Bootstrap Prompt Learning with Feature Adaptation for Vision-Language Efficient Tuning

Anonymous Authors[1]

Abstract

As popular alternatives to fine-tune vision-language foundation models such as CLIP, prompt learning and adapter tuning resort to pre-adjustment in the input space and post-adjustment on the pretrained weight matrices to optimize the task-specific objective, respectively. However, there still lacks a method to jointly exploit their benefits due to potential conflicts in optimization directions. In this paper, we propose a novel framework named ada**P**ter bootstr**A**pped prompt contrastive **T**uning (PAT) to address this problem. Specifically, we bootstrap prompt learning with adapters and achieves pre-post alignment to avoid mismatch between the optimization directions of prompter learning and adapter tuning. Furthermore, we propose a tolerance regularization that equally pushes away all negative samples and improves generalization by introducing additional categories of unlabeled data to avoid overfitting. To our best knowledge, this is the first successful attempt to simultaneously exploit the advantages of prompt learning and adapter tuning. Extensive evaluations demonstrate that PAT achieves state-of-the-art performance in various recognition tasks on three prevailing benchmarks.

1. Introduction

Vision-language models pretrained on large-scale datasets of image-text pairs exhibit strong generalization capabilities across various downstream tasks (Alayrac et al., 2022; Radford et al., 2021; Jia et al., 2021). However, pretraining these models requires massive volume of image-text pairs and substantial computational resources. To address these challenges, parameter-efficient fine-tuning (PEFT)(Han et al., 2024) has been widely studied in recent literature. Compared to full fine-tuning, PEFT achieves competitive or superior performance by tuning a minimal number of trainable parameters. Generally, existing methods for PEFT can be categorized into three types, including prompt learning(Jia et al., 2022), adapters(Chen et al., 2022), and reparameterization(Hu et al., 2021).

Prompt learning introduces trainable embeddings into the input space to guides pretrained models to adapt to downstream tasks. In the context of vision-language adaptation, prompt learning methods such as CoOp (Zhou et al., 2022b) combine classification labels with a classification template and add trainable text embeddings, thereby converting the text encoder into a classifier. However, early methods for prompt learning in VLMs are limited in the input space and restricted in representation capacity. Recent efforts facilitate prompt learning with adapter-based feature adaptation. For example, in addition to trainable prompt embeddings, TCP (Yao et al., 2024) introduces textual knowledge embedding (TKE) that serves as a specialized adapter to learn class-level features and transform them into prompts. DePT (Zhang et al., 2024) incorporates a channel adjusted transfer (CAT) head into prompt learning, which resembles an adapter in implementation. However, existing methods for vision-languages efficient tuning suffer from two problems.

- **Mismatch between prompt learning and adapter tuning.** Existing methods simply introduce adapters to enhance the representation capacity of prompt learning, and neglect the mismatch between optimization directions of adapters with trainable parameters inserted alongside parameter matrices and prompts with trainable parameters inserted into the input space.
- **Bias by exclusive cross entropy loss.** We reveal that existing methods rely heavily on exclusive cross entropy loss. They select the class with the highest similarity between fine-tuned visual and textual representations as final prediction and undermine the generalization to unseen categories. Table 1 shows that the forced-choice constraint causes incorrect bias toward the categories given in the few-shot tuning and results in a loss of information about unseen categories.

To address these problems, in this paper, we propose a novel

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

vision language efficient tuning framework, namely adaPter bootstrAp prompt contrastive Tuning (PAT), which for the first time simultaneously exploit the advantages of prompt learning and adapter tuning. PAT incorporates two novel modules, *i.e.*, pre-post alignment to address the mismatch between pre-adjustment and post-adjustment and tolerance regularization to mitigate the bias caused by exclusive cross-entropy loss. Our contributions are summarized as below.

- We develop the ada**P**ter bootstr**A**p prompt contrastive **T**uning (PAT) framework to simultaneously improve the fitting capability and generalization performance of prompt learning in downstream tasks.
- We bootstrap the pre-adjustment with prompt learning by integrating post-adjustment with adapters and introduce a pre-post alignment module to avoid mismatch between the optimization directions of prompt learning and adapter tuning.
- We propose a tolerance regularization, which equally pushes away all negative samples and improves generalization by introducing additional categories of unlabeled data to prevent the model from over-fitting on the training categories.

We conduct extensive evaluations of PAT's fitting and generalization capabilities across over ten datasets and perform detailed ablation studies, demonstrating through experimental results that PAT achieves state-of-the-art performance. Specifically, our method improves overall performance in Base-to-New by 0.9% compared to the previous state-of-the-art (79.5% vs 80.4%) and achieves a 1.5% improvement in the few-shot experiment (76.7% vs 78.2%).

2. Related Work

2.1. Vision Language Models

Vision-language models (VLMs) learn multi-modal representations by pretraining on large-scale image-text datasets, such as CLIP(Radford et al., 2021) and ALIGN(Jia et al., 2021), with 400 million and 1 billion pairs, respectively. Using contrastive loss, these models align paired features while distinguishing unpaired ones, enabling strong open-vocabulary generalization. Recent advancements enhance their descriptive and discriminative capabilities through stronger encoders(Li et al., 2023; Vaswani, 2017), deeper modality fusion, larger datasets, and techniques like Masked Language Modeling (MLM) and image masking(Kim et al., 2021; Lu et al., 2019). CLIP, a key framework with exceptional generalization, has inspired numerous CoOp-based prompt tuning approaches. In this work, we propose a novel prompt learning framework to further adapt pretrained CLIP for generalization and few-shot learning.

2.2. PEFT for Vision Language Models

Prompt learning, as a parameter-efficient fine-tuning method, aims to transfer pretrained models to downstream tasks while keeping most parameters frozen. Classical prompt learning methods achieve this by adding a small number of trainable embeddings into the input space of pretrained models without altering the pretrained weights, thereby guiding the model's outputs to adapt to downstream tasks. Due to its efficiency in terms of trainable parameters, developing more powerful prompt learning methods for adapting multimodal pretrained models like CLIP to visual or vision-text downstream tasks has garnered significant interest from both academia and industry. For example, Context Optimization (CoOp) (Zhou et al., 2022b) replaces handcrafted prompts with learnable embeddings in the input space of CLIP's text encoder to enable few-shot adaptation. Recently, Textual-based Class-aware Prompt (TCP) (Yao et al., 2024) proposed another paradigm, focusing on class-aware prompt tuning and try to combine adapter and prompt learning. To mitigate potential knowledge forgetting during fine-tuning, Knowledge-Guided Context Optimization (KgCoOp) (Yao et al., 2023) applies L2 norm constraints to the text encoder, thus enhancing generalization. All the aforementioned methods focus on single-modal encoder fine-tuning and cannot flexibly adjust the representations of both modalities based on downstream tasks. Thus Multi-modal Prompt Learning (MaPLe) (Khattak et al., 2023a) improves the consistency between visual and language representations using trainable prompts and a vision-language coupling function, thereby enhancing the generalization of prompt learning.

Unlike these methods, we observe that while both are parameter-efficient fine-tuning approaches, adapter-based methods differ from prompt learning in their focus. Instead of modifying the input space of pretrained models as prompt learning does, adapter-based methods insert a small number of trainable parameters alongside the pretrained modules. This suggests that the two approaches may employ different optimization strategies to acquire knowledge beneficial for downstream tasks. In this paper, we propose a novel approach that leverages prompt learning as a pre-adjustment, followed by a post-adjustment using adapter methods. By aligning the representations learned from both approaches, we demonstrate that the knowledge acquired through adapter methods can be utilized to further bootstrap the effectiveness of prompt learning.

3. Methodology

3.1. Revisiting Vision-Language Model

We consider the pre-trained vision-language model CLIP that comprises a text encoder g and a vision encoder f with respective pre-trained parameters θ_g and θ_f. We denote $\theta_{CLIP} = \{\theta_g, \theta_f\}$ as the collection these parameters.

Vision Encoder: An input image $X \in \mathbb{R}^{C\times H\times W}$ is first divided into M patches that are projected into M patch tokens $t_1, \cdots, t_M$. The input $\hat{X} = \{t_{cls}, t_1, \cdots, t_M\}$ to the vision encoder f is then formed by appending a learnable class token t_{cls} to the M patch tokens. Latent visual feature representation $\hat{f} = f(\hat{X}, \theta_f) \in \mathbb{R}^d$ is extracted from $\hat{X}$ with multiple transformer blocks.

Text Encoder: The class label y corresponding to the image is wrapped within a text template (*e.g.*, 'a photo of a class label') to form $\hat{Y} = \{t_{SOS}, t'_1, \cdots, t'_L, c_k, t_{EOS}\}$, where $\{t'_l\}_{l=1}^{L}$ and c_k are word embeddings for the text template and class label of the k_{th} class, respectively, and t_{SOS} and t_{EOS} are learnable start and end token embeddings. The text encoder g encodes $\hat{Y}$ via multiple transformer blocks to obtain the latent textual feature $\hat{g} = g(\hat{Y}, \theta_g) \in \mathbb{R}^d$.

Zero-shot Classification for Vision-Language Model: For zero-shot classification, textual prompts are crafted with the text template and class labels $y \in \{1, \cdots, C\}$ for C classes. The prediction $\hat{y}$ given the image feature $\hat{f}$ is calculated by cosine similarity with a temperature parameter τ.

$$p(\hat{y}|\hat{f}) = \frac{\exp(sim(\hat{f}, \hat{g}_{\hat{y}})/\tau}{\sum_{C}^{i=1} \exp(sum(\hat{f}, \hat{g}_i))}. \tag{1}$$

Limitations of Different Tuning Methods: Prompt learning inserts trainable embeddings into the model's input space without modifying its internal parameters, which can lead to instability during training. Furthermore, since these embeddings merely guide the model's output, their effectiveness in downstream tasks is highly dependent on the pretrained model's inherent capabilities. Consequently, prompt learning performs poorly in scenarios where there is a significant distribution shift between the pretraining data and downstream tasks or when handling complex tasks. In contrast, adapter-based methods introduce trainable modules alongside the model's parameter matrices, enabling stronger representational capacity. They are more robust to distribution shifts and complex datasets. However, these methods often suffer from slow convergence; for example, TCP (Yao et al., 2024) requires 50 epochs to reach its reported performance, whereas VPT (Jia et al., 2022) and MaPLe (Khattak et al., 2023a) achieve slightly poor results in only 5 epochs. Therefore, an important research question is how to efficiently integrate adapter-based methods with prompt learning to leverage their respective advantages.

3.2. Proposed Method

Existing approaches like TCP (Yao et al., 2024) and DePT (Zhang et al., 2024), while pioneering the integration of adapters into prompt learning frameworks, exhibit two critical limitations: Their concurrent optimization of distinct parameter spaces (adapter modules vs. prompt embeddings) may induce conflicting optimization trajectories due to divergent gradient propagation patterns; They inadequately exploit the visual-semantic representational capacity inherent in CLIP's pretrained text encoder for classification tasks. In this section, PAT introduces a pre-post adjustment that sequentially aligns optimization directions through constrained prompt tuning, combined with tolerance regulatization to reinforce the model's discriminative and generalization capability, ultimately achieving more robust classification performance. Figure 1 depicts the overall framework architecture, we use prompt learning as a pre-adjustment to fine-tune the pre-trained VLM, followed by adapter tuning as a post-adjustment, mathematically, this process is expressed as

$$\hat{f}_\alpha = f(\hat{X}, \{\theta_f, \alpha_f\}) \ , \ \hat{g}_\alpha = g(\hat{Y}, \{\theta_g, \alpha_g\}) \tag{2}$$

$$\hat{f}_\beta = f(\{\beta_f, \hat{X}\}, \theta_f) \ , \ \hat{g}_\beta = g(\{\beta_g, \hat{Y}\}, \theta_g) \tag{3}$$

$$\hat{f} = \hat{f}_\alpha + \hat{f}_\beta \ , \ \hat{g} = \hat{g}_\alpha + \hat{g}_\beta \tag{4}$$

Table 1. In the Base-to-New experiment, performance comparison between classification experiments using New classes labels and All classes labels on the New classes dataset. It is observed that using All classes labels results in a significant drop in performance.

Datasets	Sets	CoOp	CoCoOp	MaPLe	PromptSRC
SUN397	New	68.3	76.9	78.7	79.0
	New(All label)	57.9	67.4	69.0	68.6
EuroSAT	New	53.0	60.0	73.2	68.4
	New(All label)	41.7	49.4	46.3	54.6
UCF101	New	67.4	73.5	78.7	78.3
	New(All label)	52.3	65.6	71.3	71.6

Where α_g and α_f denote the learnable parameters of adapter inserted alongside the model for the text branch and the visual branch, respectively. β_g and β_f represent the prompt learnable parameters inserted into the input embeddings for the text branch and the visual branch, respectively. $\hat{f}_\alpha$ and $\hat{g}_\alpha$ represent the fine-tuned representations obtained after applying prompt learning (pre-adjustment) of text branch and visual branch. Similarly, $\hat{f}_\beta$ and $\hat{g}_\beta$ indicate the fine-tuned representations obtained after applying adapter tuning (post-adjustment). the final fine-tuned representations $\hat{f}$ and $\hat{g}$ are obtained by integrating both approaches. And Figure 2 illustrates the tolerance regularization mechanism.

3.2.1. Pre-Post Alignment

Considering that x is the input image, z is the latent feature, α and β is the parameterized adapter and prompt respec-

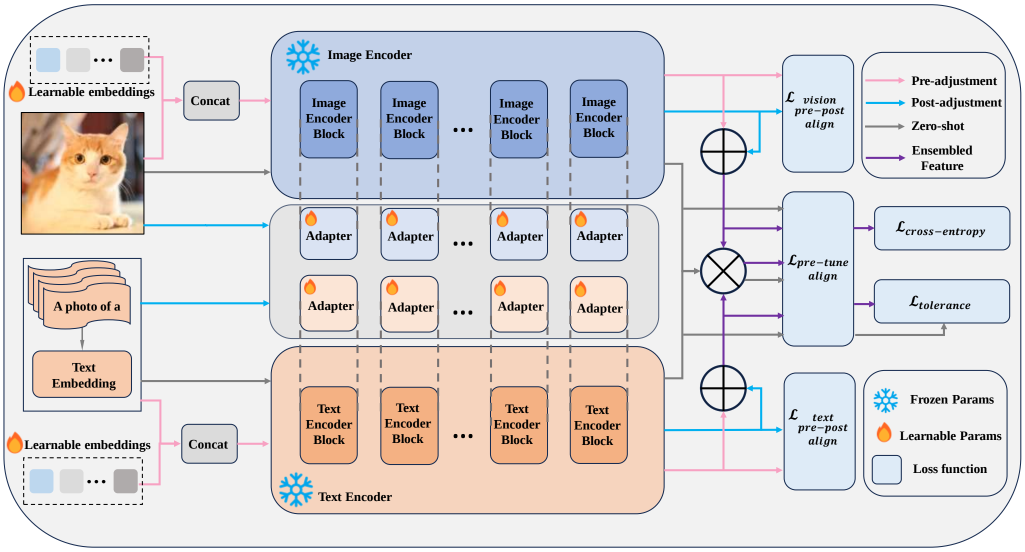

Figure 1. Overall Structure of PAT. In the figure, the blue blocks (excluding the loss function) represent the visual branch of the framework, while the orange blocks correspond to the textual branch of the model. The gray lines denote zeroshot inference, the pink lines illustrate the pre-adjustment leveraging prompt learning, and the blue lines depict the post-adjustment employing feature adaptation. Apart from the final cross-entropy loss function, the representations obtained from the pre- and post-adaptation processes of both modalities are constrained and aligned using MSE. Subsequently, the two representations are integrated through equal-weighted summation. The resulting logits and integrated representations are further aligned with zeroshot features using KL divergence and MAE, respectively. Additionally, Tolerance Regularization is computed between the final visual representation and the zeroshot textual representation.

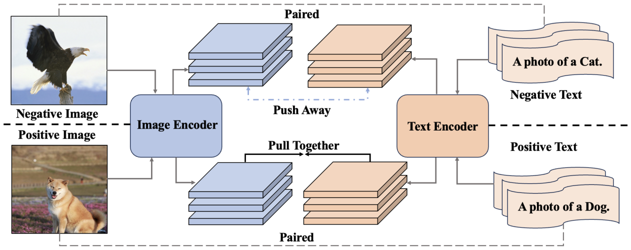

Figure 2. A schematic diagram of contrastive learning based on the tolerance regularization loss. For each image-text pair, if the pair is positive, the resulting visual representation and textual representation are pull together; otherwise, the two representations are pushed away apart.

tively. Then we have

$$p_\alpha(y|x) = \int p_\alpha(y|z)p_\alpha(z|x)\mathrm{d}z = \mathbb{E}_z[p_\alpha(y|z)], \tag{5}$$

$$p_\beta(y|x) = \int p_\beta(y|z)p_\beta(z|x)\mathrm{d}z = \mathbb{E}_z[p_\beta(y|z)]. \tag{6}$$

Assume $p_\alpha(z|x)$ and $p_\beta(z|x)$ obey Gaussian distribution

$$p_\alpha(z|x) = \mathcal{N}\left(z;\mu_\alpha(x),\sigma_\alpha^2 I\right), \tag{7}$$

$$p_\beta(z|x) = \mathcal{N}\left(z;\mu_\beta(x),\sigma_\beta^2 I\right). \tag{8}$$

$p_\alpha(y|z)$ and $p_\beta(y|z)$ are the determinant function i.e., the linear projection layer

$$p_\alpha(y|z) = \delta(y - W_\alpha z), \quad p_\beta(y|z) = \delta(y - W_\beta z) \tag{9}$$

where $\delta(\cdot)$ is the Dirac delta function and W_α, W_β is the weight matrix. Then the expectation can be simplified as

$$p_\alpha(y|x) = p_\alpha(z = \mu_\alpha(x)|x),\ p_\beta(y|x) = p_\beta(z = \mu_\beta(x)|x) \tag{10}$$

We aim to minimize the KL divergence between the prediction distribution as

$$\mathcal{D}_{KL}(p_\alpha(y|x)\|p_\beta(y|x)) = \mathbb{E}_{y\sim p_\alpha(y|x)}\left[\log\frac{p_\alpha(y|x)}{p_\beta(y|x)}\right]. \tag{11}$$

When Then we bring Eq. 10 into the above objective. Considering that $p_\alpha(z|x)$ and $p_\beta(z|x)$ obey Gaussian distribution, then the KL divergence has the analytical form:

$$\mathcal{D}_{KL}(p_\alpha\|p_\beta) = \frac{1}{2}\left[\log\frac{\sigma_\beta^2}{\sigma_\alpha^2} + \frac{\sigma_\alpha^2 + (\mu_\alpha - \mu_\beta)^2}{\sigma_\beta^2} - 1\right]. \tag{12}$$

To simplify the computation, we assume the adapter α and the prompt β have the same variance as

$$\sigma_\alpha^2 = \sigma_\beta^2 = \sigma^2 \tag{13}$$

Then the KL divergence turn into

$$\mathcal{D}_{KL}(p_\alpha\|p_\beta) = \frac{1}{2\sigma^2}\|\mu_\alpha - \mu_\beta\|^2 \tag{14}$$

Therefore, the pre-post alignment loss for each modal branch model is formalized using MSE.

$$\mathcal{L} = \mathbb{E}(\hat{f}_\alpha - \hat{f}_\beta)^2 + \mathbb{E}(\hat{g}_\alpha - \hat{g}_\beta)^2 \tag{15}$$

3.2.2. Tolerance Regularization

For each sample x_i, the corresponding visual feature is f_i, and for all the textual description, the textual embedding is $\{t_k\}_{k=1}^K$, where K is the number of categories. We then calculate the logits after softmax as

$$\hat{y}_i^{(k)} = \frac{e^{cf_i t_k + b}}{\sum_{j=1}^K e^{cf_i t_j + b}}. \tag{16}$$

where c is the constant and b is the bias. For one-hot label $y_i^{(k)} \in \{0, 1\}$, the cross-entropy loss is defined as

$$H(\hat{y}, y) = -\frac{1}{|B|} \sum_{i=1}^{|B|} \sum_{k=1}^{K} y_i^{(k)} \log \hat{y}_i^{(k)}. \tag{17}$$

In conventional vision language efficient tuning, the predicted textual description is forced to match one of the given label, making the model over-fitted on the training categories, thereby undermining the generalization to unseen classes. To avoid over-fitting on training datasets, we propose to use binary contrastive loss, *i.e.*, the sigmoid loss, as the regularization in the objective function.

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \log \frac{1}{1 + \exp(z_{ij}(-tx_i \cdot y_j + b))}, \tag{18}$$

where z_{ij} returns 1 when the i-th visual representation x_i matches the j-th textual representation y_j and -1 otherwise. Different from (Zhai et al., 2023), we fix the parameters b and t to -2 and 2, since the model possesses strong representation capability during fine-tuning.

Proposition 1. *The sigmoid loss function degenerates to the class-irrelevant binary cross entropy loss function, when considering only positive samples.*

Proof. Please refer to Appendix A. □

This proposition demonstrates that our proposed tolerance regularization yields an undifferentiated binary cross-entropy (BCE) loss. Furthermore, when incorporating images from unrelated categories without label information and randomly sampled mismatched text, this undifferentiated constraint effectively prevents the model from incorrectly assigning these samples to inappropriate categories. By avoiding the enforcement of erroneous category selection, the regularization enhances the model's robustness and generalization capability

Figure 2 shows that, during fine-tuning, the training data is divided into two categories: images within the current category space paired with their corresponding labels, and noise images outside the current category space paired with randomly sampled labels. The tolerance regularization processes each image-text pair independently.

For images within the category space, it brings their embeddings closer to the corresponding textual embeddings. For noise images, the similarity with the embeddings of all existing category texts will be pushed farther. During this process, we progressively penalize the distance between unknown samples and known text embeddings, thereby enhancing their generalization to unseen categories. Subsequent experimental results demonstrate that the application of tolerance regularization significantly improves the generalization capability of prompt learning.

3.2.3. Final Objective Function

Mathematically, based on the preceding content, the final objective function of PAT can be expressed as

$$\mathcal{L} = -0.01 \cdot \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \log \frac{1}{1 + \exp(z_{ij}(2\hat{g}_i \cdot \hat{f}_{j,zs} - 2))} + \mathbb{E}[(\hat{f}_\alpha - \hat{f}_\beta)^2] + \mathbb{E}[(\hat{g}_\alpha - \hat{g}_\beta)^2] + \mathcal{L}_{ce} + \mathcal{L}_{pre-tune} \tag{19}$$

Where $\hat{f}_{zs}$ represents the text representation obtained through zero-shot CLIP. In addition to the previously proposed pre-post align and pos-neg align, the objective function also includes a cross-entropy loss $\mathcal{L}_{ce}$ for guiding classification and an alignment loss $\mathcal{L}_{pre-tune}$ to constrain the pretrained and fine-tuned model, this type of alignment loss is widely adopted in other prompt learning approaches(Khattak et al., 2023b; Yao et al., 2023; 2024).

4. Experiments

4.1. Benchmark Settings

Datasets: Following (Khattak et al., 2023b) and (Yao et al., 2024), we conduct the Base-to-New generalization, few-shot learning, and cross-dataset generalization on 11 datasets and contains a wide range of recognition tasks. The datasets include ImageNet (Deng et al., 2009) and Caltech101 (Fei-Fei et al., 2004) for generic objects, OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and FGVC-Aircraft (Maji et al., 2013) for fine-grained classification, SUN397 (Xiao et al., 2010) for scene recogினition dataset, UCF101 (Soomro, 2012) for action recognition, DTD (Cimpoi et al., 2014) for texture classification, and the EuroSAT (Helber et al., 2019) dataset of satellite images.

Implementation Details: We utilize a pretrained CLIP with

Table 2. Performance comparison across different methods on Base-to-New Benchmark. PAT achieved state-of-the-art performance across Base, New, and H, with performance improvements of 1.5%, 0.7%, and 0.9%, respectively.

Datasets	Sets	CoOp	CoCoOp	ProGrad	ProDA	KgCoOp	RPO	PLOT	LFA	MaPLe	DePT	PromptSRC	TCP	PAT
		(ICCV22)	(CVPR22)	(ICCV23)	(CVPR22)	(ICCV23)	(ICCV23)	(ICLR23)	(ICCV23)	(CVPR23)	(CVPR24)	(ICCV23)	(CVPR24)	
Average	Base	82.4	80.5	82.5	81.6	80.7	81.1	84.0	83.6	82.3	83.6	84.1	84.1	**85.6**
	New	68.0	71.7	70.8	72.3	73.6	75.0	71.7	74.6	75.1	75.0	75.0	75.4	**76.1**
	H	74.5	75.8	76.2	76.7	77.6	77.8	77.4	78.8	78.5	79.1	79.3	79.5	**80.4**
ImageNet	Base	76.5	76.0	77.0	75.4	75.8	76.6	77.3	76.9	76.7	77.0	77.8	77.3	**78.0**
	New	66.3	70.4	66.7	70.2	70.0	**71.6**	69.9	69.4	70.5	70.1	70.7	69.9	70.5
	H	71.0	73.1	71.5	72.7	72.8	74.0	73.4	72.9	73.4	**74.1**	73.4	73.4	**74.1**
Caltech101	Base	97.8	98.0	98.0	98.3	97.7	98.0	98.5	98.4	97.7	98.3	98.1	98.2	**98.8**
	New	93.3	93.8	93.9	93.2	94.4	94.4	92.8	93.9	94.4	94.6	93.9	**94.7**	94.1
	H	95.5	95.8	95.9	95.7	96.0	96.0	95.6	96.1	96.0	**96.4**	96.0	96.0	**96.4**
OxfordPets	Base	94.5	95.2	95.1	95.4	94.7	94.6	94.5	95.1	95.4	94.3	95.5	94.7	**95.8**
	New	96.0	97.7	97.6	97.8	97.8	97.5	96.8	96.2	**97.8**	97.2	97.4	97.2	97.4
	H	95.2	96.4	96.3	**96.6**	96.2	96.1	95.7	95.7	**96.6**	95.8	96.4	95.9	**96.6**
Cars	Base	75.7	70.5	77.7	74.7	71.8	73.9	79.1	76.3	72.9	79.1	78.4	80.8	**81.5**
	New	67.5	73.6	68.6	71.2	75.0	**75.5**	74.8	74.9	74.0	**75.5**	74.7	74.1	73.5
	H	71.4	72.0	72.9	72.9	73.4	74.7	76.9	75.6	73.5	**77.3**	75.5	**77.3**	**77.3**
Flowers	Base	97.3	94.9	95.5	97.7	95.0	94.1	97.9	97.3	95.9	98.0	97.9	97.7	**98.2**
	New	67.1	71.8	71.9	68.7	74.7	76.7	73.5	75.4	72.5	76.4	76.8	75.6	**77.3**
	H	79.4	81.7	82.0	80.7	83.7	84.5	84.0	85.0	82.6	85.8	86.1	85.2	**86.5**
Food101	Base	89.4	**90.7**	90.4	90.3	90.5	90.3	89.8	90.5	**90.7**	90.5	90.6	90.6	90.5
	New	88.8	91.3	89.6	88.6	91.7	90.8	91.4	91.5	**92.1**	91.6	91.5	91.4	91.2
	H	89.1	91.0	90.0	89.4	91.1	90.6	90.6	91.0	**91.4**	91.1	91.1	91.0	90.8
Aircraft	Base	39.7	33.4	40.5	36.9	36.2	37.3	42.1	41.5	37.4	43.2	42.3	42.0	**46.2**
	New	31.2	23.7	27.6	34.1	33.6	34.2	33.7	32.3	35.6	34.8	37.0	34.4	**37.4**
	H	35.0	27.7	32.8	35.5	34.8	35.7	37.5	36.3	36.5	38.6	39.5	37.8	**41.3**
SUN397	Base	80.9	79.7	81.3	78.7	80.3	80.6	82.2	82.1	80.8	82.3	82.8	82.6	**82.9**
	New	68.3	76.9	74.2	76.9	76.5	77.8	73.6	77.2	78.7	77.8	**79.0**	78.2	78.8
	H	74.1	78.3	77.6	77.8	78.4	79.2	77.7	79.6	79.8	80.0	**80.9**	80.4	80.8
DTD	Base	80.0	77.0	77.4	80.7	77.6	76.7	82.0	81.3	80.4	82.2	82.6	82.8	**85.3**
	New	48.6	56.0	52.4	56.5	55.0	62.1	43.8	60.6	59.2	59.1	57.5	58.1	**63.5**
	H	60.5	64.9	62.5	66.4	64.4	68.6	57.1	69.5	68.2	68.8	67.8	68.3	**72.8**
EuroSAT	Base	90.1	87.5	90.1	83.9	85.6	86.6	93.7	93.4	94.1	89.0	92.4	91.6	**94.8**
	New	53.0	60.0	60.9	66.0	64.3	69.0	62.7	71.2	73.2	71.1	68.4	**74.7**	74.4
	H	66.7	71.2	72.7	73.9	73.5	76.8	75.1	80.8	82.3	79.0	78.6	82.3	**83.4**
UCF101	Base	84.5	82.3	84.3	85.2	82.9	83.7	86.6	87.0	83.0	85.8	86.9	87.1	**89.2**
	New	67.4	73.5	74.9	72.0	76.7	75.4	75.9	77.5	78.7	77.2	78.3	**80.8**	79.3
	H	75.0	77.7	79.4	78.0	79.7	79.3	80.9	82.0	80.8	81.3	82.4	83.8	**84.0**

backbone of ViT-B/16. We repeat all the experiments for 3 times and report the average results. The prompts are randomly initialized and trained for 20 epochs under all the settings. The length of prompts is set to 4. We adopt AdapterFormer (Chen et al., 2022) for post-adjustment of the text encoder and Convpass (Jie et al., 2024) for the vision encoder. For both modalities, the adapters are applied in the multi-head attention layer and the linear layer with a scaling factor of 0.1 and a hidden dimension of 16. For cross-dataset evaluation, we train the source model on all classes of ImageNet with 16 shots settings using SGD optimizer with the learning rate of 3.5e-3 and the batch size of 4. For feature ensembling, we add both feature from pre-adjustment and post-adjustment with equal weight. All experiments are conducted on RTX 2080Ti except for ImageNet on RTX 4090 and NVIDIA A100.

Baselines: We adopt most recent state-of-the-art methods without using the large language model as baselines, including CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), ProGrad (Zhu et al., 2023), ProDA (Lu et al., 2022), KgCoOp (Yao et al., 2023), PromptSRC (Khattak et al., 2023b), MaPLe (Khattak et al., 2023a), LFA (Ouali et al., 2023), DePT (Zhang et al., 2024), PLOT (Chen et al., 2023), TaskRes (Yu et al., 2023), RPO (Lee et al., 2023), VPT (Jia et al., 2022), TIP-Adapter-F (Zhang et al., 2022), and TCP (Yao et al., 2024).

4.2. Base-to-New Generalization

To evaluate the generalization ability of PAT, we equally split each dataset into base and new classes. The model is trained using the base classes in a 16-shot setting and evaluated on new classes. To simultaneously evaluate the fitting ability, generalization capability, and overall performance, we report the classification accuracy for both base classes and new classes, as well as their harmonic mean.

Table 2 shows that PAT achieves state-of-the-art perfor-

Table 3. Accuracy (%) for few-shot classification. PAT achieved state-of-the-art performance, delivering an absolute performance improvement of 1.5% compared to TCP.

Datasets	CLIP	CoOp	CoCoOp	ProGrad	KgCoOp	MaPLe	TIP-Adapter-F	DAPT	PromptSRC	PLOT	TaskRes	TCP	PAT
ImageNet	66.7	69.4	70.6	70.2	70.2	70.7	**70.8**	**70.8**	**70.8**	70.4	62.9	70.5	**70.8**
Caltech101	93.3	94.4	95.0	94.9	94.7	94.3	94.8	94.2	94.8	95.1	94.7	95.0	**95.5**
OxfordPets	89.1	91.3	93.0	93.2	93.2	92.1	92.3	92.2	93.2	92.6	92.0	91.9	**93.5**
StanfordCars	65.7	72.7	69.1	71.8	72.0	68.7	74.4	74.4	71.8	74.9	75.9	**76.3**	75.7
Flowers	70.7	91.1	82.6	90.0	90.7	80.8	93.0	92.4	91.3	92.9	91.5	**94.4**	93.7
Food101	85.9	82.6	86.6	85.8	86.6	**86.9**	86.2	83.6	86.1	86.5	86.0	85.3	86.3
Aircraft	24.9	33.2	30.9	32.9	32.5	29.0	35.5	32.5	32.8	35.3	33.8	36.2	**38.0**
SUN397	62.6	70.1	70.5	71.2	71.8	71.5	70.7	72.2	72.8	70.4	72.7	72.1	**74.0**
DTD	44.3	58.6	54.8	57.7	58.3	54.7	61.7	61.4	60.6	62.4	59.6	64.0	**65.4**
EuroSAT	48.3	68.6	63.8	70.8	71.1	54.9	78.3	72.7	75.0	80.7	72.9	77.4	**85.3**
UCF101	67.6	77.4	75.0	77.8	78.4	73.7	79.7	79.4	79.4	79.8	76.1	80.8	**81.7**
Average	65.4	73.6	72.0	74.2	74.5	70.7	76.1	75.1	75.3	76.5	74.4	76.7	**78.2**

Table 4. Accuracy (%) for cross-dataset generalization. PAT achieved state-of-the-art performance in all settings.

Parameter-Efficient Fine-Tuning On DTD Base Classes											
Methods	Caltech101	OxfordPets	Cars	Flowers	Food101	Aircraft	SUN397	EuroSAT	UCF101	ImageNet	Average
CLIP	93.3	89.1	65.6	70.7	85.9	24.7	62.6	48.3	67.6	72.4	68.0
TCP	91.6	86.8	64.7	68.5	85.2	20.5	62.1	46.4	68.2	65.6	66.0
PAT	96.7	89.4	60.3	66.3	87.9	22.4	72.0	57.5	68.6	71.6	**69.3**

Parameter-Efficient Fine-Tuning On EuroSAT Base Classes											
Methods	Caltech101	OxfordPets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	UCF101	ImageNet	Average
CLIP	93.3	89.1	65.6	70.7	85.9	24.7	62.6	44.1	67.6	72.4	67.6
TCP	86.4	82.8	61.4	65.1	83.3	16.5	51.6	34.8	63.3	58.8	60.4
PAT	96.8	87.0	60.4	63.3	88.9	21.4	70.7	53.9	68.9	69.2	**68.1**

Parameter-Efficient Fine-Tuning On ImageNet all Classes											
Methods	Caltech101	OxfordPets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP	93.3	89.1	65.6	70.7	85.9	24.7	62.6	44.1	48.3	67.6	65.2
CoOp	93.7	89.1	64.5	68.7	85.3	18.5	64.2	41.9	46.4	66.6	63.9
ProGrad	91.5	89.6	62.4	67.9	85.4	20.2	62.5	39.4	43.5	64.3	62.7
KgCoOp	93.9	89.8	65.4	70.0	86.4	22.5	66.2	46.4	46.0	68.5	65.5
DePT	94.2	90.0	65.6	70.6	86.4	23.3	66.7	46.0	43.5	69.3	65.6
VPT	93.7	89.3	65.5	70.2	86.3	22.1	66.6	46.9	47.4	67.2	65.5
PLOT	92.1	90.1	65.7	69.2	86.2	**25.0**	61.7	38.6	47.8	67.0	64.3
PromptSRC	93.6	90.3	65.7	70.3	86.2	23.9	67.1	**46.9**	45.5	68.8	65.8
MaPLe	93.5	90.5	65.6	**72.2**	86.2	24.7	67.0	46.5	48.1	68.7	66.3
DAPT	93.5	90.7	**65.9**	71.7	86.1	23.0	67.0	44.0	**52.5**	68.7	66.3
TCP	**94.0**	**91.3**	64.7	71.2	**86.7**	23.5	67.2	44.4	51.5	68.7	66.3
PAT	93.4	90.2	65.8	71.3	86.0	24.5	**67.6**	46.1	50.8	**68.9**	**66.5**

mance on 9 out of 11 datasets and is competitive on the remaining SUN397 and Flowers datasets. Compared with CoOp (Zhou et al., 2022b), PAT achieves an accuracy gain of 5.9% on average and 3.2% and 8.1% on the base and new classes, respectively. Furthermore, PAT outperforms the state-of-the-art TCP by 0.9% on average (80.4% vs. 79.5%), 1.5% on the base classes (85.6% vs. 84.1%), and 0.7% on the new classes (76.1% vs. 75.4%). These results demonstrate that PAT achieves better fitting capability and generalization ability compared to existing methods.

4.3. Few-Shot Classification

To better validate the ability of our proposed PAT to perform transfer learning with limited data, we conducted few-shot classification experiments on 11 datasets. All methods were trained using K-shot training images and corresponding class labels, and evaluated on test sets that share the same class space as the training sets. Following previous approaches, we present classification performance on 4-shots. Table 3 shows that PAT achieves the best performance in 8 out of 11 datasets. For example, in DTD, we improved performance from 64% to 65.4%; in EuroSAT, from 77.4% to 85.3%; and in SUN397, from 72.8% to 74.0%. Overall, PAT shows a 1.5% improvement compared to the previous state-of-the-art, providing strong evidence of its capability for downstream transfer learning with limited samples.

4.4. Cross-Dataset Generalization

In base-to-new generalization, the base and new classes are sampled from the same datasets, and thereby are similar in data distribution. To further evaluate the generalization of PAT, we conduct a cross-dataset generalization experiments. Unlike previous studies, we aim to verify whether models maintain strong generalization capabilities after downstream transfer under truly small-scale data with limited samples. Thus in this experiment, all methods are trained on the base classes of the DTD and EuroSAT datasets and all classes of ImageNet in 16-shots settings under three distinct random

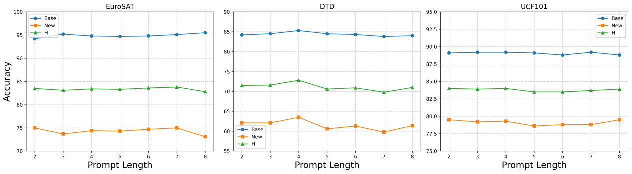

Figure 3. Ablation study of prompt length.

seeds, and subsequently evaluated on all categories of the other datasets. We compare the proposed PAT with Zero-shot CLIP and TCP (demonstrating robust performance in base-to-new scenarios).

Table 4 shows that, when trained on these cross-distribution few-shot datasets, TCP is inferior to Zero-shot CLIP in most cases. Notably, after training on the satellite imagery dataset EuroSAT, TCP exhibits a 7.2% performance gap relative to Zero-shot CLIP. This discrepancy persists at 2.0% when trained on DTD. In contrast, PAT outperforms TCP by 7.7% and 3.3% on these two datasets respectively while simultaneously surpassing Zero-shot CLIP, further demonstrating PAT's strong generalization capabilities.

4.5. Ablation Studies

Ablation studies are performed on on base-to-new generalization on the EuroSAT, DTD, and UCF101 datasets to validate the loss function, adapter configuration, and prompt length. We evaluate on each dataset using three random seeds, and report average accuracy for base classes, new classes, and their harmonic mean.

Loss Function: We first validate the effectiveness of the proposed loss function, including the alignment loss to constrain the pre-trained and fine-tuned models, alignment loss to constrain pre-adjustment and post-adjustment, and the tolerance regularization for constructing a robuster classifier. Table 5 shows that the absence of the tolerance regularization results in a 1.4% decline in overall performance on EuroSAT, with accuracy decreasing by 1.3% on the Base category and 1.4% on the New category. On DTD, the overall performance drops by 0.4%, with a 1.3% decrease in the Base category. Similarly, when the forward-backward calibration loss is removed, the accuracy on EuroSAT decreases by 2.7% in the New category and 1.6% overall. On DTD, the Base, New, and overall performance decrease by 0.3%, 2.3%, and 1.6%, respectively.

Adapter Configuration: Since PAT relies on adapters to constrain the update direction of prompt learning, we perform ablation experiments on the scaling factor α and hidden dimensions r of the adapter. Table 6 shows that the configuration with $\alpha = 0.1$ and $r = 16$ achieves the best comprehensive performance overall. However, other hyper-parameter combinations can outperform this configuration on specific datasets. For instance, $\alpha = 0.1$ and $r = 8$ perform better on UCF101, while $\alpha = 0.01$ and $r = 16$ achieve superior results on EuroSAT. Note that, compared to the hidden dimension r, the scaling factor α has a more significant impact on performance across all three datasets.

Table 5. Ablation study of Loss function. B, PP, and Tol are the abbreviations of Baseline, Pre-Post, and Tolerance respectively.

B	PP	Tol	EuroSAT			DTD			UCF101		
			Base	New	H	Base	New	H	Base	New	H
✓	✓		93.5	73.0	82.0	84.0	63.6	72.4	89.0	79.2	83.8
✓		✓	95.2	71.7	81.8	85.0	61.2	71.2	89.0	80.5	84.5
✓	✓	✓	94.8	74.4	83.4	85.3	63.5	72.8	89.2	79.3	84.0

Table 6. Ablation study of hyper-params in adapter config.

α	r	EuroSAT			DTD			UCF101		
		Base	New	H	Base	New	H	Base	New	H
0.1	2	94.1	66.5	77.8	83.9	63.6	72.4	88.3	79.2	83.5
0.1	4	93.7	72.0	81.4	84.1	63.4	72.3	88.2	78.4	83.0
0.1	8	94.1	74.4	83.1	84.3	62.0	71.5	89.5	79.6	84.2
0.1	16	94.8	74.4	83.4	85.3	63.5	72.8	89.2	79.3	84.0
10.0	16	96.0	68.0	79.6	84.8	58.9	69.5	87.0	74.7	80.4
1.0	16	96.2	65.2	77.7	83.9	57.9	68.5	87.4	78.1	82.5
0.1	16	94.8	74.4	83.4	85.3	63.5	72.8	89.2	79.3	84.0
0.01	16	92.8	75.9	83.5	83.7	60.3	70.0	86.4	78.2	82.1

Prompt Length: We investigate the impact of prompt length under the Base-to-New configuration. We compare the performance effects of prompt lengths ranging from 2 to 8. Figure 3 shows that the length of the prompt has a certain impact on performance. In previous experiments, we fixed the prompt length to 4. However, when the prompt length is set to 2, 6, or 7, PAT's performance on EuroSAT can be further improved, and a length of 2 achieves better performance on UCF101. Nevertheless, as indicated by the trend lines in the figure, PAT is generally insensitive to the choice of prompt length.

5. Conclusion

In this paper, we propose a novel prompt learning approach based on pre-adjustment, post-adjustment, and contrastive learning. To further enhance the fitting ability and generalization of current prompt learning methods, we employ adapter-based feature adaptation as a post-adjustment to refine the optimization direction of prompt learning, allowing it to acquire knowledge in the parameter space. Furthermore, we utilize a tolerance regularization to bring known samples closer to text representations while penalizing noise samples against existing text representations, resulting in a more robust multimodal classifier.Our extensive experimental results, including Base-to-New, Cross-dataset, and few-shot evaluations, demonstrate that our proposed method, PAT, achieves significant advancements in both fitting performance and generalizability compared to previous SOTA.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Bossard, L., Guillaumin, M., and Van Gool, L. 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*, pp. 446–461. Springer, 2014.

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., and Zhang, K. PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

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

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Han, Z., Gao, C., Liu, J., Zhang, J., and Zhang, S. Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

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

Jie, S., Deng, Z.-H., Chen, S., and Jin, Z. Convolutional bypasses are better vision transformer adapters. In *ECAI 2024*, pp. 202–209. IOS Press, 2024.

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

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M.-H., and Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

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

Lee, D., Song, S., Suh, J., Choi, J., Lee, S., and Kim, H. J. Read-only prompt optimization for vision-language few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1401–1411, 2023.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Lu, Y., Liu, J., Zhang, Y., Liu, Y., and Tian, X. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5215, 2022.

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

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Ouali, Y., Bulat, A., Matinez, B., and Tzimiropoulos, G. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15534–15546, 2023.

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Soomro, K. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Yao, H., Zhang, R., and Xu, C. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6757–6767, 2023.

Yao, H., Zhang, R., and Xu, C. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23438–23448, 2024.

Yu, T., Lu, Z., Jin, X., Chen, Z., and Wang, X. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Zhang, J., Wu, S., Gao, L., Shen, H. T., and Song, J. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12924–12933, 2024.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023.

A. Proof of Proposition 1

When we only consider the positive samples Then we have the sigmoid loss function as

$$\mathcal{L} = -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\kappa(cf_it_i - b). \tag{20}$$

where κ is the sigmoid function as $\kappa(u) = \frac{1}{1+e^{-u}}$ Besids, the cross entropy can be simplified as

$$H(\hat{y}, y) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\frac{e^{cf_it_{k^*}+b}}{\sum_{j=1}^{K} e^{cf_it_j+b}}. \tag{21}$$

Considering that only on positive textual label for each image sample. Besides, Similarity scores for all negative samples are constant as $cf_it_j + b = 0(j \neq k^*)$. Then we have

$$\sum_{j=1}^{K} e^{cf_it_j+b} = e^{cf_it_{k^*}+b} + (K-1)e^0 = e^{cf_it_{k^*}+b} + (K-1). \tag{22}$$

Then the cross entropy degrads into:

$$H(\hat{y}, y) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\frac{e^{cf_it_{k^*}+b}}{e^{cf_it_{k^*}+b} + (K-1)}. \tag{23}$$

When we further consider the binary classification, i.e., whether the image feature is aligned with the textual description, we have

$$H(\hat{y}, y) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\frac{e^{cf_it_{k^*}+b}}{e^{cf_it_{k^*}+b} + 1} = -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\kappa(cf_it_{k^*} + b). \tag{24}$$

B. Domain Generalization

Domain generalization experiments involve training a model on the source domain and testing it on the target domain, making them useful for evaluating model generalization. Therefore, we train PAT under the ImageNet 16-shot setting and test it on ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. The final results are reported as the average performance across these five datasets. As shown in Table 7, PAT still holds state-of-the-art performance.

Datasets	ImageNet	-V2	-S	-A	-R	Avg.
CoCoOp	71.0	64.1	48.8	50.6	76.2	62.1
ProGrad	72.2	64.7	47.6	49.4	74.6	61.7
KgCoOp	71.2	64.1	49.0	50.7	76.7	62.3
MaPLe	70.7	64.1	49.2	50.9	77.0	62.4
DAPT	71.7	64.5	49.5	51.1	76.3	62.6
TCP	71.2	64.6	49.5	**51.2**	76.7	62.6
PromptSRC	71.3	64.4	**49.6**	50.9	**77.8**	62.8
PAT	**72.8**	**66.5**	49.4	49.0	77.1	**63.0**

Table 7. Performance comparison across different methods on Domain Generalization Experiment. PAT achieved state-of-the-art performance, delivering an absolute performance improvement of 0.2% compared to PromptSRC.