DA-VPT: Semantic-Guided Visual Prompt Tuning for Vision Transformers

Supplementary Material

Contents

A Details About the Experiments 1

B Discussion and Comparison with Self-SPT 2

C DA-VPT+: Integration with Bias Tuning 2

D Supplemental Empirical Studies 3

E The Proof and Detial of theorem 1 4

F. Limitations 5

G Future Works 5

H Broader Impact 5

I. More Examples of Attention Maps on Prompts 5

A. Details About the Experiments

A.1. Datasets

Classification Datasets. **FGVC** encompasses five fine-grained visual classification datasets: CUB-200-2011 [27], NABirds [25], Oxford Flowers [23], Stanford Dogs [13], and Stanford Cars [7]. Following Jia et al. [11], we split each dataset into `train` (90%) and `val` (10%) subsets.

VTAB-1K contains 19 diverse visual tasks across three categories: (i) *Natural* tasks involving standard camera images for object classification and scene recognition; (ii) *Specialized* tasks using domain-specific imagery such as medical scans and satellite data; and (iii) *Structured* tasks focusing on spatial relationships and object properties.

Segmentation Datasets. We evaluate on two semantic segmentation benchmarks: **ADE20K** with 150 fine-grained semantic concepts, and **PASCAL Context** providing pixel-wise annotations across 60 object classes. For dataset partitioning, we strictly follow the protocol established in VPT [11]. Complete dataset statistics and task details are provided in Table 2.

A.2. Implementation Details

Classification Tasks. For FGVC datasets, we employ standard data augmentation: random resizing and cropping to 224 × 224 pixels with random horizontal flipping. For VTAB-1K, following Zhai et al. [32] and Jia et al. [11], images are directly resized to 224 × 224 pixels without additional augmentation.

Model training utilizes the AdamW optimizer with a batch size of 32 over 100 epochs. The learning rate follows a combined schedule: a 10-epoch linear warm-up followed by cosine decay [17] from the initial value to 1e-8. We determine optimal hyperparameters through cross-validation on the validation set. Following established protocols [6, 11, 16], we report mean accuracy across three runs with different random seeds.

Segmentation Tasks. We implement our experiments using the SETR framework [33] through MMSegmentation. We adopt the SETR-PUP configuration, utilizing one primary head and three auxiliary heads to process features from transformer layers 9, 12, 18, and 24. Training follows Zheng et al. [33]: 160k iterations for ADE20K and 80k iterations for PASCAL Context, with hyperparameter optimization mirroring our classification approach.

For multi-class segmentation samples, we adapt the class assignment strategy by randomly selecting one non-background class as the target class for visual prompt assignment during each iteration, accounting for the pixel-wise multi-class nature of segmentation tasks.

A.3. Hyperparameter Configuration

Parameter Search Space. Table 1 details our hyperparameter search space for each task, including learning rate, weight decay, and the number and location of layers guided by semantic metrics loss. For the main results, we maintain default values for Proxy-Anchor loss parameters to narrow the parameter searching space.

Configuration	Value
Optimizer	AdamW [18]
Base learning rate range	{1e-3, 5e-4, 1e-4, 5e-5}
Weight decay range	{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0}
Learning rate schedule	Cosine Decay [17]
Layers applied guidance	{ 12, 10, 8, 6, 4, 2, 0}
Num of prompts applied guidance	{ 5, 10, 20, 40}
Proxy-Anchor δ	32.0
Proxy-Anchor τ	10.0
Batch size	32
Warmup epoch	10
Total epoch	100 (ViT-B/16)
Augmentation	RandomResizedCrop, RandomHorizontalFlip

Table 1. Hyper Parameters Searching Space and Training configuration in our experiments

Prompt Configuration for Tasks with Limited Classes. For tasks with limited classes ($C < 5$), we augment the visual prompts by adding extra unassigned prompts that are

not guided by semantic metrics loss. This approach empirically improves prompt transferability, particularly in tasks with very few classes (e.g., Patch Camelyon, Retinopathy, or KITTI-Dist in VTAB). The extra prompts help maintain an effective number of visual prompts in the guiding layer while preserving the semantic structure of the original class-assigned prompts.

B. Discussion and Comparison with Self-SPT

Methodological Distinctions. While both Self-SPT and our work leverage prompt distributions to enhance representation learning, they differ fundamentally in their approaches. Self-SPT attempts to align prompt and visual token distributions through initialization, using mean or max pooling of input data to set background values. In contrast, our method achieves distribution matching at the semantic level throughout the optimization process. This semantic-level matching enables our visual prompts to capture discriminative features by explicitly considering class relationships during training.

Our Key Advantages. Our approach demonstrates several significant advantages over Self-SPT:

- **Continuous Optimization:** Our method maintains distribution regularization throughout the entire optimization process, while Self-SPT only applies distribution alignment during initialization.
- **Discriminative Feature Learning:** Through metric guidance, our prompts explicitly capture class-specific discriminative features by comparing tokens from the same and different classes. In contrast, Self-SPT's uniform background value initialization does not differentiate between class-specific features.
- **Computational Efficiency:** Our method significantly reduces pre-processing overhead by clustering class representations rather than entire visual token sets, avoiding the computational burden of Self-SPT's k-means clustering approach on the full dataset.

Empirical Analysis. To thoroughly evaluate the relationship between background value initialization and metric guidance learning, we attempted to reproduce Self-SPT's results and assess its performance when integrated with our method. As illustrated in Figure 1, our experiments revealed two key findings: (1) we were unable to reproduce the performance metrics reported in the original Self-SPT paper, and (2) the background value initialization strategy did not yield measurable improvements when combined with our metric guidance approach. These results further validate our focus on semantic-level distribution matching as the primary mechanism for improving prompt optimization.

The comparative analysis demonstrates that while both methods address prompt distribution optimization, our approach offers *more robust theoretical foundations*, *better computational efficiency*, and *superior empirical performance*. The continuous nature of our distribution regularization, combined with explicit semantic guidance, provides a more principled framework for learning discriminative prompt representations.

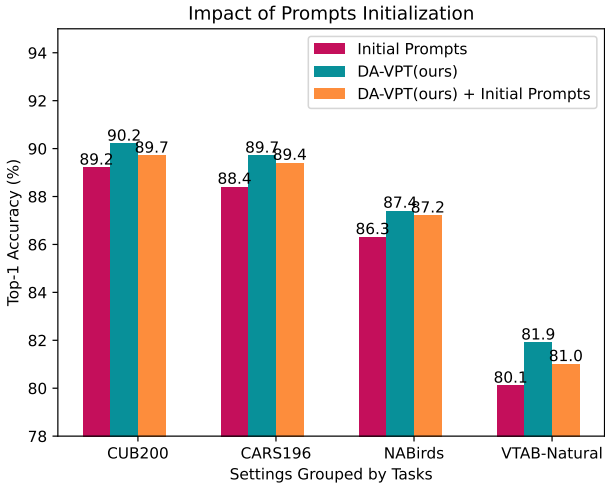


Figure 1. **Impact of Prompt Initialization Strategies.** Comparative analysis of model performance under different prompt initialization schemes. Results demonstrate that the background value initialization method proposed in Self-SPT [28], which uses mean pooled visual tokens, shows no significant performance gains when combined with our distribution-aware guidance approach. This suggests that our method's effectiveness stems from its continuous distribution matching during training rather than initial prompt configurations.

C. DA-VPT+: Integration with Bias Tuning

This section examines the synergy between our proposed metric learning guidance and bias tuning in PEFT models. Our investigation is motivated by an intriguing observation from the original VPT work [11], which reported that bias tuning adversely affects vanilla VPT optimization. We present a novel perspective on this interaction and demonstrate how our approach effectively addresses these limitations.

Theoretical Motivation. The distribution of visual tokens in transformer layers is inherently constrained by the representations from previous layers and their associated visual prompts. We hypothesize that unfreezing bias terms, particularly in Key and Value projections, introduces additional flexibility in token representation. This flexibility becomes especially significant when combined with our metric learning guidance, as it allows for more nuanced distribution alignment between visual prompts and tokens.

Datasets	Task Description	Classes	Train Size	Val Size	Test Size
Fine-Grained Visual Classification (FGVC) [11]					
CUB-200-2011 [27]	Fine-grained Bird Species Recognition	200	5,394	600	5,794
NABirds [25]	Fine-grained Bird Species Recognition	55	21,536	2,393	24,633
Oxford Flowers [23]	Fine-Grained Flower Species recognition	102	1,020	1,020	6,149
Stanford Dogs [13]	Fine-grained Dog Species Recognition	120	10,800	1,200	8,580
Stanford Cars [7]	Fine-grained Car Classification	196	7,329	815	8,041
Visual Task Adaptation Benchmark (VTAB-1k) [32]					
Caltech101 [5]	Natural-Tasks (7) Natural images captured using standard cameras.	102	800/1000	200	6,084
CIFAR-100 [14]		100			10,000
DTD [3]		47			1,880
Oxford-Flowers102 [22]		102			6,149
Oxford-PetS [24]		37			3,669
SVHN [21]		10			26,032
Sun397 [29]		397			21,750
Patch Camelyon [26]	Special-Tasks (4) Images captured via specialized equipments	2	800/1000	200	32,768
EuroSAT [9]		10			5,400
Resisc45 [2]		45			1,880
Retinopathy [4]		5			42,670
Clevr/count [12]	Structured-Tasks (8) Require geometric comprehension	8	800/1000	200	15,000
Clevr/distance [12]		6			15,000
DMLab [1]		6			22,735
KITTI-Dist [8]		4			711
dSprites/location [19]		16			73,728
dSprites/orientation [19]		16			73,728
SmallNORB/azimuth [15]		18			12,150
SmallNORB/elevation [15]		18			12,150
Image Semantic Segmentation					
ADE20K [34]	Fine-grained images with pixel-wise semantic annotations	150	20210	2000	3352
PASCAL Context [20]		60	4998	5105	—-

Table 2. **The details and specifications of the downstream task datasets we selected to evaluate our proposed framework.**

Advantages over Vanilla VPT. Unlike vanilla VPT, DA-VPT explicitly manages distribution alignment between visual prompts and tokens through metric learning guidance. This explicit alignment makes our method more responsive to distribution shifts in visual tokens, which are substantially influenced by projection layer bias terms. By simultaneously optimizing bias terms and maintaining distribution alignment, DA-VPT+ achieves more robust and effective feature representations. This integration of bias tuning with DA-VPT demonstrates how our method's distribution-aware approach can transform a previously problematic component (bias tuning) into a complementary enhancement.

D. Supplemental Empirical Studies

We conduct additional empirical investigations to explore potential extensions of our proposed framework in two key directions: alternative metric learning approaches and modified architectural connections.

Alternative Metric Learning Losses. We investigate the effectiveness of different metric learning losses for guiding visual prompt distributions. Specifically, we compare our proposed Proxy-Anchor (PA) loss against two alternatives: vanilla Proxy-NCA loss and triplet loss. For these vanilla losses, we treat the selected prompts as individual data points, with class assignments determined by the current training epoch. This comparison helps us understand the relative advantages of our PA loss formulation in the context of prompt optimization. Details are listed in Table 3.

Modified Connection Structures. We examine two architectural modifications to the base framework: 1) Cross-layer prompt connections (DA-VPT+Conn), which enable information flow between prompts at different layers, and 2) Learnable gated connections following GateVPT [30] (DA-VPT+Gate), which introduce adaptive control over prompt interactions.

These architectural studies provide insights into the role of prompt connectivity in our distribution-aware framework. The experimental results and detailed analysis of these varia-

Methods	CUB200	Cars	NABirds	VTAB-Natural
VPT (baseline)	88.6	87.4	85.7	78.48
DA-VPT	90.2 (+1.6)	89.7 (+2.3)	87.4 (+1.7)	80.25 (+1.77)
DA-VPT+Conn	89.6 (+1.0)	88.1 (+0.7)	86.8 (+1.1)	79.11 (+0.63)
DA-VPT+Gate	89.8 (+1.2)	88.4 (+1.0)	87.0 (+1.3)	79.48 (+1.00)
DA-VPT (PNCA)	89.2 (+0.6)	88.2 (+0.8)	86.9 (+1.2)	79.22 (+0.74)
DA-VPT (triplet)	87.9 (-0.7)	87.1 (-0.3)	85.4 (-0.3)	78.61 (+0.13)

Table 3. **Analysis of Connection Structures and Metric Learning Variants.** Empirical evaluation of (a) different visual prompt connection architectures across transformer layers and (b) alternative metric learning approaches. Our results indicate that neither cross-layer connections nor learnable activation gates provide substantial improvements over our base method. Furthermore, experiments with alternative metric learning losses show less stable fine-tuning performance compared to our approach, with some variants performing below the VPT baseline. These findings suggest that the effectiveness of our method primarily stems from its distribution-aware prompt optimization rather than architectural modifications or alternative metric formulations.

tions are presented in Table 3.

Algorithm 1 Distribution Aware Visual Prompt Tuning (DA-VPT)

Input: Pre-trained ViT model f_θ, Dataset $\mathcal{D} = (x_i, y_i)_{i=1}^N$,
number of prompts M, β, λ, learning rate and other related hyperparameters
Output: Fine-tuned ViT model f_θ
Initialize M prompts $\mathbf{p}^l$ for each layer l
Get class tokens $\mathbf{S} \in \mathbb{R}^{C \times D}$ by Forward passing f_θ
Create a mapping from C classes to M prompts ($C \to M$) using k-means clustering on $\mathbf{S}$
while *stop criteria is not satisfied* **do**
  Obtain a batch $\{x_i, y_i\}_{i=1}^n$ from $\mathcal{D}$
  Forward pass $\mathbf{x}_i$ through ViT f_θ with prompts $\mathbf{p}^l$
  Select saliency patch $\mathbf{x}$ right after attention layer in last selected blocks
  Calculate metric learning losses $\mathcal{L}_{\text{ML}}(\mathbf{x}, \mathbf{p})$ and $\mathcal{L}_{\text{ML}}(\mathbf{p}, \mathbf{x}_{\text{cls}})$
  Calculate cross-entropy loss $\mathcal{L}_{\text{CE}}$
  Minimize loss: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{ML}}(\mathbf{x}, \mathbf{p}) + \lambda\mathcal{L}_{\text{ML}}(\mathbf{p}, \mathbf{x}_{\text{cls}})$
  Update $\mathbf{p}$ and other learnable parameters from Backward of $\mathcal{L}$
  Update class tokens $\mathbf{S}$ and class-prompt mapping $C \to M$ after certain steps
return Fine-tuned ViT model f_θ

E. The Proof and Detial of theorem 1

Theorem 1. *For a weight perturbation Δa_i calculated using the softmax function, there is an approximate relationship:*

$$\Delta a_i \approx a_i(1 - a_i)\Delta s_i,$$

where Δs_i is a small change in the attention score s_i, and

$$\Delta s_i = \frac{\Delta p^\top v_i}{\sqrt{d}}.$$

Proof. The attention weights a_i are calculated using the softmax function applied to the attention scores s_i:

$$a_i = \frac{e^{s_i}}{\sum_j e^{s_j}}.$$

The partial derivative of a_i with respect to s_j is given by:

$$\frac{\partial a_i}{\partial s_j} = a_i(\delta_{ij} - a_j),$$

where δ_{ij} is the Kronecker delta function:

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

For small perturbations Δs_j, we can approximate the change in a_i using a first-order Taylor expansion:

$$\Delta a_i \approx \sum_j \frac{\partial a_i}{\partial s_j}\Delta s_j.$$

Then we substitute the expression for the derivative:

$$\Delta a_i \approx \sum_j a_i(\delta_{ij} - a_j)\Delta s_j.$$

Split the summation into two parts:

$$\Delta a_i \approx a_i(1 - a_i)\Delta s_i - a_i \sum_{j \neq i} a_j \Delta s_j.$$

We assume that the weighted sum of the perturbations Δs_j for $j \neq i$ is negligible:

$$\sum_{j \neq i} a_j \Delta s_j \approx 0.$$

This approximation is reasonable when:

- The perturbations Δs_j for $j \neq i$ are small and uncorrelated, so they average out.
- The attention weights a_j for $j \neq i$ are small (i.e., a_i is dominant).

Under this assumption, the expression simplifies to:

$$\Delta a_i \approx a_i(1 - a_i)\Delta s_i.$$

□

F. Limitations

Our Parameter-Efficient Fine-Tuning (PEFT) approach, while effective, still faces a few key challenges.

Hyperparameter Sensitivity. The introduction of metric learning losses alongside the standard cross-entropy loss creates additional complexity in hyperparameter optimization. The performance of our method depends significantly on the weight ratios β and λ, which require careful tuning for each combination of backbone model and downstream task. This dependency can make the optimization process more time-intensive compared to simpler PEFT approaches.

Computational Overhead. Our method introduces additional computational costs through the metric learning losses and their associated operations. While the increased latency remains within practical bounds (typically 5% higher than baseline PEFT methods), it may impact applications with strict real-time requirements or resource constraints.

G. Future Works

To address these limitations, we plan to develop automated hyperparameter optimization strategies, potentially leveraging meta-learning or Bayesian optimization techniques. We will also investigate more efficient metric learning formulations that maintain performance while reducing computational overhead. Additionally, our research will explore hardware-specific optimizations to minimize the latency impact in practical deployments.

Despite these challenges, our experimental results demonstrate that the performance improvements offered by our method consistently outweigh its limitations across diverse tasks and model architectures.

H. Broader Impact

Distribution Aware Visual Prompt Tuning (DA-VPT) has significant implications for both technical advancement and societal applications.

Technical Contributions. Our method advances the performance in parameter-efficient fine-tuning by enabling more efficient adaptation of large vision models to specific domains. The reduced computational requirements for model specialization, coupled with improved performance on fine-grained visual tasks, make sophisticated vision models more accessible and practical for real-world applications.

Potential Applications. DA-VPT could enable significant advances in several high-impact domains. In healthcare, it can facilitate more accurate medical image analysis with limited training data, potentially improving diagnostic accuracy and treatment planning. Environmental protection efforts could benefit from enhanced wildlife monitoring and biodiversity assessment capabilities. The method's efficiency also enables deployment of sophisticated vision models on edge devices, advancing assistive technologies for accessibility applications. Furthermore, industrial applications such as quality control and visual inspection systems could see substantial improvements in accuracy and reliability.

Table 4. Summary of notation used throughout the paper.

Symbol	Domain	Description
$\mathbf{I}$	$\mathbb{R}^{H\times W\times C}$	Input image
N	$\mathbb{N}$	Number of image patches
D	$\mathbb{N}$	Dimension of embedding space
L	$\mathbb{N}$	Number of Transformer layers
l	$\{1,\dots,L\}$	Layer index
$\mathbf{x}_{\text{cls}}$	$\mathbb{R}^D$	Class [CLS] token
$\mathbf{X}$	$\mathbb{R}^{(N+1)\times D}$	Sequence of embeddings
$\mathbf{H}_i$	$\mathbb{R}^{D\times D}$	Attention head i
$\mathbf{Q},\mathbf{K},\mathbf{V}$	$\mathbb{R}^{N\times D}$	Query, Key, Value matrices
M	$\mathbb{N}$	Number of prompt tokens
$\mathbf{P}$	$\mathbb{R}^{M\times D}$	Set of prompt tokens
$\mathbf{p}_k$	$\mathbb{R}^D$	k-th prompt token
$\hat{\mathbf{x}}$	$\mathbb{R}^D$	L2-normalized vector of $\mathbf{x}$
y_i	$\{1,\dots,C\}$	Class label for sample i
C	$\mathbb{N}$	Number of classes
δ	$\mathbb{R}^+$	Margin in metric learning
τ	$\mathbb{R}^+$	Temperature parameter
$\mathcal{P}$	-	Set of all prompts
$\mathcal{P}^+$	-	Set of positive prompts
$\mathcal{X}_p^+$	-	Set of positive visual tokens
$\mathcal{X}_p^-$	-	Set of negative visual tokens
β,λ	$\mathbb{R}^+$	Loss weighting hyperparameters
$\mathbf{S}$	$\mathbb{R}^{C\times D}$	Class representations
$\mathbf{W}_Q^l$	$\mathbb{R}^{D\times D}$	Query projection matrix at layer l
$\mathbf{b}_K,\mathbf{b}_V$	$\mathbb{R}^D$	Bias terms for Key and Value projections

I. More Examples of Attention Maps on Prompts

We also demonstrate some representative attention visualizations from CUB-200-2011 and Stanford Dogs datasets. For each image, we display the attention map corresponding to its class-assigned prompt. The attention patterns demonstrate how our method learns to focus on class-discriminative regions.

References

[1] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016. 3

[2] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, pages 1865–1883, 2017. 3

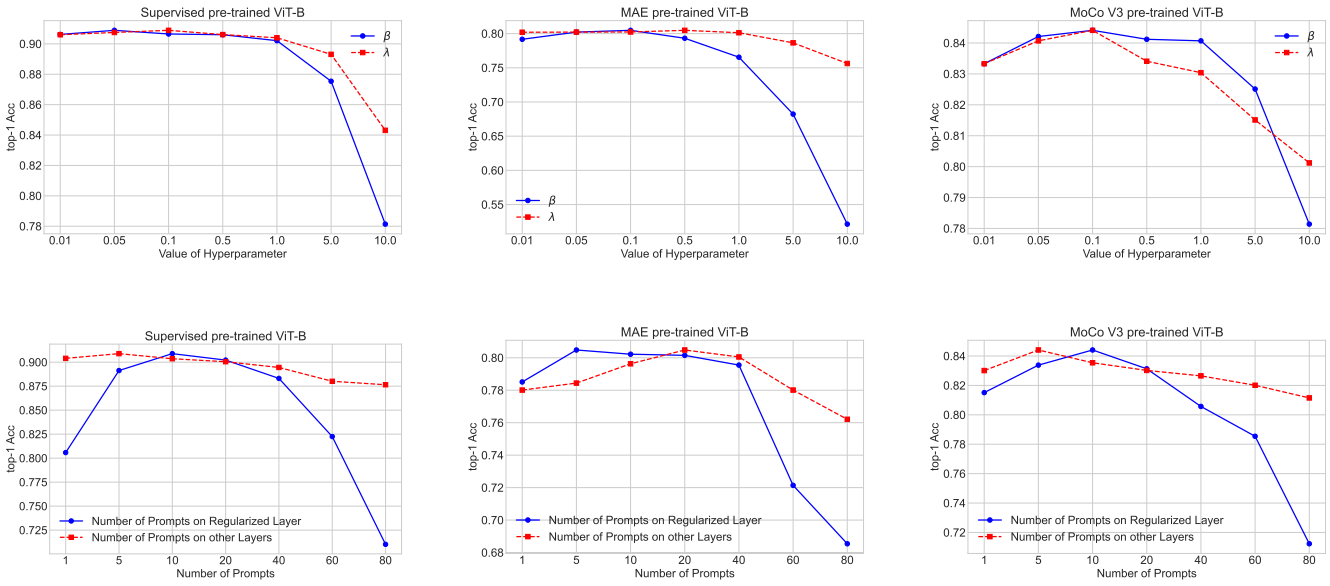


Figure 2. **Impact of Hyperparameters in Three Pre-trained Models on CUB-200-2011**: This figure illustrates the impact of hyperparameters on the performance of our proposed method across three pre-trained models (Supervised ViT, MAE, and MoCo-v3) on the CUB-200-2011 dataset. The hyperparameters investigated include the weight factors β and λ for the two proposed $\mathcal{L}_{\text{ML}}$ losses, the number of prompts in the metric guidance layer, and the number of prompts in other layers. The results show that the optimal weight factors are less than 1.0, indicating that a balanced contribution from the $\mathcal{L}_{\text{ML}}$ losses is beneficial for performance. Furthermore, the number of prompts in the guidance layer exhibits higher sensitivity compared to the number of prompts in other layers, suggesting that the choice of prompt configuration in the guidance layer plays a crucial role in the effectiveness of our method. These findings provide insights into the importance of carefully tuning the hyperparameters to achieve optimal performance across different pre-trained models.

	Natural (7)							*Specialized* (4)				*Structured* (8)							
Methods	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
ViT-B with supervised pre-trained on ImageNet-21K																			
Full fine-tuning [11]	68.9	87.7	64.3	97.2	86.9	87.4	38.8	79.7	93.7	84.2	73.9	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1
Linear probing [11]	63.4	85.0	63.2	97.0	86.3	36.6	51.0	78.5	87.5	68.6	74.0	34.3	30.6	33.2	55.4	12.5	20.0	9.6	19.2
Adapter [10]	74.1	86.1	63.2	97.7	87.0	34.6	50.8	76.3	88.0	73.1	70.5	45.7	37.4	31.2	53.2	30.3	25.4	13.8	22.1
Bias [31]	72.8	87.0	59.2	97.5	85.3	59.9	51.4	78.7	91.6	72.9	69.8	61.5	55.6	32.4	55.9	66.6	40.0	15.7	25.1
VPT-Deep [11]	**78.8**	90.8	65.8	98.0	88.3	78.1	49.6	81.8	**96.1**	83.4	68.4	68.5	60.0	46.5	72.8	73.6	**47.9**	32.9	37.8
DA-VPT+ (ours)	74.4	**92.7**	**74.3**	**99.4**	**91.3**	**91.5**	**86.2**	**96.2**	87.2	**87.2**	**76.3**	**81.3**	**62.58**	**52.82**	**65.3**	**84.9**	51	33.11	**48.7**
ViT-B with MAE pre-trained on ImageNet-1K																			
Full fine-tuning [11]	24.6	84.2	56.9	72.7	74.4	86.6	15.8	81.8	94.0	72.3	70.6	67.0	59.8	45.2	75.3	72.5	47.5	30.2	33.0
DA-VPT+ (ours)	38.7	87.6	64.6	83.5	86.1	83.6	22	85	94.6	79.0	73.2	77.6	63.8	46.9	65.7	90.8	53.0	28.8	47.7
ViT-B with MoCo-V3 pre-trained on ImageNet-1K																			
Full fine-tuning [11]	57.6	91.0	64.6	91.6	79.9	89.8	29.1	85.1	96.4	83.1	74.2	55.2	56.9	44.6	77.9	63.8	49.0	31.5	36.9
DA-VPT+ (ours)	63.5	90.7	69.8	92.5	90.6	90.5	40.5	85.8	96.0	83.9	73.2	80.5	62.3	49.8	63.7	84.2	52.2	30.3	48.9

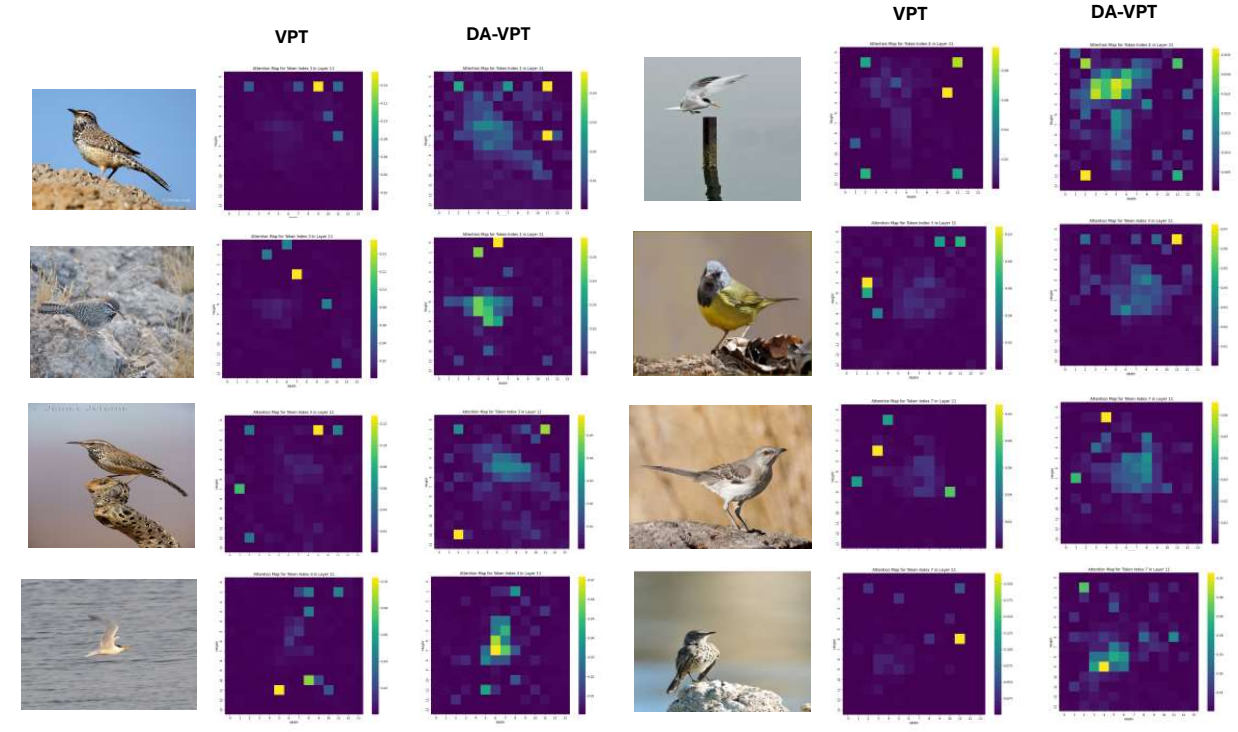
Table 5. Results of details of performance comparisons on the VTAB-1k benchmark with ViT-B/16 models with supervised, MAE and MoCo-V3 pre-training.

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 3

[4] Emma Dugas, Jorge Jared, and Will Cukierski. Diabetic retinopathy detection, 2015. 3

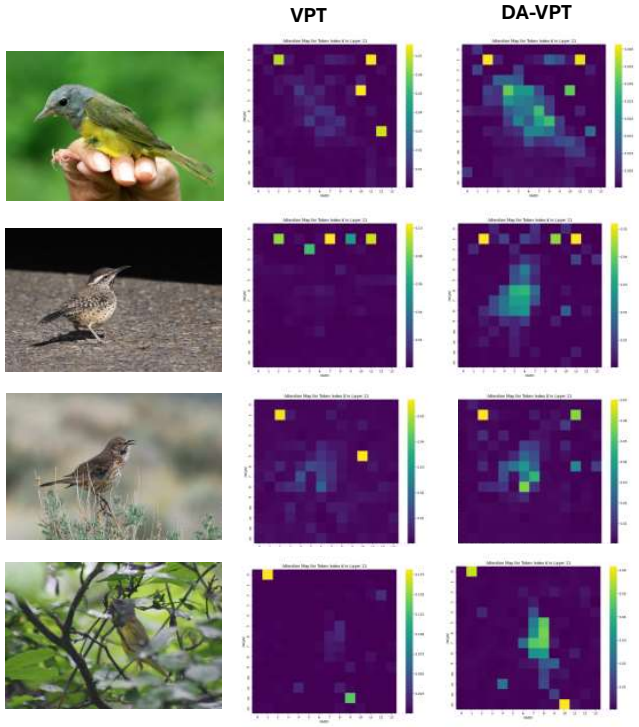
[5] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE TPAMI*, 28(4):594–611, 2006. 3

[6] Mingze Gao, Qilong Wang, Zhenyi Lin, Pengfei Zhu, Qinghua Hu, and Jingbo Zhou. Tuning pre-trained model



(a) Part 1

(b) Part 2



(c) Part 3

(d) Part 4

Figure 3. **Visualization of Class-Specific Attention Maps.** 3a,3b,3c: Examples from CUB-200-2011 showing fine-grained bird features. 3d: Examples from Stanford Dogs highlighting breed-specific characteristics. These visualizations illustrate the model's ability to capture class-relevant visual features across different fine-grained classification tasks.

via moment probing. In *CVPR*, pages 11803–11813, 2023. 1

[7] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, 2017. 1, 3

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, pages 1231–1237, 2013. 3

[9] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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*, pages 2217–2226, 2019. 3

[10] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799. PMLR, 2019. 6

[11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 1, 2, 3, 6

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017. 3

[13] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*. Citeseer, 2011. 1, 3

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[15] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR*. IEEE, 2004. 3

[16] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Neurips*, 35:109–123, 2022. 1

[17] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[19] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset, 2017. 3

[20] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014. 3

[21] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 7. Granada, Spain, 2011. 3

[22] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, pages 1447–1454. IEEE, 2006. 3

[23] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*. IEEE, 2008. 1, 3

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 3

[25] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, pages 595–604, 2015. 1, 3

[26] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, pages 210–218. Springer, 2018. 3

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 3

[28] Yuzhu Wang, Lechao Cheng, Chaowei Fang, Dingwen Zhang, Manni Duan, and Meng Wang. Revisiting the power of prompt for visual tuning. *arXiv preprint arXiv:2402.02382*, 2024. 2

[29] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 3

[30] Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sungroh Yoon. Improving visual prompt tuning for self-supervised vision transformers. In *ICML*, pages 40075–40092. PMLR, 2023. 3

[31] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 6

[32] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1, 3

[33] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 1

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127:302–321, 2019. 3