TTE: Two Tokens are Enough to Improve Parameter-Efficient Tuning

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Xian Gao, Suncheng Xiang, Ting Liu, Yuzhuo Fu*

Shanghai Jiao Tong University, China
jackchenruan@sjtu.edu.cn

Abstract

Existing fine-tuning paradigms are predominantly characterized by Full Parameter Tuning (FPT) and Parameter-Efficient Tuning (PET). FPT fine-tunes all parameters of a pre-trained model on downstream tasks, whereas PET freezes the pre-trained model and employs only a minimal number of learnable parameters for fine-tuning. However, both approaches face issues of overfitting, especially in scenarios where downstream samples are limited. This issue has been thoroughly explored in FPT, but less so in PET. To this end, this paper investigates overfitting in PET, representing a pioneering study in the field. Specifically, across 19 image classification datasets, we employ three classic PET methods (*e.g.*, VPT, Adapter/Adaptformer, and LoRA) and explore various regularization techniques to mitigate overfitting. Regrettably, the results suggest that existing regularization techniques are incompatible with the PET process and may even lead to performance degradation. Consequently, we introduce a new framework named TTE (Two Tokens are Enough), which effectively alleviates overfitting in PET through a novel constraint function based on the learnable tokens. Experiments conducted on 24 datasets across image and few-shot classification tasks demonstrate that our fine-tuning framework not only mitigates overfitting but also significantly enhances PET's performance. Notably, our TTE framework surpasses the highest-performing FPT framework (DR-Tune), utilizing significantly fewer parameters (0.15M *vs.* 85.84M) and achieving an improvement of 1%.

Code — https://github.com/JCruan519/TTE

Introduction

The fine-tuning paradigm utilizes parameters learned during the pre-training phase as the initialization for the fine-tuning process, leveraging prior knowledge for enhanced downstream adaptation. Existing fine-tuning paradigms can be categorized into Full Parameter Tuning (FPT) and Parameter-Efficient Tuning (PET) based on whether the model's pre-trained parameters are frozen. In FPT, the pre-trained parameters serve as the initialization, and all model parameters are fine-tuned to accommodate the distribution changes of downstream tasks. In PET, the backbone of the pre-trained model remains frozen, and only a small portion of parameters is introduced or adjusted to acquire the downstream knowledge. Benefiting from large-scale, pre-trained datasets and the diversity of downstream tasks, this fine-tuning paradigm has been extensively developed in various fields, including computer vision (Zhong et al. 2020; Zhang et al. 2021; Jia et al. 2022; Chen et al. 2022), natural language processing (Houlsby et al. 2019; Hu et al. 2021; Zaken, Ravfogel, and Goldberg 2021), and multimodal domains (Gao et al. 2021; Khattak et al. 2023; Gao et al. 2024).

However, both fine-tuning paradigms are afflicted by overfitting issues (Zheng et al. 2023; Zhuang et al. 2020), which detrimentally affect the fine-tuning performance on downstream tasks. In FPT, a variety of regularization frameworks have been proposed to mitigate this issue (Zhong et al. 2020; Zhang et al. 2021; You et al. 2020; Zhou, Chen, and Huang 2023). For instance, DR-Tune (Zhou, Chen, and Huang 2023), the leading FPT framework, integrates distribution regularization and semantic calibration techniques to avoid overfitting and improve performance. In PET, although considerable research has focused on optimizing performance via well-designed modules (Zhang, Zhou, and Liu 2022; Lian et al. 2022; Jie and Deng 2022; Luo et al. 2023a; Dong et al. 2023; Jiang et al. 2023), the overfitting issues and regularization frameworks remains underexplored.

In this paper, we conduct a pioneering study on the overfitting phenomena in PET, as shown in Figure 1. Specifically, we examine the performance of three classic PET architectures: VPT (Jia et al. 2022), Adapter/AdaptFormer[1] (Houlsby et al. 2019; Chen et al. 2022), and LoRA (Hu et al. 2021), on the VTAB-1K benchmark (Zhai et al. 2019). The results reveal that all these PET structures exhibit significant overfitting issues. In (Han et al. 2024), the authors addressed the overfitting of VPT by increasing the number of training samples. Correspondingly, we intuitively reduced the number of parameters introduced by PET while maintaining a fixed number of samples; however, this approach failed to significantly mitigate overfitting and led to decreased performance. Furthermore, we explored the use of various regularization strategies to mitigate the overfitting problem in

*Corresponding Author.

[1]Because Adapter and Adaptformer incorporate bottleneck modules in serial and parallel configurations, respectively, in this paper, we refer to Adapter as S-Ada. and Adaptformer as P-Ada. for simplicity.

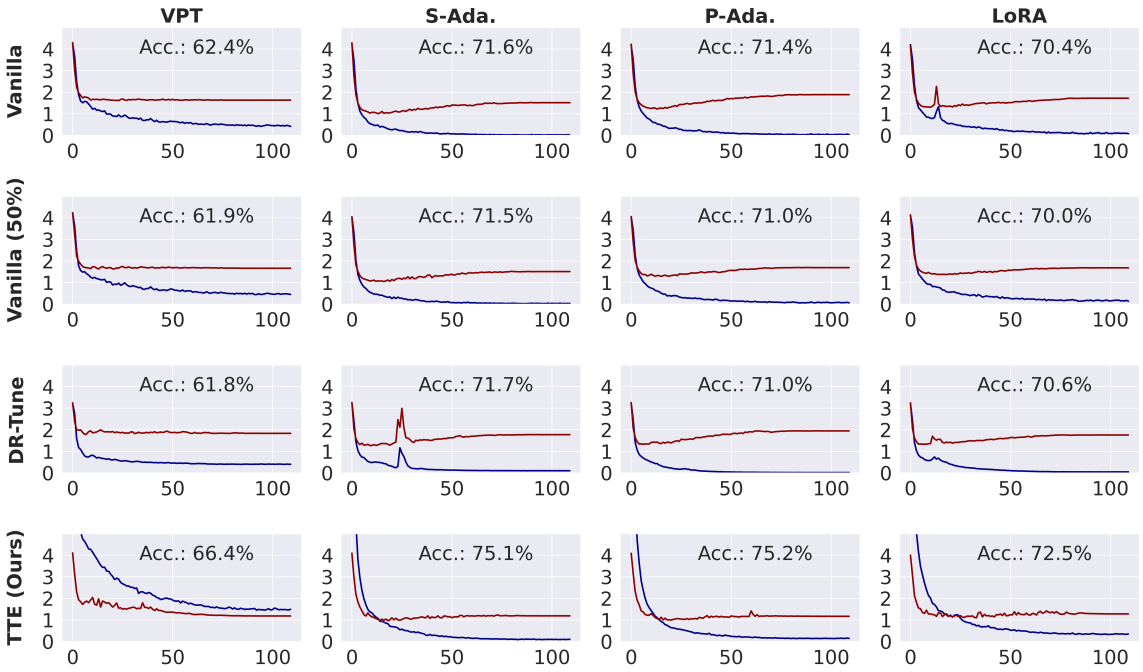

Figure 1: Performance and Average Train loss (Blue) vs. Eval loss (Red) on the VTAB-1K benchmark with different PET methods. The y-axis represents the loss value, and x-axis is the epoch. Vanilla (50%) indicates that the parameter count introduced by PET is reduced by half. For instance, Vanilla VPT introduces 20 learnable tokens, while Vanilla (50%) VPT introduces 10 learnable tokens. Besides, DR-Tune and TTE indicates that employing various regularization frameworks based on Vanilla.

PET. Unfortunately, the advanced DR-Tune (Zhou, Chen, and Huang 2023) regularization framework proved ineffective in adapting to the PET process and subsequently impaired the fine-tuning performance. Thus, the development of a specialized regularization framework tailored for the PET process has emerged as the central focus of this study.

Consequently, we propose a simple yet effective regularization framework named TTE (Two Tokens are Enough), designed to mitigate overfitting in PET methods and enhance performance on downstream tasks. Our TTE framework is built upon a learnable token module and an innovative constraint function, facilitating seamless integration into existing PET techniques. Specifically, the learnable token module comprises a globally learnable token (t^g), an instance-specific token (t^{in}) tailored according to the input, and an adaptive dropout operation. After concatenating the two tokens followed by adaptive dropout, they are further appended to the input sequence of the Transformer, contributing to both forward and backward computations alongside the learnable parameters introduced by PET. Furthermore, in classification tasks, in addition to utilizing cross-entropy loss for optimization, we also develop a regularization loss termed the Parameter-Free Cross Attention (PFCA) loss, based on t^g, t^{in}, and the inherent CLS token of the Transformer. This approach effectively alleviates overfitting and enhances the fine-tuning performance of PET methods. For instance, as shown in Figure 1, our TTE framework has achieved significant mitigation of the overfitting issue and an average improvement in performance by 3.36% on the VTAB-1K benchmark.

The principal contributions of this paper are summarized as follows: **1)** For the first time, we present an investigation of the overfitting issue in PET methods, and explore mitigation strategies through various regularization frameworks. Preliminary experiments suggest that existing regularization frameworks are incompatible with PET techniques and may even lead to diminished fine-tuning performance. **2)** We introduce a novel framework named TTE, which includes a learnable token module and a regularization constraint function, designed to alleviate the overfitting problem in existing PET methods. **3)** Extensive experiments are conducted on the VTAB-1K benchmark, encompassing 19 image classification datasets, and the FGVC benchmark, which includes 5 fine-grained few-shot datasets. The results demonstrate that our framework integrates seamlessly with existing PET processes, enhancing fine-tuning performance by an average of 3.36% on the VTAB-1K benchmark without a significant increase in parameter count. It is noteworthy that, with a significantly lower total number of trainable parameters compared to DR-Tune, the currently optimal FPT framework (85.84M *vs.* 0.15M), our TTE achieves superior performance (1% ↑).

Related Works

Regularization frameworks in FPT

In the study of the FPT framework, applying regularization techniques is a popular approach. In (Xuhong, Grandvalet,

and Davoine 2018), the authors utilize L2 penalty regularization to retain more pretrained model features, thereby enhancing transfer learning in convolutional networks. DR-Tune (Zhou, Chen, and Huang 2023) represents state-of-the-art work in the design of the FPT framework, introducing distribution regularization and semantic alignment to improve the fine-tuning process of pretrained visual models.

However, the regularization terms introduced in these methods are designed for the FPT framework, requiring all parameters to be updatable during fine-tuning. In the PET process, only a small fraction of parameters are updatable, posing challenges to designing compatible regularization terms. Thus, in this paper, we leverage the prior knowledge of pre-trained models along with downstream task knowledge to introduce more comprehensive and refined regularization constraints for PET methods.

Parameter-Efficient Tuning

Parameter-Efficient Tuning (PET) is a novel fine-tuning paradigm introduced in recent years, widely applied to the fine-tuning of Transformer-based models. PET techniques freeze the pre-trained backbone and incorporate learnable modules to obtain the specific knowledge of downstream tasks. According to (Yu et al. 2023), existing PET methods mainly include Prompt Tuning, Adapter Tuning, and Parameter Tuning, represented by VPT (Jia et al. 2022), S-Ada./P-Ada. (Houlsby et al. 2019; Chen et al. 2022), and LoRA (Hu et al. 2021), respectively.

Moreover, numerous innovative PET techniques such as SSF (Lian et al. 2022), FacT (Jie and Deng 2023), Res-Tuning (Jiang et al. 2023), RLRR (Dong et al. 2024), and others (Luo et al. 2023a; Fu, Zhu, and Wu 2024; Dong et al. 2023; Zhang et al. 2023; Yin, Li, and Zhang 2023; Jie, Wang, and Deng 2023; Ruan et al. 2024b,a,c) have also been proposed to improve performance. However, existing techniques mainly focus on well-designed modules but often overlook the overfitting issue and the importance of regularization constraints during fine-tuning. Therefore, we propose TTE framework, which introduces regularization constraints based on learnable tokens to further unleash the potential of PET methods.

Empirical Studies

In this section, we conduct an exploration of the overfitting phenomenon for three classical PET methods, including VPT (Jia et al. 2022), S-Ada./P-Ada (Houlsby et al. 2019; Chen et al. 2022). and LoRA (Hu et al. 2021) on the VTAB-1K benchmark (Zhai et al. 2019). The experimental settings are the same as illustrated in the Experiments Section.

As shown in Figure 1, the changes in loss values of these classic PET techniques during the training process are visualized. For the VTAB benchmark's 19 image classification tasks, each task is equipped with only 1,000 training samples and approximately 20,000 test samples. Thus, it is evident that due to the limited training samples for the downstream tasks, these PET techniques all suffer from significant overfitting issues. This phenomenon is similarly observed in (Han et al. 2024). Previous work (Han et al. 2024) has addressed overfitting by fixing the parameter count of VPT and increasing the number of training samples. Correspondingly, we first attempt the most intuitive strategy of fixing the number of training samples while reducing the parameter count of PET methods, with the results shown in the second row of Figure 1. It can be observed that reducing the number of parameters does not significantly impact overfitting and can lead to a degradation in performance. Thus, we attempt to introduce regularization constraint for mitigating the overfitting issues.

PET \ Reg.	Vanilla	L1	L2	USKD	DR-Tune	TTE
VPT	62.41	63.01	62.84	61.70	61.60	66.44
S-Ada.	71.55	72.13	72.03	72.91	71.41	75.06
P-Ada.	71.35	69.86	70.72	70.60	70.85	75.24
LoRA	70.43	70.17	70.01	68.46	70.56	72.45

Table 1: Mean Accuracy (%) on VTAB-1K. The existing regularization methods are insufficient to effectively alleviate the issue of overfitting in PET, and may even lead to decreased performance.

Further, we investigate three distinct regularization strategies to address the overfitting issue in PET: traditional methods (*e.g.*, L1 and L2 penalty), self-knowledge distillation-based regularization (*e.g.*, USKD (Yang et al. 2023)), and advanced frameworks employed in FPT, such as DR-Tune (Zhou, Chen, and Huang 2023). As demonstrated in Table 1, the application of conventional regularization techniques during the PET process generally results in varying degrees of performance degradation. For instance, even the advanced DR-Tune framework fails to impose effective constraints on PET, particularly for VPT, where the constraints of DR-Tune lead to a 0.81% performance decline. Furthermore, as illustrated in Figure 1, this observation underscores that DR-Tune is ineffective in mitigating the overfitting issue in PET.

One possible reason could be that most regularization methods are typically applied under the condition that all model parameters are updatable, a scenario suitable for pre-training and FPT. However, in PET, typically only about 0.5% of the parameters are learnable. In scenarios with limited training samples, when only a small number of parameters are updatable, the use of conventional regularization constraints may lead to local optima due to insufficiently comprehensive and precise constraints, thus failing to effectively mitigate overfitting (Xu et al. 2023; Luo et al. 2023b). Therefore, in this paper, we introduce learnable and instance-based tokens to capture global and instance-specific information, thereby formulating more comprehensive regularization constraints that effectively alleviate overfitting in the PET process and enhance performance.

Methods

TTE framework

In the process of Parameter-Efficient Tuning (PET) for a Vision Transformer (ViT) (Dosovitskiy et al. 2020) with N layers, the input image is transformed through an embedding

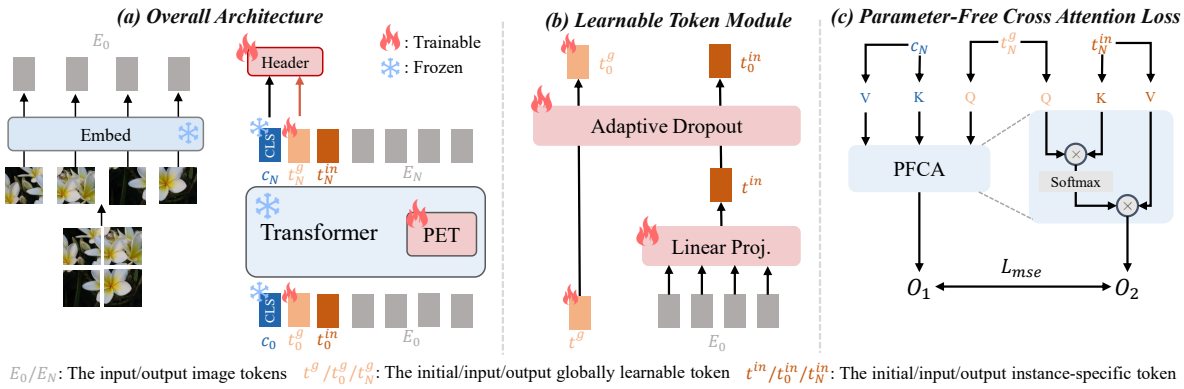

Figure 2: The overall architecture of our TTE framework, which consists of a Learnable Token Module and a Parameter-Free Cross Attention Loss.

operation into $E_0 \in \mathbb{R}^{l \times d}$, where l represents the sequence length and d stands for the embedding dimension. Subsequently, E_0 and the CLS token $c_0 \in \mathbb{R}^{1 \times d}$ are concatenated to form the input. The entire process can be represented as follows.

$$\begin{aligned} [c_i, E_i] &= L_i([c_{i-1}, E_{i-1}]; \Theta_i, \theta_i) \qquad i = 1, 2, \cdots, N \\ s_c &= \mathrm{Head}(c_N) \end{aligned} \tag{1}$$

where each layer L_i comprises a multi-head self-attention module (MHSA) and a feed-forward network (FFN). Θ_i denotes the backbone parameters of the i-th layer, which are frozen during the PET process. θ_i represents newly introduced learnable parameters, employed to acquire specific knowledge for downstream tasks. $[\cdot, \cdot]$ signifies the concatenation operation. Head is the trainable linear classification head, which generates logits (s_c) from the CLS token of the last layer (c_N).

In the TTE framework, as illustrated in Figure 2 (a), the PET process involves the introduction of two tokens: a globally learnable token (t^g) and a token derived from the image instance (t^{in}). These two tokens, together with the inherent CLS token of the ViT, facilitate the computation of the Parameter-Free Cross Attention loss, serving effectively as a regularization to mitigate overfitting. This process is delineated as follows:

$$\begin{aligned} [c_i, t_i^g, t_i^{in}, E_i] &= L_i([c_{i-1}, t_{i-1}^g, t_{i-1}^{in}, E_{i-1}]; \Theta_i, \theta_i) \qquad i = 1, 2, \cdots, N \\ s_c &= \mathrm{Head}(c_N); \quad s_g = \mathrm{Head}(t_N^g) \\ \mathcal{L}_{all} &= \mathcal{L}_{ce}(s_c, y) + \alpha \mathcal{L}_{ce}(s_g, y) + \beta \mathcal{L}_{pfca}(c_N, t_N^g, t_N^{in}) \end{aligned} \tag{2}$$

where the logits s_g is derived from the last layer's t_N^g via the classification head. $\mathcal{L}_{all}$, $\mathcal{L}_{ce}$, and $\mathcal{L}_{pfca}$ respectively denote the total loss, cross-entropy loss, and the Parameter-Free Cross Attention loss, which provide regularization constraints in the TTE framework. α and β represent the weight coefficients. Note that we only utilize s_c as the final decision for classification during inference.

Learnable Token Module

As depicted in Figure 2(b), we present our Learnable Token Module. It comprises a linear projection $W_{in} \in \mathbb{R}^{1 \times l}$ that extracts instance information from the input image, yielding $t^{in} \in \mathbb{R}^{1 \times d}$. Additionally, a globally learnable token ($t^g \in \mathbb{R}^{1 \times d}$) is introduced to capture the representation of the entire downstream dataset. A dropout operation is then applied to these two tokens, thereby enhancing the robustness of the instance information embedded in t^{in} and the global information inherent in t^g. However, given the diversity of downstream tasks, searching for a suitable mask ratio for each dataset proves impractical. Consequently, an adaptive dropout is introduced, applied to t^{in} and t^g, resulting in t_0^g and t_0^{in}, which are utilized in the subsequent input sequence. As illustrated in Figure 3, the adaptive dropout has a learnable mask ratio, enabling adaptive adjustment of the dropout intensity according to different downstream tasks.

Parameter-Free Cross Attention Loss

As illustrated in Figure 2 (c), we introduce the Parameter-Free Cross Attention Loss (PFCA loss). The loss function takes three inputs: $c_N \in \mathbb{R}^{1 \times d}$, $t_N^g \in \mathbb{R}^{1 \times d}$, and $t_N^{in} \in \mathbb{R}^{1 \times d}$, all derived from the final layer of the ViT. Without incorporating any linear transformation layers, c_N serves directly as both key and value, with t_N^g as the query. The query, key and value are input into the Parameter-Free Cross Attention (PFCA) mechanism, yielding the output $O_1 \in \mathbb{R}^{1 \times d}$. Similarly, t_N^{in} acts as both key and value with t_N^g as the query, subsequently entered into the PFCA to produce the output $O_2 \in \mathbb{R}^{1 \times d}$. Finally, the mean squared error between O_1 and O_2 is computed to serve as the output of PFCA loss.

How does TTE work?

The TTE framework introduces the globally learnable token t^g and the instance-based token t^{in}, which is derived from the input image instance. Moreover, the CLS token, encapsulating the pre-trained knowledge, is also utilized in the calculation of our PFCA loss. The CLS token is learnable during the pre-training phase while frozen during fine-tuning, hence the pre-trained knowledge it embodies remains unchanged during the PET process. Additionally, for t^g, an ex-

The Pytorch-style code for Adaptive Dropout

```
class AdaptiveDropout(nn.Module):
    def __init__(self, p=0.9):
        super(AdaptiveDropout, self).__init__()
        p = torch.log(torch.tensor(p / (1 - p)))
        self.p = nn.Parameter(p)
    def forward(self, x):
        if self.training:
            p = torch.sigmoid(self.p)
            mask_ratio = (1-p) * torch.ones_like(x)
            binary_mask = torch.Bernoulli(mask_ratio)
            return binary_mask * x / (1-p)
        return x
```

Figure 3: The Pytorch-style code for Adaptive Dropout.

tra cross-entropy loss is introduced between it and the true labels, enabling it to learn a more comprehensive representation of the downstream task. Finally, t^{in} varies according to the input image, capturing more specific and detailed instance information. In other words, through pre-training, the CLS token acquires the most extensive and general knowledge from the large-scale pre-training data. Via fine-tuning, t^{g} obtains global knowledge of the downstream task, which is less extensive than that captured by the CLS token. Furthermore, t^{in} represents the instance information of each input image, embodying the most detailed knowledge.

Subsequently, using t^{g}, which is positioned centrally in terms of knowledge breadth, as a medium, we perform a cross-attention calculation between t^{g} and CLS token to facilitate the interaction between downstream global information and pre-trained knowledge. The output is denoted as O_1. Another cross-attention calculation is conducted between t^{g} and t^{in} to promote the interaction between downstream global information and instance-specific information, with the resulting output denoted as O_2. Finally, a mean square error is calculated between O_1 and O_2 to construct a more comprehensive and detailed regularization constraint. This helps alleviate overfitting during the PET process and enhances fine-tuning performance.

Experiments

Datasets and metrics

Image classification tasks. We utilize the VTAB-1K benchmark (Zhai et al. 2019) to validate our TTE framework for image classification tasks. Specifically, VTAB-1K includes 19 different datasets, which can be categorized into three groups: Natural, Specialized, and Structured. Each dataset consists of 1,000 samples for training, with an average of 20,000 samples for testing, making it a highly challenging benchmark. Following the empirical setting (Lian et al. 2022), for each dataset, we report the Top-1 accuracy on the test set. For the entire benchmark, we present the arithmetic mean of the Top-1 accuracy.

Fine-grained few-shot tasks. In a few-shot setting, we validate the performance of our framework in the low-data regime using Food-101 (Bossard, Guillaumin, and Van Gool 2014), OxfordPets (Parkhi et al. 2012), Stanford Cars (Krause et al. 2013), Oxford-Flowers102 (Nilsback and Zisserman 2006), and FGVC-Aircraft (Maji et al. 2013) datasets. Following the empirical setting (Zhang, Zhou, and Liu 2022; Jie and Deng 2023), we conduct validation under $\{1, 2, 4, 8, 16\}$-shot settings and report the Top-1 accuracy.

Implementation details

For the VTAB-1K benchmark and FGVC datasets, we employ the ViT-B/16 (Dosovitskiy et al. 2020) model, pre-trained on the ImageNet-21K dataset (Deng et al. 2009), as the backbone. For PET methods, unless specifically stated otherwise, we fix the length of the learnable token introduced by VPT (Jia et al. 2022) at 20, set the hidden layer dimensions of S-Ada. (Houlsby et al. 2019) and P-Ada. (Chen et al. 2022) to 4, fix the scaling factor at 0.1, set the rank of LoRA (Hu et al. 2021) at 4, and the scaling factor at 1. In terms of training configurations, we follow the work of predecessors (Lian et al. 2022; Jie and Deng 2023; Luo et al. 2023a), to ensure fairness and reproducibility.

Regarding our TTE framework, to avoid redundancy brought about by further hyperparameter adjustment, we fix temperature α at 0.5, and only allow β to be searched from $\{0.25, 0.5, 0.75\}$. Pytorch (Paszke et al. 2019) and Transformers (Wolf et al. 2020) are utilized to implement experiments on NVIDIA A100 GPUs.

Main results

Comparative Results on VTAB-1K We conduct a comprehensive validation of the TTE framework using the VTAB-1K benchmark, with the results presented in Table 2. Initially, we apply the TTE framework to three distinct PET techniques—Prompt Tuning (VPT (Jia et al. 2022)), Adapter Tuning (S-Ada. (Houlsby et al. 2019) and P-Ada. (Chen et al. 2022)), and Parameter Tuning (LoRA (Hu et al. 2021))—as described in (Yu et al. 2023), during the fine-tuning phase. The results indicate that our framework significantly enhances these classical techniques by an average of 3.36%, while maintaining a comparable parameter count. Moreover, to our knowledge, P-Ada.† has surpassed recent state-of-the-art methods. For instance, when compared to RLRR (Dong et al. 2024), P-Ada.† achieves an improvement of 0.13%, whilst necessitating only 40% of the parameters.

Comparative Results on FGVC As shown in Figure 4, comprehensive validation is conducted in a few-shot scenario. The PET methods employed include VPT (Jia et al. 2022), S-Ada (Houlsby et al. 2019), and LoRA (Hu et al. 2021), each fine-tuned both with and without the proposed TTE framework. Overall, even within this constrained few-shot setting, various PET methods combined with our TTE framework are shown to improve performance without an increase in the number of trainable parameters. On average, for three different PET methods, our TTE framework achieves improvements of 2.27%, 1.52%, 1.35%, 1.41%, and 1.22% under the settings of {1,2,4,8,16}-shot.

Comparative Results with Full Parameter Tuning Framework As illustrated in Table 3, we conduct comparisons of TTE with Full Parameter Tuning frameworks in transfer learning (e.g., Core-tuning (Zhang et al. 2021) and DR-Tune (Zhou, Chen, and Huang 2023)). The results suggest that our P-Ada.† surpasses DR-Tune by 1%, currently

Methods	Comments	Natural							Specialized				Structured								All Mean	Δ	Param. (M)
		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			
Recent SOTA methods																							
ARC	NeurIPS23	71.2	90.9	75.9	99.5	92.1	90.8	52.0	87.4	96.5	87.6	76.4	83.3	61.1	**54.6**	81.7	81.0	**57.0**	30.9	41.3	74.27	-	0.13
Res-T.	NeurIPS23	75.2	92.7	71.9	99.3	91.9	86.7	**58.5**	86.7	95.6	85.0	74.6	80.2	63.6	50.6	80.2	85.4	55.7	31.9	42.0	74.09	-	0.55
SPT	Arxiv24	**79.3**	92.6	73.2	99.5	91.0	89.1	51.2	85.4	**96.8**	84.9	74.8	70.3	64.8	54.2	75.2	79.3	49.5	**36.5**	41.5	73.11	-	0.22
LAST	Arxiv24	66.7	93.4	76.1	99.6	89.8	86.1	54.3	86.2	96.3	86.8	75.4	81.9	**65.9**	49.4	82.6	**87.9**	46.7	32.3	**51.5**	74.15	-	0.66
RLRR	Arxiv24	76.7	92.7	76.3	99.6	**92.6**	**91.8**	56.0	87.8	96.2	**89.1**	76.3	80.4	63.3	54.5	83.3	83.0	53.7	32.0	41.7	75.11	-	0.33
Prompt tuning methods																							
VPT	ECCV22	60.5	90.6	70.6	99.1	89.3	50.1	50.8	82.2	93.8	82.5	74.9	50.6	58.9	41.0	68.1	39.0	32.4	22.3	29.1	62.41	-	0.06
VPT†	Ours	66.5	92.8	73.2	99.2	89.9	84.0	50.5	81.5	94.0	83.4	73.8	48.4	63.1	43.0	73.6	57.0	35.7	24.3	28.5	66.44	**4.03**↑	0.06
Adapter tuning methods																							
S-Ada.	ICML19	70.1	93.5	74.9	99.5	91.7	87.2	51.4	86.9	96.6	87.8	**76.9**	84.3	34.5	53.8	80.2	72.8	54.8	22.4	40.2	71.55	-	0.13
S-Ada.†	Ours	77.4	**94.5**	77.1	**99.7**	**92.6**	90.0	55.2	**88.4**	96.2	**89.1**	76.2	**85.5**	62.7	52.6	83.0	80.1	54.0	28.7	43.1	75.06	**3.51**↑	0.13
P-Ada.	NeurIPS22	70.4	92.2	74.5	99.4	91.3	79.1	51.5	83.2	96.4	87.7	76.2	83.7	60.3	53.5	74.6	56.6	54.4	28.3	42.3	71.35	-	0.13
P-Ada.†	Ours	78.6	94.0	**77.5**	**99.7**	92.3	90.0	55.8	87.8	96.5	88.5	76.8	84.3	62.6	53.3	**83.8**	81.1	52.9	28.9	45.1	**75.24**	**3.89**↑	0.13
Parameter tuning methods																							
LoRA	ICLR21	65.9	91.3	73.6	99.3	91.7	83.2	51.0	84.0	96.2	87.3	76.2	71.2	57.8	50.5	78.1	58.4	53.2	28.1	41.1	70.43	-	0.19
LoRA†	Ours	67.4	91.9	75.4	99.5	91.1	87.2	51.5	86.7	95.6	86.5	76.3	76.7	60.5	49.1	82.2	76.3	50.2	29.4	43.1	72.45	**2.02**↑	0.19

† denotes employing PET techniques within our TTE framework.

Table 2: Performance and efficiency comparison on the VTAB-1K benchmark with ViT-B/16 (Dosovitskiy et al. 2020) pre-trained on ImageNet-21K (Deng et al. 2009). "All Mean" denotes the average accuracy of 19 tasks.

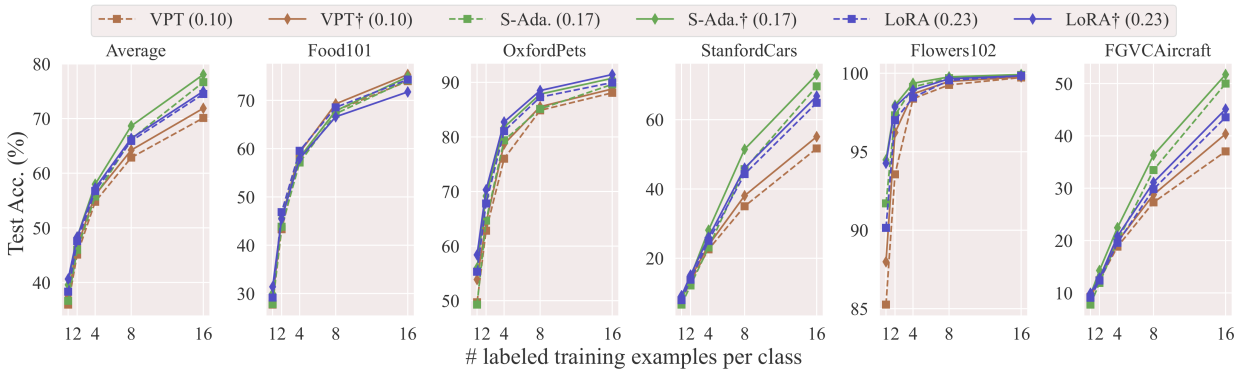

Figure 4: Top-1 accuracy of few-shot learning on FGVC datasets. The trainable parameters (M) is shown in parentheses.

Method	Params. (M)	Mean	CIFAR-100	Caltech101	DTD	Flowers102	Pets
P-Ada.†(ours)	**0.15**	**88.4**	78.6	**94.0**	**77.5**	**99.7**	92.3
Core-tuning	85.84	83.6	66.3	89.7	70.9	99.0	92.3
DR-Tune	85.84	87.4	**81.1**	92.8	71.4	99.3	**92.4**

Table 3: Comparative results with other FPT frameworks.

recognized as the state-of-the-art FPT framework. Furthermore, it is significant that we achieve a reduction in the trainable parameters during the fine-tuning process by **572x**.

Ablation studies

Extensive ablation experiments are conducted on the VTAB-1K benchmark. Unless otherwise specified, the ViT-B/16, pre-trained on the ImageNet-21K, is employed as the backbone, and P-Ada. (Chen et al. 2022) is utilized as the PET method. Furthermore, the symbol † indicates the use of the PET method within our TTE framework, and the arithmetic mean of the Top-1 accuracy is displayed.

The impact of loss function As shown in Equation 2, in addition to the standard classification loss function $\mathcal{L}_{ce}(S_c, y)$, our TTE framework employs two additional loss functions to optimize the PET process: $\mathcal{L}_{ce}(S_g, y)$ and

L_1	L_2	L_3	Acc.
✓			71.38
✓	✓		72.88
✓		✓	74.30
✓	✓	✓	75.24

(a)

Loss	Acc.
$\mathcal{L}_{mse}$	75.24
$\mathcal{L}_{mae}$	74.19
$\mathcal{L}_{cos}$	73.27

(b)

Len.	Param. (K)	Acc.
1	1.0	75.24
5	4.8	72.15
10	9.7	69.97

(c)

Ratio	Acc.
0.1	72.92
0.5	74.35
0.9	75.24

(d)

Method	Param. (M)	Acc.
S+P-Ada.	0.07	70.92
S+P-Ada.†	0.07	**72.52**
L+P-Ada.	0.30	71.46
L+P-Ada.†	0.30	**75.36**

(e)

Method	Param. (M)	Acc.
FT	86.7	72.46
LP	0.1	58.19
P-Ada.	0.21	73.16
P-Ada.†	0.21	**75.43**

(f)

Table 4: The main ablation studies for our TTE framework: (a) loss function utilization in TTE (L_1 is $\mathcal{L}_{ce}(S_c, y)$, L_2 is $\mathcal{L}_{ce}(S_g, y)$, L_3 is $\mathcal{L}_{pfca}$). (b) various losses in PFCA loss. (c) token length for t^g and t^{in}. (d) initialization for mask ratio in Adaptive Dropout. (e) different size of backbone (S is ViT-S/16, L is ViT-L/16). (f) Backbone is Swin-B. Acc.: Top-1 Mean accuracy on the VTAB-1K (%). For (a)∼(d), our baseline is Vanilla P-Ada., which achieves 71.35% Acc. on the VTAB-1K.

$\mathcal{L}_{pfca}$. As demonstrated in Table 4a, we present the performance of different combinations. Clearly, as the number of loss terms increases, the effectiveness of our method also improves. Firstly, by introducing $\mathcal{L}_{ce}(S_g, y)$ on top of $\mathcal{L}_{ce}(S_c, y)$, we achieve a 1.5% gain, indicating that providing real labels to the globally learnable token (t^g) allows it to acquire more accurate global information. Secondly, the introduction of $\mathcal{L}_{pfca}$ on top of $\mathcal{L}_{ce}(S_c, y)$ results in nearly a 3% improvement. This significant enhancement further underscores the importance of regularization constraints between pretrained knowledge (represented by the CLS token), global information (represented by t^g), and instance information (represented by t^{in}).

Furthermore, as depicted in Figure 2, we use the mean squared error loss at the end of the PFCA loss to narrow the gap between general knowledge and downstream knowledge. As alternatives to the mean squared error loss, we employ mean absolute error loss and cosine loss to demonstrate the versatility of PFCA loss. As shown in Table 4b, even with the simpler $\mathcal{L}_{mae}$ and $\mathcal{L}_{cos}$, our TTE framework achieves improvements of 2.84% and 1.92% compared to the Vanilla P-Ada., respectively. This adequately demonstrates that constructing regularization constraints to bridge the gap between global and instance information is highly effective in the PET process.

The impact of token length We increase the token lengths of t^g and t^{in}, as detailed in Table 4c. Observations indicate that as the introduced token length increases, the fine-tuning performance significantly decreases. A possible explanation for this phenomenon is that the PET introduces a relatively small amount of trainable parameters during fine-tuning. If our TTE framework further introduces an excessive amount of parameters, it may lead to insufficient learning, resulting in constraint deviation and even a decline in performance.

The impact of the initialization of mask ratio In Table 4d, we compare the effects of various initial values of the mask ratio for Adaptive Dropout. The results suggest that as the initial mask ratio increases, our method exhibits enhanced performance. Consequently, in this paper, we have selected an initial mask ratio of 0.9.

The impact of different backbones First, to illustrate the versatility of our TTE across models of varying sizes, we substitute ViT-B/16 with ViT-S/16 and ViT-L/16, as detailed in Table 4e. Next, to highlight our framework's adaptability to different network structures, we conduct experiments using Swin-B (Liu et al. 2021) as the backbone, as presented in Table 4f. As evident from Tables 4e and 4f, regardless of whether we modify the model size or transition to an alternate backbone, our TTE consistently bolsters performance without an increase in parameters.

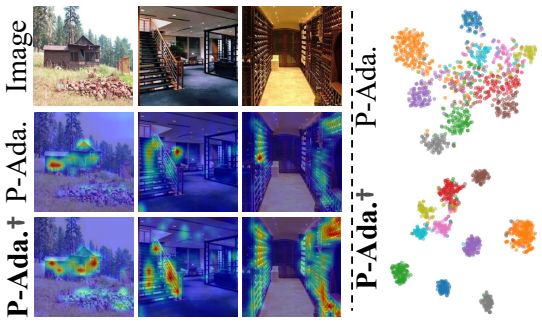

Figure 5: *Left:* The attention map visualization on Sun397 dataset. *Right:* The t-SNE visualization on SVHN dataset.

Visualization

We perform attention map and t-SNE (Van der Maaten and Hinton 2008) visualization analysis, as depicted in Figure 5. For this purpose, we extract the CLS token following the final Transformer layer and preceding the linear classification head. Notably, upon integrating the TTE framework, attention becomes more focused on the target object, and the classification clusters become more condensed.

Conclusions

In this paper, we first conduct a preliminary exploration of the overfitting phenomenon during the PET process and find that existing regularization methods are incompatible with PET. Further, we propose TTE framework, which utilizes both global and instance tokens to fully capture downstream information and construct regularization constraints with pre-trained knowledge. The TTE effectively mitigates overfitting in the PET process and further enhances fine-tuning performance, while introducing minimal additional parameters. Extensive experiments demonstrate the effectiveness and universality of our framework.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61977045).

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 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*, 446–461. Springer.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, W.; Yan, D.; Lin, Z.; et al. 2023. Efficient Adaptation of Large Vision Transformer via Adapter Re-Composing. In *NeurIPS*.

Dong, W.; Zhang, X.; Chen, B.; Yan, D.; Lin, Z.; Yan, Q.; Wang, P.; and Yang, Y. 2024. Low-Rank Rescaled Vision Transformer Fine-Tuning: A Residual Design Approach. *arXiv preprint arXiv:2403.19067*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fu, M.; Zhu, K.; and Wu, J. 2024. Dtl: Disentangled transfer learning for visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12082–12090.

Gao, J.; Ruan, J.; Xiang, S.; Yu, Z.; Ji, K.; Xie, M.; Liu, T.; and Fu, Y. 2024. LAMM: Label Alignment for Multi-Modal Prompt Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1815–1823.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Han, C.; Wang, Q.; Cui, Y.; Wang, W.; Huang, L.; Qi, S.; and Liu, D. 2024. Facing the Elephant in the Room: Visual Prompt Tuning or Full Finetuning? *arXiv preprint arXiv:2401.12902*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2790–2799. PMLR.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Jiang, Z.; Mao, C.; Huang, Z.; et al. 2023. Res-Tuning: A Flexible and Efficient Tuning Paradigm via Unbinding Tuner from Backbone. In *NeurIPS*.

Jie, S.; and Deng, Z.-H. 2022. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*.

Jie, S.; and Deng, Z.-H. 2023. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1060–1068.

Jie, S.; Wang, H.; and Deng, Z.-H. 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.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Lian, D.; Zhou, D.; Feng, J.; and Wang, X. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Luo, G.; Huang, M.; Zhou, Y.; Sun, X.; Jiang, G.; Wang, Z.; and Ji, R. 2023a. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*.

Luo, X.; Wu, H.; Zhang, J.; Gao, L.; Xu, J.; and Song, J. 2023b. A closer look at few-shot classification again. In *International Conference on Machine Learning*, 23103–23123. PMLR.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nilsback, M.-E.; and Zisserman, A. 2006. A visual vocabulary for flower classification. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, 1447–1454. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ruan, J.; Gao, J.; Xie, M.; Dong, D.; Xiang, S.; Liu, T.; and Fu, Y. 2024a. iDAT: inverse Distillation Adapter-Tuning. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Ruan, J.; Gao, J.; Xie, M.; Xiang, S.; Yu, Z.; Liu, T.; Fu, Y.; and Qu, X. 2024b. GIST: Improving Parameter Efficient Fine-Tuning via Knowledge Interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8835–8844.

Ruan, J.; Gao, X.; Xiang, S.; Xie, M.; Liu, T.; and Fu, Y. 2024c. Understanding Robustness of Parameter-Efficient Tuning for Image Classification. *arXiv preprint arXiv:2410.09845*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

Xu, C.; Yang, S.; Wang, Y.; Wang, Z.; Fu, Y.; and Xue, X. 2023. Exploring efficient few-shot adaptation for vision transformers. *arXiv preprint arXiv:2301.02419*.

Xuhong, L.; Grandvalet, Y.; and Davoine, F. 2018. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, 2825–2834. PMLR.

Yang, Z.; Zeng, A.; Li, Z.; Zhang, T.; Yuan, C.; and Li, Y. 2023. From Knowledge Distillation to Self-Knowledge Distillation: A Unified Approach with Normalized Loss and Customized Soft Labels. *arXiv preprint arXiv:2303.13005*.

Yin, D.; Li, L. H. B.; and Zhang, Y. 2023. Adapter is All You Need for Tuning Visual Tasks. *arXiv preprint arXiv:2311.15010*.

You, K.; Kou, Z.; Long, M.; and Wang, J. 2020. Co-Tuning for Transfer Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 17236–17246. Curran Associates, Inc.

Yu, B. X.; Chang, J.; Wang, H.; Liu, L.; Wang, S.; Wang, Z.; Lin, J.; Xie, L.; Li, H.; Lin, Z.; et al. 2023. Visual Tuning. *arXiv preprint arXiv:2305.06061*.

Zaken, E. B.; Ravfogel, S.; and Goldberg, Y. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Zhai, X.; Puigcerver, J.; Kolesnikov, A.; Ruyssen, P.; Riquelme, C.; Lucic, M.; Djolonga, J.; Pinto, A. S.; Neumann, M.; Dosovitskiy, A.; et al. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*.

Zhang, Q.; Zou, B.; An, R.; et al. 2023. Split & Merge: Unlocking the Potential of Visual Adapters via Sparse Training. *arXiv preprint arXiv:2312.02923*.

Zhang, Y.; Hooi, B.; Hu, D.; Liang, J.; and Feng, J. 2021. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. *Advances in Neural Information Processing Systems*, 34: 29848–29860.

Zhang, Y.; Zhou, K.; and Liu, Z. 2022. Neural prompt search. *arXiv preprint arXiv:2206.04673*.

Zheng, H.; Shen, L.; Tang, A.; Luo, Y.; Hu, H.; Du, B.; and Tao, D. 2023. Learn from model beyond fine-tuning: A survey. *arXiv preprint arXiv:2310.08184*.

Zhong, J.; Wang, X.; Kou, Z.; Wang, J.; and Long, M. 2020. Bi-tuning of pre-trained representations. *arXiv preprint arXiv:2011.06182*.

Zhou, N.; Chen, J.; and Huang, D. 2023. DR-Tune: Improving Fine-tuning of Pretrained Visual Models by Distribution Regularization with Semantic Calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1547–1556.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.