Lessons and Insights from a Unifying Study of Parameter-Efficient Fine-Tuning (PEFT) in Visual Recognition

Zheda Mai[1], **Ping Zhang**[1], **Cheng-Hao Tu**[1],
Hong-You Chen[1], **Quang-Huy Nguyen**[1], **Li Zhang**[2], **Wei-Lun Chao**[1]
[1]The Ohio State University, [2]Google Research.

Abstract

Parameter-efficient fine-tuning (PEFT) has attracted significant attention due to the growth of pre-trained model sizes and the need to fine-tune (FT) them for superior downstream performance. Despite a surge in new PEFT methods, a systematic study to understand their performance and suitable application scenarios is lacking, leaving questions like "when to apply PEFT" *and* "which method to use" *largely unanswered, especially in visual recognition. In this paper, we conduct a unifying empirical study of representative PEFT methods with Vision Transformers. We systematically tune their hyperparameters to fairly compare their accuracy on downstream tasks. Our study offers a practical user guide and unveils several new insights. First, if tuned carefully, different PEFT methods achieve similar accuracy in the low-shot benchmark VTAB-1K. This includes simple approaches like FT the bias terms that were reported inferior. Second, despite similar accuracy, we find that PEFT methods make different mistakes and high-confidence predictions, likely due to their different inductive biases. Such an inconsistency (or complementarity) opens up the opportunity for ensemble methods, and we make preliminary attempts at this. Third, going beyond the commonly used low-shot tasks, we find that PEFT is also useful in many-shot regimes, achieving comparable or better accuracy than full FT while using significantly fewer parameters. Lastly, we investigate PEFT's ability to preserve a pre-trained model's robustness to distribution shifts (*e.g.*, CLIP). Perhaps not surprisingly, PEFT approaches outperform full FT alone. However, with weight-space ensembles, full FT can better balance target distribution and distribution shift performance, suggesting a future research direction for robust PEFT*[1].

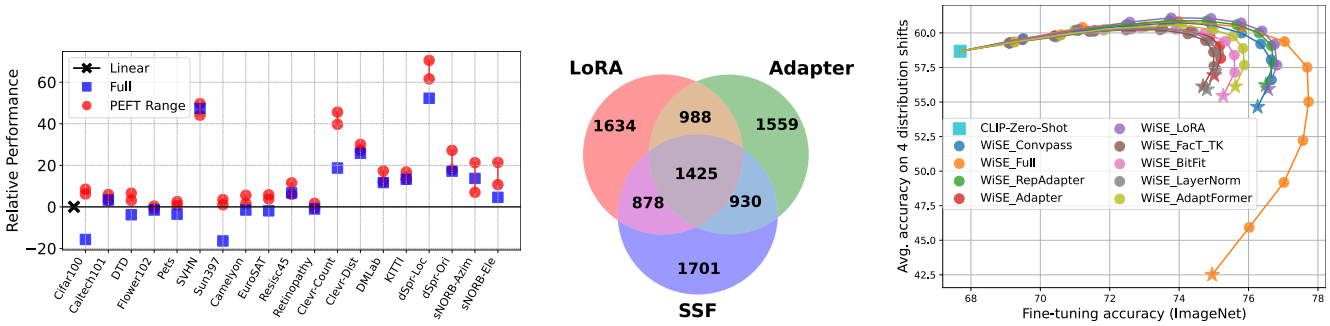
1. Introduction

Pre-training and then fine-tuning (FT) has become the standard practice to tackle visual recognition problems [9]. The community-wide enthusiasm for open-sourcing has made it possible to access large, powerful pre-trained models learned from a gigantic amount of data, e.g., ImageNet-21K [79] or LAION-5B [81]. More research focus has thus been on how to FT such large models [101]. Among existing efforts, parameter-efficient transfer learning (PEFT), has attracted increasing attention lately [17, 28]. Instead of FT the whole model (*i.e.*, full FT) or the last fully connected layer (*i.e.*, linear probing), PEFT approaches seek to update or insert a relatively small number of parameters to the pre-trained model [99]. Doing so has several noticeable advantages. First, as named, PEFT is parameter-efficient. For one downstream task (*e.g.*, recognizing bird species or car brands), it only needs to learn and store a tiny fraction of parameters on top of the pre-trained model. Second, accuracy-wise, PEFT has been shown to consistently outperform linear probing and often beat full FT, as reported on the commonly used low-shot image classification benchmark VTAB-1K [104].

To date, a plethora of PEFT approaches have been proposed, bringing in inspiring ideas and promising results. Along with this come several excellent surveys that summarize existing PEFT approaches [17, 99, 101]. *Yet, a systematic understanding of the PEFT paradigm seems still missing.* For example, with so many PEFT approaches, there is a lack of unifying references for when and how to apply them. Though superior accuracy was reported on the low-shot benchmark VTAB-1K, there is not much discussion on how PEFT approaches achieve it. Does it result from PEFT's ability to promote transferability or prevent over-fitting? The current evaluation also raises the question of whether PEFT is useful beyond a low-shot scenario. Last but not least, besides superior accuracy, do existing PEFT approaches offer different, ideally, complementary information?

Attempting to answer these questions, we conduct a unifying empirical study of representative PEFT methods for Vision Transformers [19], including Low-Rank Adaptation (LoRA) [37], Visual Prompt Tuning (VPT) [42], Adapter [36], and *ten* other approaches. We systematically tune their hyperparameters to fairly compare their accuracy on the low-shot benchmark VTAB-1K. This includes

[1] https://github.com/OSU-MLB/ViT_PEFT_Vision.



(a) Accuracy gain vs. linear probing on VTAB-1K (19 tasks) (b) Prediction overlaps (5K most confident) (c) Target distribution vs. distribution shifts

Figure 1. Highlights of our insights. **(a) Downstream accuracy:** with proper implementation and fair tuning, different PEFT methods achieve similar accuracy (●-●: the range from the most to the least accurate methods) and consistently outperform linear probing (×) and full FT (■) on VTAB-1K. **(b) Diverse predictions:** despite reaching similar downstream performance, different PEFT methods produce diverse predictions. This opens new opportunities for ensemble approaches and other learning paradigms (*e.g.* semi-supervised learning) that can exploit the prediction discrepancies. **(c) Distribution shift accuracy:** FT a CLIP ViT-B/16, known for its generalizability across domains, with PEFT on ImageNet-1K (100 samples/class) better preserves the distribution shift accuracy (Y-axis, averaged across ImageNet-(V2, S, R, A) than full FT, evidenced by the ⋆ points. Interestingly, *weight-space ensembles (WiSE) [96] is applicable between PEFT's FT model and the pre-trained model* (■), but not as effective as applying it to the fully FT model. Details are in section 3, section 4 and section 7.

learning rate, weight decay, and method-specific parameters like the PEFT module sizes. Besides VTAB-1K, we examine PEFT methods on full-size downstream datasets such as CIFAR-100 [49], RESISC for remote sensing [12], and Clevr-Distance for depth classification [45, 104]. We also conduct a study on ImageNet [15] and its variants with domain shifts [25, 34, 35, 78] for robustness evaluation.

We summarize our key findings and analyses as follows:

Representative PEFT methods perform similarly on VTAB-1K, when properly implemented and tuned (Figure 1a). This includes methods previously considered less effective, such as BitFit [103], which FT only the bias terms of the frozen backbone. Methods originally proposed for NLP, like Adapter [36] and LoRA [37] also exhibit impressive performance when their bottleneck dimensions are carefully tuned. Among all the hyperparameters, we find the drop path rate [38] particularly important. Ignoring it (*i.e.*, setting it to 0) significantly degrades the performance, potentially due to over-fitting. Overall, PEFT methods consistently outperform linear probing and full FT on all 19 image classification tasks (with $1,000$ training samples) in VTAB-1K.

While similarly accurate on average, PEFT approaches make different predictions (Figure 1b). The above finding seems daunting: *if existing PEFT approaches all perform similarly in terms of accuracy, do we learn anything useful beyond a single approach?* This is particularly worrisome given that they FT the same backbone using the same downstream data. Fortunately, our analysis shows that different PEFT methods learn differently from the same data, resulting in *diverse prediction errors and confidence*. We attribute this to their inductive bias differences [70] — they explicitly specify different parameters to be updated or inserted. This opens up the door to leverage their discrepancy for improvement, *e.g.*, through *ensemble methods [16, 110] or co-training [6, 8]* and we provide preliminary studies.

PEFT is also effective in many-shot regimes. We extend PEFT beyond low-shot regime and find it effective even with ample downstream training data — PEFT can be on par or surpass full FT. This suggests that adjusting only a fraction of parameters in a suitably pre-trained backbone (*e.g.*, pre-trained on ImageNet-21K [19]) could already offer a sufficient capacity [105] to reach a performant hypothesis for downstream tasks.

PEFT appears more robust than full FT to distribution shifts, but WiSE overturns this advantage (Figure 1c). We also assess PEFT's robustness to distribution shifts, following [96]. We consider a CLIP backbone [75], known for its superior generalizability to distribution shifts, and FT it with PEFT on ImageNet-1K. We found that PEFT retains CLIP's generalizability (*e.g.*, to samples from ImageNet-(V2, S, R, A)) better than full FT. This may not be surprising. What is interesting is that the weight-space ensembles (WiSE) between the FT and pre-trained models [96] is compatible with PEFT as well to further improve the robustness without sacrificing the target accuracy. To the best of our knowledge, *we are the first to explore WiSE for PEFT.* Nevertheless, full FT with WiSE can achieve even higher accuracy in both downstream and distribution shift data than PEFT, suggesting a further research direction in robust PEFT.

What lead to PEFT's success? We attempt to answer this fundamental question by analyzing the findings in our study. On VTAB-1K with 19 tasks, we identify two scenarios: (1) in certain tasks, full FT outperforms linear probing, suggesting the need to update the backbone; (2) in other tasks, linear probing outperforms full FT, suggesting either the backbone is good enough or updating it risks over-fitting.

The superior accuracy of PEFT in both scenarios suggests that PEFT acts as an *effective regularizer* during low-shot training. Still using VTAB but with ample training data, we find that for scenario (1) tasks, PEFT performs similarly with full FT, suggesting that its regularization role does not impede model learning from abundant data. For scenario (2) tasks, PEFT can surprisingly still outperform full FT, indicating that it effectively transfers (or preserves) crucial pre-trained knowledge that full FT may discard. Overall, PEFT succeeds as a **high-capacity learner** equipped with an **effective regularizer**.

Contributions. Instead of chasing the leaderboard, we systematically scrutinize existing methods via a unifying study. Our contribution is thus not a technical novelty, but: (1) a *systematic framework* for reproducible evaluations of PEFT methods; (2) a set of *empirical recommendations* on when and how to use PEFT methods for practitioners (section 3, section 5); (3) *new insights for future research* including leveraging PEFT's prediction differences (section 4) and exploring robust fine-tuning (section 7).

2. Background

2.1. Large pre-trained models

Building upon networks with millions (or billions) of parameters and massive training data, large pre-trained models have led to groundbreaking results in various downstream tasks [56, 69] and shown emerging capabilities not observed previously [9, 47, 53]. For example, a Vision Transformer (ViT) [19] trained with ImageNet-21K (14M images) leads to consistent gains v.s. a ViT trained with ImageNet-1K (1.3M images) [19]. ViTs pre-trained with millions of image-text pairs via a contrastive objective function (*e.g.*, CLIP-ViT) [13, 75] show an unprecedented zero-shot capability and robustness to distribution shifts [75]. We focus on the ImageNet-21K-ViT and CLIP-ViT in this paper.

Vision Transformer (ViT). A ViT contains M Transformer layers consisting of a multi-head self-attention (MSA) block, a multi-level perceptron (MLP) block, two Layer Normalization (LN) blocks [4], and two residual links. The m-th Transformer layer can be formulated as

$$\boldsymbol{Z}'_m = \mathrm{MSA}\left(\mathrm{LN}\left(\boldsymbol{Z}_{m-1}\right)\right) + \boldsymbol{Z}_{m-1}, \tag{1}$$

$$\boldsymbol{Z}_m = \mathrm{MLP}\left(\mathrm{LN}\left(\boldsymbol{Z}'_m\right)\right) + \boldsymbol{Z}'_m, \tag{2}$$

where $\boldsymbol{Z}_{m-1}$ is the output of the preceding $(m-1)$-th Transformer layer. Without loss of generality, let us consider a single-head MSA where the input $\boldsymbol{Z}$ is first projected into three matrices, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. The output of this block is then formulated as:

$$\boldsymbol{Q} = \boldsymbol{W}_Q \boldsymbol{Z}, \quad \boldsymbol{K} = \boldsymbol{W}_K \boldsymbol{Z}, \quad \boldsymbol{V} = \boldsymbol{W}_V \boldsymbol{Z}, \tag{3}$$

$$\boldsymbol{V} \times \mathsf{Softmax}(\frac{\boldsymbol{K}^\top \boldsymbol{Q}}{\sqrt{D}}). \tag{4}$$

2.2. Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning is arguably the most common way to tailor a pre-trained model for downstream tasks [89, 96, 111]. As the size of pre-trained models gets larger, updating and storing all the parameters for one downstream task becomes inefficient. PEFT has thus emerged as a promising paradigm. PEFT was originally developed in NLP [3, 29, 33, 51, 58, 68, 83, 84, 91, 103, 109] and has attracted increasing attention in vision [11, 42, 43, 55, 59, 107]. Existing approaches can generally be categorized into four groups: prompt-based, adapter-based, direct selective tuning, and efficient selective tuning. *We focus on visual recognition and compare representative PEFT approaches applicable to ViTs.*

Prompt-based. Prompt-based method emerged in NLP [54, 57] whose core concept is augmenting the input data with task-specific prompts. **Visual Prompt Tuning (VPT)** [42] adapts such an idea to ViTs. Its deep version (VPT-Deep) prepends a set of soft prompts to the input tokens of each Transformer layer (*i.e.*, $\{\boldsymbol{Z}_m\}_{m=0}^{M-1}$) and only optimizes the prompts during FT. Other representative works in this category include [5, 27, 87, 88, 94, 102].

Adapter-based. They typically introduce additional trainable parameters to a frozen pre-trained model [54]. Initially developed for domain adaptation [76, 77] and continual learning [66, 80], they have been extended to adapt Transformer-based models in NLP and vision [36, 59, 99, 102, 108]. We consider five popular methods. As the first adapter-based method, **Houl. Adapter** [36] inserts two Adapters (a two-layer bottleneck-structured MLP with a residual link) into each Transformer layer, one after the MSA block and one after the MLP block. **Pfeif. Adapter** [74] inserts the Adapter solely after the MLP block, shown effective in recent studies [37]. **AdaptFormer** [11] adds an Adapter in parallel with the original MLP block, different from the sequential design of previous methods. **ConvPass** [43] inserts a convolutional-based Adapter that explicitly encodes visual inductive biases by performing 2D convolution over nearby patch tokens. This module is inserted parallel to the MSA and MLP blocks. **RepAdapter** [63] introduces linear Adapters with group-wise transformations [62], placed sequentially after MSA and MLP.

Direct selective tuning. They selectively update a subset of parameters of the backbone, striking a balance between full FT and linear probing. We consider three approaches. **BitFit** [103] updates the bias terms, including those in the patch embeddings projection, the Q/K/V weights, the MLP and LN blocks. **LayerNorm** [7] updates the parameters of the LN blocks. **DiffFit** [97] updates both the bias terms and the LN blocks and inserts learnable factors to scale the features after the MSA and the MLP blocks. Instead of updating parameters, **SSF** [55] linearly adapts intermediate features, motivated by feature modulation [39].

Efficient selective tuning. Instead of directly updating pa-

Method	Natural								Specialized					Structured										
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Mean	Camelyon	EuroSAT	Resisc45	Retinopathy	Mean	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Elev	Mean	Overall Mean	Tunable Params
Linear	78.1	86.6	65.7	98.9	89.3	41.5	53.2	72.5	83.1	90.0	74.9	74.6	80.6	37.5	35.1	36.5	64.6	16.2	29.4	17.3	23.7	32.5	61.9	0
Full	62.4	89.9	61.9	97.4	85.8	88.9	36.8	76.7	81.6	88.1	81.6	73.6	81.2	56.2	60.9	48.2	77.9	68.5	46.6	31.0	28.3	52.2	70.0	85.8
VPT-Shallow	80.2	88.7	67.9	99.1	89.6	77.0	54.2	79.4	81.8	90.3	77.2	74.4	80.9	42.2	52.4	38	66.5	52.4	43.1	15.2	23.2	41.6	67.3	0.07
VPT-Deep	84.8	91.5	69.4	99.1	91.0	85.6	54.7	81.8	86.4	94.9	84.2	73.9	84.9	79.3	62.4	48.5	77.9	80.3	56.4	33.2	43.8	60.2	75.6	0.43
BitFit	86.5	90.5	70.3	98.9	91.0	91.2	54.2	82.6	86.7	95.0	85.3	75.5	85.6	77.2	63.2	51.2	79.2	78.6	53.9	30.1	34.7	58.5	75.6	0.1
DiffFit	86.3	90.2	71.2	99.2	91.7	91.2	56.1	83.2	85.8	94.1	80.9	75.2	84.0	80.1	63.4	50.9	81.0	77.8	52.8	30.7	35.5	59.0	75.4	0.14
LayerNorm	86.0	89.7	72.2	99.1	91.4	90.0	56.1	83.0	84.7	93.8	83.0	75.2	84.2	77.5	62.2	49.9	78.1	78.0	52.1	24.3	34.4	57.1	74.7	0.04
SSF	86.6	89.8	68.8	99.1	91.4	91.2	56.5	82.8	86.1	94.5	83.2	74.8	84.7	80.1	63.6	53.0	81.4	85.6	52.1	31.9	37.2	60.6	76.0	0.21
Pfeif. Adapter	86.3	91.5	72.1	99.2	91.4	88.5	55.7	83.0	86.2	95.5	85.3	76.2	85.8	83.1	65.2	51.4	80.2	83.3	56.6	33.8	41.1	61.8	76.9	0.67
Houl. Adapter	84.3	92.1	72.3	98	91.7	90.0	55.4	83.2	88.7	95.3	86.5	75.2	86.4	82.9	63.6	53.8	79.6	84.4	54.3	34.2	44.3	62.1	77.2	0.77
AdaptFormer	85.8	91.8	70.5	99.2	91.8	89.4	56.7	83.2	86.8	95.0	86.5	76.3	86.2	82.9	64.1	52.8	80.0	84.7	53.0	33.0	41.4	61.5	76.9	0.46
RepAdapter	86.0	92.5	69.1	99.1	90.9	90.9	55.4	82.9	86.9	95.3	86.0	75.4	85.9	82.5	63.5	51.4	80.2	85.4	52.1	35.7	41.7	61.6	76.8	0.53
Convpass	85.0	92.1	72.0	99.3	91.3	90.8	55.9	83.5	87.7	95.8	85.9	75.9	86.3	82.3	65.2	53.8	78.1	86.5	55.3	38.6	45.1	63.1	77.6	0.49
LoRA	85.7	92.6	69.8	99.1	90.5	88.5	55.5	82.6	87.5	94.9	85.9	75.7	86.0	82.9	63.9	51.8	79.9	86.6	47.2	33.4	42.5	61.0	76.5	0.55
FacT_TT	85.8	91.8	71.5	99.3	91.1	90.8	55.9	83.4	87.7	94.9	85.0	75.6	85.8	83.0	64.0	49.0	79.3	85.8	53.1	32.8	43.7	61.3	76.8	0.13
FacT_TK	86.2	92.5	71.8	99.1	90.1	91.2	56.2	83.4	85.8	95.5	86.0	75.7	85.8	82.7	65.1	51.5	78.9	86.7	53.1	27.8	40.8	60.8	76.6	0.23
Relative Std Dev	0.81	1.13	1.78	0.34	0.54	1.82	1.24	0.54	1.20	0.59	1.95	0.83	0.94	2.67	1.50	3.22	1.37	4.11	4.46	11.02	9.30	2.70	1.09	-

Table 1. PEFT methods performances on VTAB-1K (19 tasks from 3 groups) by TOP-1 ACCURACY. Based on the results across PEFT, linear probing, and full FT, we identify two task groups (purple and orange), as discussed in section 6.

rameters, these methods learns *additive residuals* (*e.g.*, $\Delta \boldsymbol{W}$) to the original parameters (*e.g.*, $\boldsymbol{W}$). By injecting a low-rank constraint to the residuals, this category effectively reduces the learnable parameters. **LoRA** [37], arguably the most well-known approach, parameterizes the residuals by low-rank decomposition to update the weights. Concretely, to update a $\boldsymbol{W} \in \mathbb{R}^{D\times D}$ matrix, LoRA learns $\boldsymbol{W}_{\text{down}} \in \mathbb{R}^{r\times D}$ and $\boldsymbol{W}_{\text{up}} \in \mathbb{R}^{D\times r}$ with $r \ll D$, and forms the additive residual by $\Delta \boldsymbol{W} = \boldsymbol{W}_{\text{up}}\boldsymbol{W}_{\text{down}} \in \mathbb{R}^{D\times D}$. **FacT** [44] extends the idea of matrix decomposition into tensor decomposition. It stacks the $D \times D$ learnable matrices in all the Transformer layers into a 3D tensor and learns an additive residual parameterized by Tensor-Train (TT) [72] and Tucker (TK) [14] formulations. More detailed summary of ViT and a survey of PEFT methods can be found in Appendix B.

2.3. Related work and comparison

The community-wide enthusiasm for PEFT has led to multiple survey articles [28, 99, 101]. Meanwhile, several empirical and theoretical studies were presented, mostly based on NLP tasks, attempting to provide a holistic understanding. [29, 67] provided unified views to methodologically connect PEFT methods. [10, 17, 31] and [32, 98] empirically compared PEFT methods on NLP and vision tasks, respectively, while [22] offered a theoretical stability and generalization analysis. Accuracy-wise, [10, 17, 31] found that PEFT is robust to over-fitting and quite effective in NLP tasks under low-data regimes. *This is, however, not the case for vision tasks: [32] showed that representative PEFT methods like LoRA and Adapter can't consistently outperform either full FT or linear probing.* In terms of why PEFT works, [22] framed PEFT as sparse fine-tuning and showed that it imposes a regularization by controlling stability; [17, 32] framed PEFT as (subspace) optimization; [17] further discussed the theoretical principle inspired by optimal control.

Our study strengthens and complements the above studies and offers new insights. First, we compared over *ten* PEFT methods and carefully tuned the hyperparameters, aiming to accurately assess each method's performance. This is particularly crucial for the vision community, where comprehensive references are still limited, and simpler methods like BitFit have often been deemed inferior, while other approaches have shown discrepancies compared to NLP studies. Second, we go beyond a *competition* perspective to investigate a *complementary* perspective of PEFT approaches. We show that different PEFT approaches offer *effective base learners for model ensembles*. Third, we go beyond downstream accuracy to investigate PEFT's effectiveness in maintaining *out-of-distribution robustness*. Finally, we analyze both *low-shot* and *many-shot* settings, revealing distinct patterns among PEFT, full FT, and linear probing, extending the understanding of PEFT.

3. PEFT Methods in Low-Shots Regime

Pre-trained models are meant to ease downstream applications. One representative scenario is low-shot learning: supervised FT of the pre-trained model with a small number of samples per class. Indeed, low-shot learning has been widely used to evaluate PEFT performance.

Dataset. VTAB-1K [104] consists of 19 classification tasks from three groups. **Natural** comprises natural images captured with standard cameras. **Specialized** contains images captured by specialist equipment for remote sensing and medical purposes. **Structured** evaluates scene structure comprehension, including object counting and depth estimation. Following [104], we split the **1000** training image

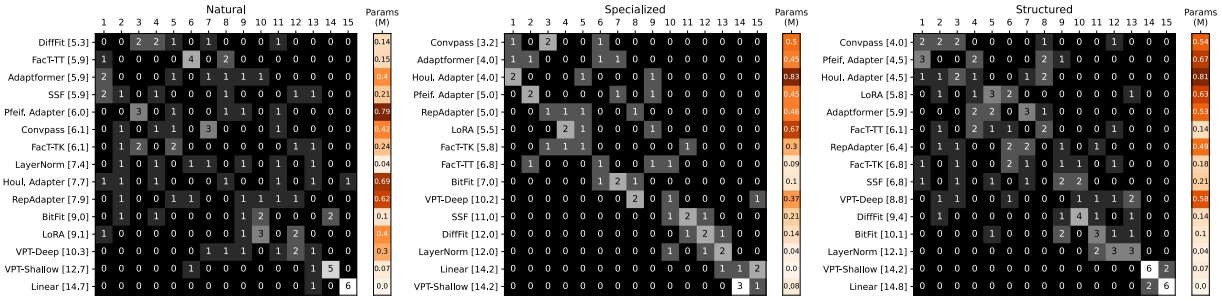


Figure 2. Ranking frequency of 15 methods (14 PEFT + linear probing) for three groups in VTAB-1K. Element (i, j) is the number of times method i ranks j^{th} in each group. Methods are ordered by mean ranks (in brackets). The parameters column shows the # of trainable parameters in millions. More details are in Appendix C.

80/20 for hyperparameter tuning. The reported TOP-1 ACCURACY is obtained after training over the 1000 images and evaluating on the original test set.

Methods. We consider linear probing, full FT, and **14** PEFT methods: 2 prompt-based [42], 5 adapter-based [11, 36, 43, 63, 74], 4 direct selective [7, 55, 97, 103], and 3 efficient selective [37, 44]. Please refer to subsection 2.2 and Appendix B for details.

Setup. We employ the ViT-B/16 [19] pre-trained on ImageNet-21K [15] as the backbone. The prediction head is randomly initialized for each dataset. We systematically tune 1) learning rate, 2) weight decay, and 3) method-specifics like the PEFT parameter sizes, which are often left intact in previous studies. We set a cap for PEFT size $\leq 1.5\%$ of ViT-B/16. We also turn the drop path rate [38] on (0.1) or off (0). A detailed experiment setup is provided in Appendix A, and more experiment results, including DINOv2 [71] and larger backbones (ViT-L and ViT-H), are provided in Appendix C.

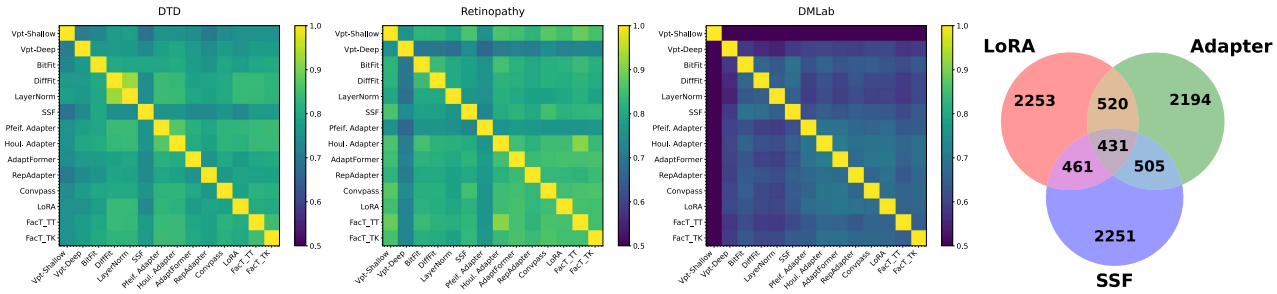
Results. As shown in Figure 1a and Table 1, PEFT methods generally outperform both linear probing and full FT across datasets. Additionally, with proper implementation and fair hyperparameter tuning, we surprisingly found that most PEFT methods perform similarly as the relative standard deviations (divided by the means) in all three groups are quite low. Simple methods (*e.g.*, Bitfit) and PEFT methods originally proposed for NLP (*e.g.*, LoRA and Adapter), which were previously reported as inferior due to unoptimized implementations and hyperparameter tuning, now demonstrate competitive performance with SOTA visual PEFT methods. To understand the relative advantages of different approaches, we provide the ranking frequency of PEFT methods across different groups in Figure 2, where the element (i, j) in each ranking matrix represents the frequency that method i ranks j^{th} in each group. Methods are ordered by their mean ranks (in brackets), and the parameters column indicates the number of trainable parameters in millions. In natural group, simpler methods with fewer trainable parameters—such as DiffFit and Fact-TT—offer a cost-effective solution without compromising performance. Conversely, in specialized and structured groups, methods with more parameters generally yield better performance. We hypothesize that this performance discrepancy arises from the **domain similarity** between the pre-trained domain (ImageNet) and the downstream domains. The natural group, sharing a stronger affinity with ImageNet, allows simpler methods like BitFit to adjust the features effectively. In contrast, the specialized and structured groups necessitate more complex methods with more trainable parameters to bridge the domain gap.

Recipes. In low-shot regimes, when the downstream data are similar to the pre-trained data, simple methods (*e.g.*, DiffFit) offer solid accuracy with fewer parameters. Conversely, if there is a substantial domain gap, more complex methods with more parameters often achieve higher accuracy. Since low-shot training is especially prone to over-fitting, we recommend activating a nonzero drop path rate—commonly set to zero by default—that stochastically drops a Transformer block per sample [38]. *All* methods can benefit from such a randomization-based regularization, as shown in Figure 11 in the Appendix.

4. Different PEFT Approaches Offer Complementary Information

The previous section demonstrates that all PEFT methods perform similarly across various domains. Given that different PEFT methods are trained on the same downstream data using the same backbone and achieve comparable accuracy, one might expect them to learn similar knowledge from the data, resulting in similar predictions. Contrary to this expectation, our findings below reveal that different PEFT methods acquire **distinct** and **complementary** knowledge from the *same* downstream data, even when built upon the *same* backbone, leading to **diverse** predictions.

We start by analyzing their prediction similarity on the same dataset in VTAB-1K. It is expected that their predictions are similar for datasets with very high accuracy, such as Flowers102 (avg 99.1%) and Caltech101 (avg 91.4%). Beyond them, we find that most PEFT methods show diverse predictions in other datasets in VTAB-1K. Figure 3a shows the prediction similarities between 14 PEFT methods



(a) (b)

Figure 3. (a) Prediction similarity analysis: element (i, j) shows the percentage of samples that method i and j predict the same. Although different methods achieve similar accuracy, they have diverse predictions. (b)The wrong prediction overlaps of LoRA, Adapter, and SSF for the 5K least confident data. Correct prediction overlaps for the 5K most confident data are shown in Figure 1b. They are FT on CIFAR100 (VTAB-1K). More results for (a) and (b) are in Appendix C.

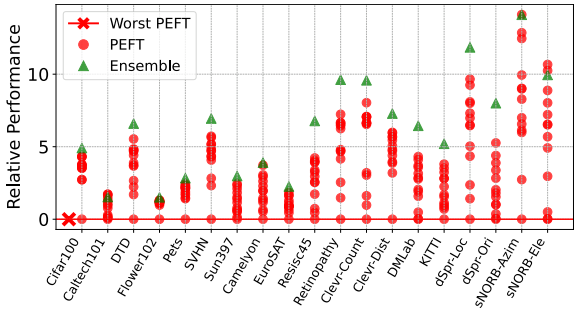


Figure 4. Ensemble (majority vote) shows consistent gain on most datasets thanks to the diverse predictions.

in DTD, Retinopathy, and DMLab, which belong to natural, specialized, and structured groups, respectively. In DTD and Retinopathy, most methods differ in about 20% of their predictions, while in DMLab, this difference increases to approximately 35%, even though they achieve similar accuracies. This prediction diversity may be attributed to the different *inductive biases* [70] of PEFT methods — they explicitly select specific parameters to update or insert different modules at various locations within the model. More analyses and details are offered in Appendix C.

Such diverse predictions across methods open up the possibility of leveraging their heterogeneity for further improvement. The most straightforward approach is ensemble [26], *e.g.*, majority vote over methods. Figure 4 demonstrates the ensemble performance gain over all the PEFT methods in each dataset, where we use the worst PEFT method as the baseline. Thanks to the diverse predictions across methods, the ensemble can provide consistent gain.

Also, we analyze if PEFT methods make similar correct predictions for high-confidence samples and similar mistakes for low-confidence samples. Figure 1b and Figure 3b show the correct prediction overlap for the 5K most confident samples (per method) and the wrong prediction overlap for the 5K least confident samples (per method). For demonstration purposes, we select one method from each PEFT category (LoRA, Adapter, SSF) and they are FT on CIFAR-100 in VTAB-1K. Methods within the same category also show diverse predictions (Appendix C). Since they make different predictions in both high and low-confidence regimes, this paves the way for new possibilities of using different PEFT methods to generate **diverse pseudo-labels** for semi-supervised learning [23, 24, 100], domain adaptation [20, 21, 86], and continual learning [60, 64–66, 82]. For example, in semi-supervised learning, we can FT different PEFT methods on the labeled data to generate *diverse and accurate pseudo-labels* by selecting highly confident predictions from each PEFT method.

5. PEFT Methods in Many-Shot Regime

Recent works in NLP [10] have indicated that PEFT methods may not perform as competitively as full FT when data is abundant. We thus aim to investigate PEFT's performance in many-shot regimes by addressing the following questions: (1) *Should we use PEFT or full FT when data is sufficient?* (2) *How should we adjust the number of trainable parameters for PEFT methods in many-shot regimes?*

Dataset. We select one representative dataset from each group in VTAB: (1) CIFAR-100 [49], a natural image dataset comprising 50K training images across 100 classes; (2) RESISC [12], a remote sensing dataset for scene classification with 25.2K training samples across 45 classes; and (3) Clevr-Distance [45, 104], a synthetic image dataset for predicting the depth of the closest object with 6 depth classes and 70K samples. The reported results are obtained by training on the **full** training set and evaluating on the original test set.

Setup. The model setup mostly follows the VTAB-1K experiment. More details about setup and hyperparameter search are provided in Appendix A.

Results. In many-shot regimes, with sufficient downstream data, full FT may catch up and eventually outperform PEFT methods. However, from Figure 5, we found that even in many-shot regimes, PEFT can achieve **comparable results with full FT**, even just using 2% of fine-tuning parameters. The performance gain, however, quickly **diminishes and**

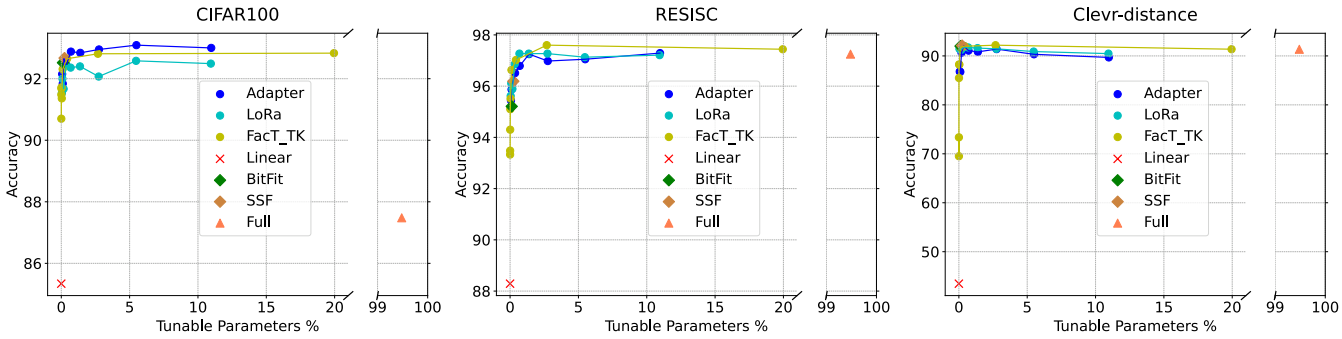


Figure 5. PEFT accuracy in many-shot regimes, with different parameter sizes (X-axis) on three datasets from different domains. Even 2%-5% trainable parameters allow the models to have sufficient capacity to learn from full data. Details are in Appendix C).

plateaus after 5% of tunable parameters. By comparing the results on the domain-close CIFAR-100 and domain-different RESISC and Clevr, we have some further observations. Downstream tasks with larger domain gaps often require updating more parameters to achieve high accuracy. With sufficient downstream data, full FT is less prone to over-fitting and, indeed, attains a high accuracy. But interestingly, PEFT methods, with only **2%∼5%** of tunable parameters, achieve similar accuracy, suggesting that its design principle does offer **sufficient effective capacity** for the model to learn [105]. Downstream tasks with smaller domain gaps suggest that the pre-trained model had learned sufficient knowledge about them; full FT thus risks washing such knowledge away. In fact, we found that PEFT notably outperforms full FT on CIFAR-100, suggesting it as a more *robust transfer learning* algorithm for downstream tasks.

Recipes. In many-shot regimes, PEFT methods with sufficient parameters (2∼5%) appear more favorable than full FT and linear probing. On the one hand, they achieve comparable and even better accuracy than full FT. On the other hand, the tunable parameters remain manageable. The parameter efficiency of PEFT also often implies *less training GPU memory usage and training time*, making PEFT methods a favorable alternative in many-shot regimes. For a downstream domain that is close to the pre-training domain, PEFT shows much pronounced *transferability*. For a downstream domain that is quite different, the limited tunable parameters (controversially, 2∼5% already amount to a few million) already allow the model to learn sufficiently.

6. Why Do PEFT Methods Work?[2]

Putting together section 3 and section 5, we identify two distinct patterns regarding the performance among linear probing, full FT, and PEFT. Within 19 VTAB-1K tasks, we see: (1) *Full FT outperforms linear probing.* As linear probing reflects the pre-trained feature quality for downstream tasks, case (1) suggests the necessity to update the backbone to close the gap between pre-trained and downstream domains. (2) *Linear probing surpasses full FT,* suggesting the pre-trained features are good enough (at least in a low-shot scenario). Recklessly updating them may risk over-fitting. Figure 6 (a-b) summarizes the low-shot accuracy comparison based on the categorization above; each line corresponds to one task. Linear probing, PEFT, and full FT are located in order, from left to right, to reflect their tunable parameter sizes. PEFT's superiority in both cases showcases its **capacity** to learn and its **regularization role** to prevent over-fitting.

We also draw the many-shot accuracy in Figure 6 (c-d) based on the same categorization: RESISC and Clevr in case (1), and CIFAR-100 in case (2). In the many-shot setting, full FT consistently outperforms linear probing, which seems to suggest *no more risk of over-fitting*. However, on CIFAR-100 (Figure 6 (d)), we again see a noticeable gap between PEFT and full FT, just like in Figure 6 (b). Such a concave shape reminds us of the long-standing under-fitting-over-fitting curve, suggesting that even with sufficient downstream data, full FT still risks over-fitting.

Considering PEFT's comparable performance to full FT on RESISC and Clevr with large domain gaps, we conclude that PEFT succeeds as a **high-capacity** learner equipped with an **effective regularizer**. The two roles trade-off well such that PEFT can excel in both low- and high-similarity domains under both low-shot and many-shot settings.

7. How Robust are PEFT Methods to Distribution Shifts?

Large pre-trained models such as CLIP [75] have demonstrated unprecedented zero-shot accuracy across diverse data distributions. However, recent studies [75, 96] have shown that FT on downstream data, while significantly boosting performance on the target distribution, often compromises the model's robustness to distribution shifts. Given that PEFT only updates a limited number of parameters in the model, we investigate whether PEFT can offer a more robust

[2]Our intention is not to offer a definitive explanation about why PEFT works. As discussed in subsection 2.3, there is no broadly accepted theory for PEFT's success. We hope our empirical findings can support the ongoing efforts to uncover the fundamental principles behind PEFT.

	Full	BitFit	Layer-Norm	Houl. Adapter	Adapt-Former	Rep-Adapter	Convpass	LoRA	FacT_TK
100-shot ImageNet	75.0	75.27	74.8	75.0	75.6	76.5	76.3	76.6	74.7
Avg. distribution shift Acc	42.5	55.4 (12.9)↑	55.9 (13.4)↑	56.9 (14.4)↑	56.1 (13.6)↑	56.2 (13.7)↑	54.7 (12.2)↑	55.9 (13.4)↑	56.1 (13.6)↑

Table 2. The "Avg. distribution shift Acc" denotes the average performance of ImageNet-(V2, S, R, A) evaluated on the CLIP model FT on ImageNet. (↑) indicates the gain over full FT.

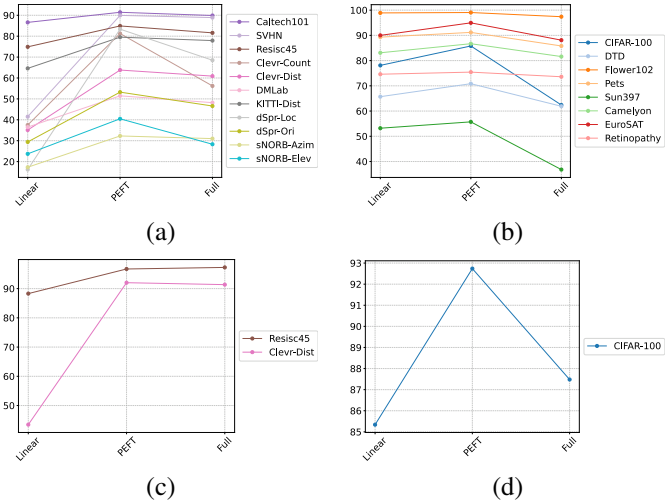


Figure 6. (a): VTAB-1K tasks in case (1), PEFT > full > linear. (b) VTAB-1K tasks in case (2), PEFT > linear > full. (c) RESISC & Clevr in case (1) with enough data, PEFT ≈ full > linear. (d) CIFAR in case (2) with enough data, PEFT > full > linear. Within each figure, left for linear, middle for PEFT, and right for full. More details are in Appendix C.

alternative to full FT.

Dataset. We use 100-shot ImageNet-1K as our target distribution, with each class containing 100 images. Following [96], we consider 4 natural distribution shifts from ImageNet: **ImageNet-V2** [78], a new ImageNet test set collected with the original labeling protocol; **ImageNet-R** [34], renditions for 200 ImageNet classes; **ImageNet-S** [25], sketch images for 1K ImageNet classes; **ImageNet-A** [35], a test set of natural images misclassified by a ImageNet pre-trained ResNet-50 [30] for 200 ImageNet classes.

Setup. We focus on the CLIP ViT-B/16 model, which comprises a visual encoder and a text encoder, pre-trained via contrastive learning on image-text pairs. Following [96], we add an FC layer as the head initialized using the class label text embedded by the text encoder. Subsequently, we discard the text encoder and apply PEFT methods to the visual encoder, FT only the PEFT modules and the head. More details about the CLIP model and experiment setup can be found in Appendix A.

Results. As shown in Table 2, while some PEFT methods may not surpass full fine-tuning on the target distribution, they consistently demonstrate more robust performance on distribution-shifted data. This robustness can be attributed to PEFT updating only a small fraction of the parameters, thereby preserving the robust features of the foundational models. Given the similar target distribution performance, *should we blindly use PEFT methods for more robustness?*

Weight-space ensembles (WiSE) for PEFT. WiSE [96], which linearly interpolates the weights of a fully FT model with those of the original backbone, is a popular approach for boosting robustness. We explore whether WiSE can also enhance the robustness of PEFT. To apply WiSE to PEFT, we first linearly interpolate the prediction head with a mixing coefficient α. For direct selective tuning methods (*e.g.* BitFit), this involves merging the PEFT-tuned parameters and the original model. Since most Adapter-based methods include residual connections (Appendix B.2.2), we can adjust their impact by scaling the adapter modules with α. A similar approach applies to efficient selective methods (*e.g.* LoRA) as they learn additive residuals to the original parameters. To the best of our knowledge, we are the **first** to study WiSE for PEFT. As shown in Figure 1c (more results in Appendix C), WiSE consistently improves both target and distribution shift performances of PEFT methods. For Adapter-based methods, WiSE can be considered feature ensembles, where α controls how strong the domain-specific features from the adapter module blend with the domain-agnostic ones from original backbones. For selective tuning, WiSE functions similarly to its application in full FT—exploiting the fact that the fine-tuned parameters remain near the original loss basin, allowing an ensemble that captures the best of both [2, 40].

Interestingly, even though full FT is generally less robust than PEFT methods, WiSE elevates full FT's performance above that of PEFT on both target distribution and distribution shift data, which suggests promising research directions to investigate the underlying mechanism and how to further improve the robustness of PEFT methods.

8. Conclusion

Instead of chasing the leaderboard, we conduct a unifying empirical study of PEFT, an emerging topic in the large model era. We provide an extendable framework for reproducible evaluations of PEFT methods in computer vision. We also have several new insights and implications, including PEFT methods' complementary expertise, suitable application regimes, and robustness to domain shifts. We expect our study to open new research directions and serve as a valuable user guide in practice.

Acknowledgment

This research is supported by grants from the National Science Foundation (ICICLE: OAC-2112606). We are grateful for the generous support from the Google gift fund and the Ohio Supercomputer Center.

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online, 2021. Association for Computational Linguistics. 21

[2] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*. 8

[3] Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. Attentional mixtures of soft prompt tuning for parameter-efficient multi-task knowledge sharing. *arXiv preprint arXiv:2205.11961*, 2022. 3

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3, 16

[5] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[6] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. *Advances in neural information processing systems*, 17, 2004. 2

[7] Samyadeep Basu, Shell Hu, Daniela Massiceti, and Soheil Feizi. Strong baselines for parameter-efficient few-shot fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11024–11031, 2024. 3, 5, 20, 21

[8] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 2

[9] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 3

[10] Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. Revisiting parameter-efficient tuning: Are we really there yet? *arXiv preprint arXiv:2202.07962*, 2022. 4, 6

[11] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 3, 5, 19

[12] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2, 6

[13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 3

[14] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000. 4, 21

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5, 15

[16] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 2

[17] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023. 1, 4

[18] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 21

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 5, 14

[20] Brunó B Englert, Fabrizio J Piva, Tommie Kerssies, Daan De Geus, and Gijs Dubbelman. Exploring the benefits of vision foundation models for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1172–1180, 2024. 6

[21] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021. 6

[22] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12799–12807, 2023. 4

[23] Kai Gan and Tong Wei. Erasing the bias: fine-tuning foundation models for semi-supervised learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 14453–14470, 2024. 6

[24] Kai Gan, Bo Ye, Min-Ling Zhang, and Tong Wei. Semi-supervised clip adaptation by enforcing semantic and trapezoidal consistency. In *The Thirteenth International Conference on Learning Representations*. 6

[25] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7457–7476, 2023. 2, 8

[26] Raphael Gontijo-Lopes, Yann Dauphin, and Ekin Dogus Cubuk. No one representation to rule them all: Overlapping features of training methods. In *International Conference on Learning Representations*, 2021. 6

[27] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 3, 17

[28] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024. 1, 4

[29] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 3, 4

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 8

[31] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021. 4

[32] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient fine-tuning for vision transformers. *arXiv preprint arXiv:2203.16329*, 3, 2022. 4

[33] Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, et al. Hyperprompt: Prompt-based task-conditioning of transformers. In *International Conference on Machine Learning*, pages 8678–8690. PMLR, 2022. 3

[34] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021. 2, 8, 15

[35] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples, 2021. 2, 8, 15

[36] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 1, 2, 3, 5, 19

[37] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 4, 5, 19, 20, 21

[38] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 2, 5, 22

[39] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 3, 22

[40] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*. 8

[41] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018. 21

[42] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 1, 3, 5, 14, 16, 17

[43] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 3, 5, 19

[44] Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1060–1068, 2023. 4, 5, 14, 16, 20, 21

[45] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2, 6

[46] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 15, 21

[47] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 3

[48] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 20

[49] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 6

[50] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2021. 20

[51] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics, 2021. 3

[52] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. 21

[53] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024. 3

[54] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023. 3, 17, 19

[55] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022. 3, 5, 14, 16, 20, 21

[56] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. *arXiv preprint arXiv:2403.14735*, 2024. 3

[57] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35, 2023. 3, 17

[58] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022. 3

[59] Yen-Cheng Liu, Chih-Yao Ma, Junjiao Tian, Zijian He, and Zsolt Kira. Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. *arXiv preprint arXiv:2210.03265*, 2022. 3

[60] Vincenzo Lomonaco, Lorenzo Pellegrini, Pau Rodriguez, Massimo Caccia, Qi She, Yu Chen, Quentin Jodelet, Ruiping Wang, Zheda Mai, David Vazquez, et al. Cvpr 2020 continual learning in computer vision competition: Approaches, results, current challenges and future directions. *Artificial Intelligence*, 303:103635, 2022. 6

[61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 14

[62] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Yan Wang, Liujuan Cao, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Towards lightweight transformer via group-wise transformation for vision-and-language tasks. *IEEE Transactions on Image Processing*, 31:3386–3398, 2022. 3, 20

[63] Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*, 2023. 3, 5, 14, 16, 19, 20

[64] Zheda Mai, Hyunwoo Kim, Jihwan Jeong, and Scott Sanner. Batch-level experience replay with review for continual learning. *arXiv preprint arXiv:2007.05683*, 2020. 6

[65] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3589–3599, 2021.

[66] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022. 3, 6, 19

[67] Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen-tau Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning. *arXiv preprint arXiv:2110.07577*, 2021. 4

[68] Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. UniPELT: A unified framework for parameter-efficient language model tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6253–6264. Association for Computational Linguistics, 2022. 3

[69] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 3

[70] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014. 2, 6

[71] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 5

[72] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011. 4, 21

[73] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 22

[74] JonLayer Noras Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *16th Conference of the European Chapter of the Associationfor Computational Linguistics, EACL 2021*, pages

487–503. Association for Computational Linguistics (ACL), 2021. 3, 5, 19

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 7

[76] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. 3, 19

[77] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018. 3, 19

[78] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. 2, 8, 15

[79] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 1

[80] Amir Rosenfeld and John K Tsotsos. Incremental learning through deep adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):651–663, 2018. 3, 19

[81] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[82] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9630–9638, 2021. 6

[83] Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, et al. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, 2022. 3

[84] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021. 3

[85] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 19

[86] Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23711–23720, 2024. 6

[87] Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. Convolutional visual prompt for robust visual perception. *Advances in Neural Information Processing Systems*, 36:27897–27921, 2023. 3

[88] Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7725–7735, 2023. 3, 17

[89] Cheng-Hao Tu, Hong-You Chen, Zheda Mai, Jike Zhong, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, and Wei-Lun Harry Chao. Holistic transfer: towards non-disruptive fine-tuning with partial target data. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 14

[91] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059. Association for Computational Linguistics, 2022. 3

[92] Benyou Wang, Yuxin Ren, Lifeng Shang, Xin Jiang, and Qun Liu. Exploring extreme parameter compression for pre-trained language models. In *International Conference on Learning Representations*, 2022. 21

[93] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 15

[94] Haixin Wang, Jianlong Chang, Yihang Zhai, Xiao Luo, Jinan Sun, Zhouchen Lin, and Qi Tian. Lion: Implicit vision prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5372–5380, 2024. 3

[95] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 14

[96] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 2, 3, 7, 8, 14, 15

[97] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via

simple parameter-efficient fine-tuning. *arXiv preprint arXiv:2304.06648*, 2023. 3, 5, 20, 21

[98] Yi Xin, Siqi Luo, Xuyang Liu, Haodi Zhou, Xinyu Cheng, Christina E Lee, Junlong Du, Haozhe Wang, MingCai Chen, Ting Liu, et al. V-petl bench: A unified visual parameter-efficient transfer learning benchmark. *Advances in Neural Information Processing Systems*, 37:80522–80535, 2024. 4

[99] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024. 1, 3, 4

[100] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022. 6

[101] Bruce XB Yu, Jianlong Chang, Haixin Wang, Lingbo Liu, Shijie Wang, Zhiyu Wang, Junfan Lin, Lingxi Xie, Haojie Li, Zhouchen Lin, et al. Visual tuning. *ACM Computing Surveys*, 2023. 1, 4

[102] Bruce XB Yu, Jianlong Chang, Haixin Wang, Lingbo Liu, Shijie Wang, Zhiyu Wang, Junfan Lin, Lingxi Xie, Haojie Li, Zhouchen Lin, et al. Visual tuning. *arXiv preprint arXiv:2305.06061*, 2023. 3, 17, 19

[103] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022. 2, 3, 5, 20, 21

[104] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1, 2, 4, 6, 14, 20

[105] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 2, 7

[106] Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Minivit: Compressing vision transformers with weight multiplexing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12145–12154, 2022. 21

[107] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. 3, 14, 16

[108] Henry Hengyuan Zhao, Pichao Wang, Yuyang Zhao, Hao Luo, Fan Wang, and Mike Zheng Shou. Sct: A simple baseline for parameter-efficient fine-tuning via salient channels. *International Journal of Computer Vision*, 132(3):731–749, 2024. 3

[109] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Panda: Prompt transfer meets knowledge distillation for efficient model adaptation. *arXiv preprint arXiv:2208.10160*, 2022. 3

[110] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012. 2

[111] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 3, 20