params. (M)
S+Adapter	68.1	92.3	73.1	99.4	90.3	81.9	52.2	87.3	96.1	85.8	77.0	81.9	60.3	49.7	75.6	67.2	50.4	25.5	42.2	71.39	0.07
<b>S+Adapter*</b>	70.4	92.4	74.2	99.4	90.9	85.2	52.2	86.4	96.2	85.6	76.5	83.1	61.2	52.3	77.7	70.4	50.8	28.2	43.8	72.47	0.07
L+Adapter	72.8	91.8	74.4	99.5	92.2	84.0	54.0	87.1	96.2	89.1	75.6	78.6	57.0	52.6	77.5	68.2	53.4	25.7	34.8	71.81	0.30
<b>L+Adapter*</b>	77.3	91.7	77.5	99.6	92.9	88.2	58.5	87.5	96.6	89.8	76.4	81.5	55.7	54.8	81.9	73.7	54.1	27.6	38.5	73.89	0.30

**Table 10: Detailed results for ViT-S/16 (S) and ViT-L/16 (L) [42] on the VTAB-1K benchmark. The symbol \* indicates employing the PEFT method within our GIST framework.**

Method	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean	Params. (M)
Adapter	70.2	93.3	77.3	99.6	92.4	82.1	54.9	87.9	96.1	88.3	76.8	84.6	56.2	52.8	83.6	78.2	54.2	24.4	37.9	73.19	0.21
<b>Adapter*</b>	71.6	93.5	77.9	99.6	92.6	85.4	55.6	88.9	96.7	88.7	77.0	84.6	60.4	54.3	85.3	78.9	53.1	26.8	38.0	74.15	0.21

**Table 11: Detailed results for Swin-B [28] on the VTAB-1K benchmark. The symbol \* indicates employing the PEFT method within our GIST framework.**

Method	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean
Adapter	70.2	92.6	74.6	99.4	91.2	80.4	51.4	84.1	96.3	88.0	75.6	84.2	59.6	53.2	76.3	60.7	51.9	27.8	40.2	71.46
Adapter+BYOT	62.8	92.2	71.3	99.2	89.2	83.4	46.8	85.2	96.3	86.9	76.5	81.8	49.7	52.9	59.0	53.8	76.5	23.2	37.4	69.70
Adapter+CS-KD	77.6	94.0	75.2	99.7	91.7	88.3	51.8	83.5	96.7	88.3	76.4	80.1	28.1	51.7	74.4	52.9	79.5	24.9	38.7	71.24
Adapter+USKD	73.0	92.2	72.2	99.5	91.4	80.5	52.1	85.9	96.7	88.0	76.6	83.4	58.9	53.6	57.6	53.5	77.9	26.2	37.4	71.40
<b>Adapter*</b>	74.5	92.3	76.9	99.5	92.3	85.7	54.6	88.2	96.5	87.9	77.4	83.6	61.2	54.0	81.2	72.3	52.1	29.3	41.0	73.71

**Table 12: Detailed results with different self-knowledge distillation methods on the VTAB-1K benchmark. The symbol \* indicates employing the PEFT method within our GIST framework.**

in MaPLe [23]. Specifically, we conduct 16-shot training and tested on full test sets. We use a pre-trained ViT-B/16 CLIP model as the backbone. Our batch size is set to 4, with a learning rate of 0.0035, and we utilize SGD as the optimizer. Training is conducted over five epochs. For text modality inputs, we consistently use the template ‘a photo of a {CLASS}’. Regarding the model settings, we exclusively adopt the default configurations from the official MaPLe code.

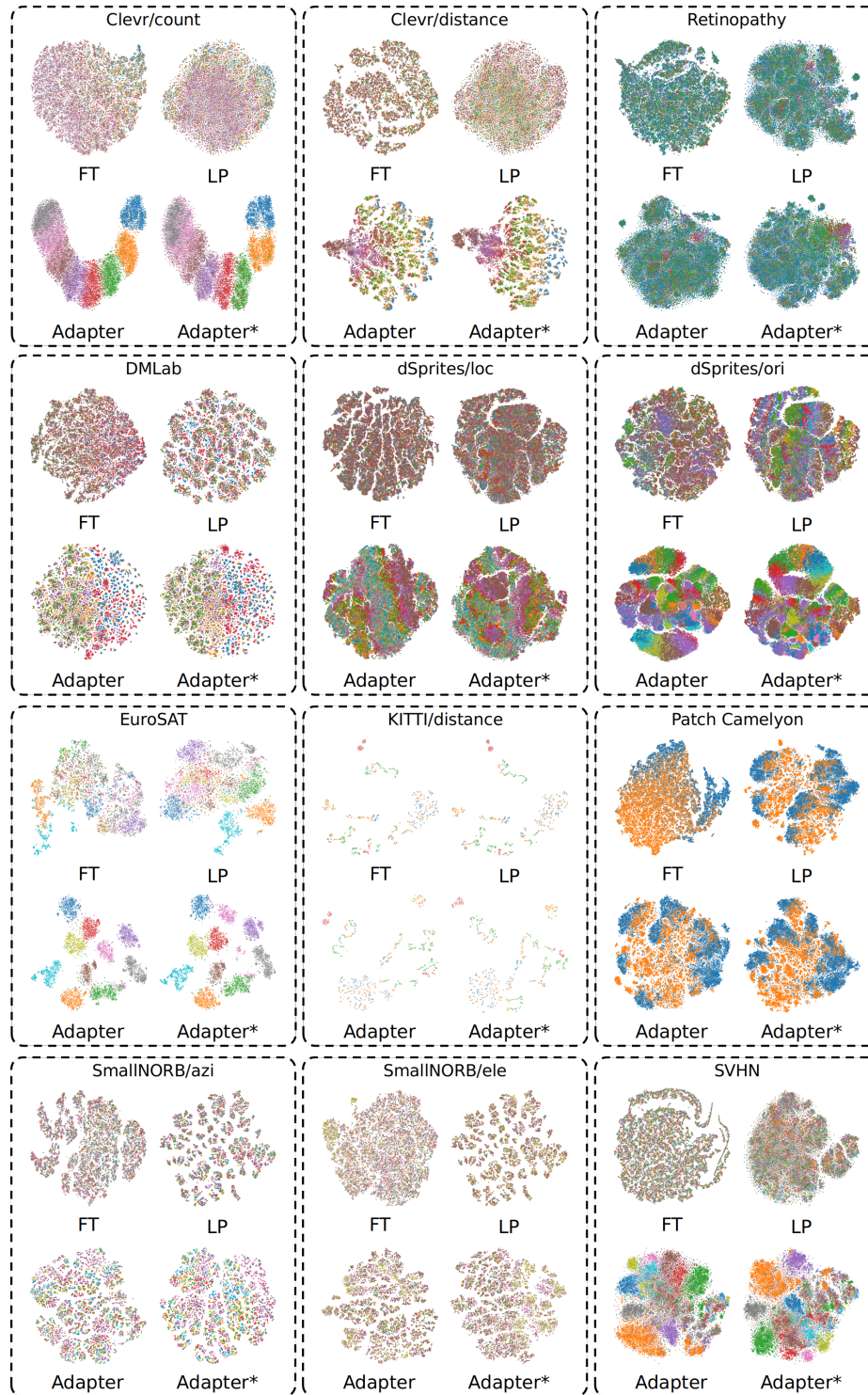
## 7 MORE EXPERIMENTAL RESULTS

Due to space constraints, when conducting ablation experiments on the VTAB-1K benchmark, we only present the arithmetic mean of the Top-1 accuracy. Therefore, we display the complete results of all experiments, as shown in Tables 6, 9, 7, 8, 10, 11, and 12.

## 8 VISUALIZATION

Due to space constraints in the main paper. Thus, we present the more visualization results for the VTAB-1K benchmark, as depicted in Figure 2.





**Figure 2: The results of visualization. We selected the datasets from the VTAB-1K benchmark with fewer than 20 categories for visualization. FT stands for full parameter fine-tuning, and LP stands for Linear Probing. The symbol \* indicates employing the PEFT method within our GIST framework.**

## REFERENCES

- [1] Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. 2022. Attempt: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 6655–6672.
- [2] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Victor Valdés, Amir Sadik, et al. 2016. Deepmind lab. *arXiv preprint arXiv:1612.03801* (2016).
- [3] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. *TAC* 7 (2009), 8.
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101-mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI* 13. Springer, 446–461.
- [5] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055* (2017).
- [6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* 105, 10 (2017), 1865–1883.
- [7] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3606–3613.
- [8] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*. Springer, 177–190.
- [9] Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.
- [10] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*. IEEE, 178–178.
- [11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.
- [12] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*. 1–9.
- [13] Ben Graham. 2015. Kaggle diabetic retinopathy detection competition report. *University of Warwick* 22 (2015).
- [14] R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szepes. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Vol. 7. 785–794.
- [15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226.
- [16] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
- [17] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- [18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.
- [19] Shibo Jie and Zhi-Hong Deng. 2023. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1060–1068.
- [20] Shibo Jie, Haoqing Wang, and Zhi-Hong Deng. 2023. Revisiting the parameter efficiency of adapters from the perspective of precision redundancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17217–17226.
- [21] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.
- [22] Rabeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems* 34 (2021), 1022–1035.
- [23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19113–19122.
- [24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*. 554–561.
- [25] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [26] Yann LeCun, Fu Jie Huang, and Leon Bottou. 2004. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Vol. 2. IEEE, II–104.
- [27] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems* 35 (2022), 109–123.
- [28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
- [29] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
- [30] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [31] Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. 2023. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106* (2023).
- [32] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* (2013).
- [33] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. 2017. dsprites: Disentanglement testing sprites dataset.
- [34] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).
- [35] M-E Nilsback and Andrew Zisserman. 2006. A visual vocabulary for flower classification. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2. IEEE, 1447–1454.
- [36] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 722–729.
- [37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3498–3505.
- [38] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3498–3505.
- [39] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).
- [40] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.
- [41] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [43] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II* 11. Springer, 210–218.
- [44] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).
- [45] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7 (2019), 625–641.
- [46] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).
- [47] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3485–3492.
- [48] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867* (2019).