SA^{2}VP: Spatially Aligned-and-Adapted Visual Prompt

Wenjie Pei[1]*, Tongqi Xia[1]*, Fanglin Chen[1], Jinsong Li[2], Jiandong Tian[3], Guangming Lu[1]†

[1]Harbin Institute of Technology, Shenzhen
[2]Shenzhen Jiang & Associates Creative Design Co., Ltd
[3]Shenyang Institute of Automation, Chinese Academy of Sciences
wenjiecoder@outlook.com, tongqixia325@gmail.com, chenfanglin@hit.edu.cn, as@jaid.cn, tianjd@sia.cn,
luguangm@hit.edu.cn

Abstract

As a prominent parameter-efficient fine-tuning technique in NLP, prompt tuning is being explored its potential in computer vision. Typical methods for visual prompt tuning follow the sequential modeling paradigm stemming from NLP, which represents an input image as a flattened sequence of token embeddings and then learns a set of unordered parameterized tokens prefixed to the sequence representation as the visual prompts for task adaptation of large vision models. While such sequential modeling paradigm of visual prompt has shown great promise, there are two potential limitations. First, the learned visual prompts cannot model the underlying spatial relations in the input image, which is crucial for image encoding. Second, since all prompt tokens play the same role of prompting for all image tokens without distinction, it lacks the fine-grained prompting capability, i.e., individual prompting for different image tokens. In this work, we propose the Spatially Aligned-and-Adapted Visual Prompt model (*SA*^{2}*VP*), which learns a two-dimensional prompt token map with equal (or scaled) size to the image token map, thereby being able to spatially align with the image map. Each prompt token is designated to prompt knowledge only for the spatially corresponding image tokens. As a result, our model can conduct individual prompting for different image tokens in a fine-grained manner. Moreover, benefiting from the capability of preserving the spatial structure by the learned prompt token map, our *SA*^{2}*VP* is able to model the spatial relations in the input image, leading to more effective prompting. Extensive experiments on three challenging benchmarks for image classification demonstrate the superiority of our model over other state-of-the-art methods for visual prompt tuning. Code is available at *https://github.com/tommy-xq/SA2VP*.

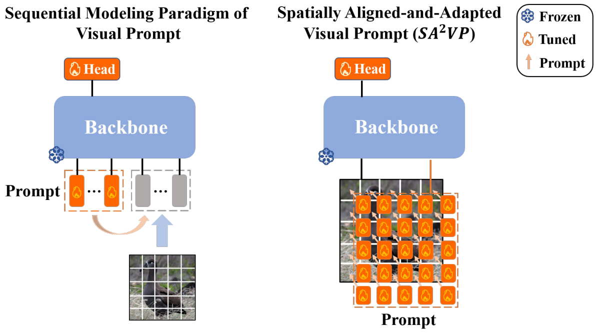
Sequential Modeling Paradigm of Visual Prompt
Spatially Aligned-and-Adapted Visual Prompt (SA^2VP)
Frozen
Tuned
Prompt
Head
Backbone
Prompt
Head
Backbone
Prompt

Figure 1: Typical methods for visual prompting tuning follow the sequential modeling paradigm which learns an unordered sequence of prompts prefixed to the image representation. All prompt tokens play the same role of prompting for all image tokens. In contrast, our *SA*^{2}*VP* learns a 2D prompt token map, which aligns with the image token map spatially. As a result, our *SA*^{2}*VP* is able to 1) conduct individual prompting for different image tokens in a fine-grained manner and 2) model the spatial relations in the input image by preserving the spatial structure in the prompt map.

Introduction

Adapting large pre-trained vision models to a specific downstream vision task via parameter-efficient fine-tuning (PEFT) has proven to be an effective and efficient strategy, particularly in the scenarios where the training data is limited. It not only distills the prior knowledge relevant to the downstream task from the pre-trained model to achieve better performance, but also substantially improves the training efficiency by incurring only a small amount of parameter-tuning overhead. As a prominent PEFT method, prompt tuning has achieved remarkable success in NLP field and is being explored its potential in computer vision (Bahng et al. 2022; Huang et al. 2023; Jia et al. 2022; Das et al. 2023; Han et al. 2023; Bar et al. 2022).

Most of existing methods for visual prompt tuning (Jia et al. 2022; Das et al. 2023; Han et al. 2023) follow the sequential modeling paradigm, which stems from the learning way of prompt in NLP, as shown in Figure 1. A prominent example is VPT (Jia et al. 2022), which encodes an input image as a sequential representation using pre-trained vision Transformer models and then learns a set of prompt tokens as auxiliary input prefixed to the image representation. The learned vision prompts are expected to distill the relevant knowledge from the pre-trained model, thereby adapting the pre-trained model to the downstream target task. While such sequential modeling paradigm of visual prompts has shown great promise, it suffers from two potential limitations. First,

*These authors contributed equally.
†Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

since no spatial structure can be captured by the simple unordered sequence structure of the prompt tokens, **it cannot model the underlying spatial relations of the input image which is crucial to feature learning of images**. Second, all visual prompts in the sequential modeling paradigm play the same role of prompting for all image tokens without distinction, thus **it is unable to conduct individual prompting for different image tokens to have a fine-grained prompting**. Nevertheless, different image tokens could have different semantics and thus may require specific prompting.

To address above two limitations, in this work we propose the Spatially Aligned-and-Adapted Visual Prompt model (*SA*^{2}*VP*). As illustrated in Figure 1, instead of learning an unordered sequence of visual prompts in the sequential modeling paradigm, our *SA*^{2}*VP* learns a two-dimensional prompt token map which has equal (or scaled) size as the image token map. Besides, the learned positional embeddings of the pre-trained model are incorporated into the prompt token map to preserve the spatial structure. As a result, *SA*^{2}*VP* is able to align the prompt token map with the image token map spatially, which yields two essential advantages. First, it can learn the spatial relations of the input image since the spatial structure is well preserved in the prompt token map, which allows our model to distill more effective knowledge by the visual prompts. Second, each prompt token is designated to prompt only for the spatially corresponding image tokens, which enables our model to perform prompting individually for different image tokens in a fine-grained manner. To conclude, we make following contributions.

- We design a novel visual prompt structure, which can align with the image token map in 2D spatial space. Such design enables our proposed *SA*^{2}*VP* to not only model the spatial relations in the input image, but also conduct individual prompting in a fine-grained manner.
- We devise a simple yet effective technical implementation for our *SA*^{2}*VP*. It follows the siamese dual-pathway architecture, with bilateral interaction between two pathways based on the proposed spatially-aligned cross attention operation.
- We conduct extensive experiments on three challenging benchmarks for image classification, which 1) reveal that our model compares favorably with other state-of-the-art methods for visual prompting tuning; 2) validate the effectiveness of each essential design by ablation study.

Related Works

Parameter-Efficient Fine-Tuning. Parameter-efficient fine-tuning (PEFT) methods perform task adaptation by introducing a small amount of additional learnable parameters while freezing the pre-trained model, which is more efficient than the classical full fine-tuning paradigm. PEFT methods typically fall into two categories: prompt tuning (elaborated later) and adapter tuning. Adapter tuning (Houlsby et al. 2019; Pfeiffer et al. 2020; Rücklé et al. 2020; Karimi Mahabadi, Henderson, and Ruder 2021; Rebuffi, Bilen, and Vedaldi 2017) generally inserts lightweight transformation module into the middle layers of large pre-trained models to perform feature transformation for task adaptation. Since only the parameters of adapters are learnable, it is more efficient than the standard fine-tuning methods. A prominent example is AdaptFormer (Chen et al. 2022), which designs a plug-and-play module named AdaptMLP to replace the primitive MLP block in ViT. We draw inspiration from Adapter and devise the prompt adapter in our *SA*^{2}*VP* to adapt the prompted knowledge to the downstream task.

Prompt Tuning in NLP. Prompt tuning is initially investigated in NLP and aims to guide the pre-trained large language models to extract the relevant knowledge to the downstream task for task adaptation. The prompts can be handcrafted (Petroni et al. 2019; Brown et al. 2020), whose performance relies heavily on the human experience, or automatically generated either in the textual form (Shin et al. 2020; Wallace et al. 2019) or in the parametric modeling way (Li and Liang 2021; Liu et al. 2021a; Lester, Al-Rfou, and Constant 2021).

Visual Prompt Tuning. Inspired by the great success of prompting tuning in NLP, it is being explored extensively in computer vision. It is difficult to handcraft the visual prompt directly, thus the visual prompts is either represented as some kind of prior information like object bounding box, mask, or salient points within the target objects (Bar et al. 2022; Kirillov et al. 2023) or modeled as parameterized embeddings automatically (Jia et al. 2022; Das et al. 2023; Han et al. 2023; Huang et al. 2023; Bahng et al. 2022). In the latter case, the visual prompts are typically learned as a sequence of embeddings prefixed to the image representations. Such sequential modeling paradigm cannot model the spatial structure of image data. Meanwhile, it is unable to conduct fine-grained prompting for different image tokens since all prompt tokens play the same role of prompting for all image tokens. Our *SA*^{2}*VP* is proposed to address these limitations.

Approach

Preliminaries

Visual prompt tuning aims to adapt a large pre-trained vision model to a downstream task with little overhead of parameter tuning. It typically learns prompts as auxiliary input to guide the model to mine the relevant knowledge to the downstream task from the pre-trained model. A prominent method for visual prompt tuning is VPT (Jia et al. 2022), which performs sequential modeling paradigm of visual prompt.

Sequential Modeling Paradigm of Visual Prompt. Given an input image I, sequential modeling paradigm for prompt tuning first divides the image I into N equal-sized patches and encodes I into a flattened sequence of token embeddings $\mathbf{E} \in \mathbb{R}^{N\times d}$, where each of N tokens corresponds to one image patch and d is the feature dimension. Then a set of p parameterized token embeddings $\mathbf{P} \in \mathbb{R}^{p\times d}$ are learned as visual prompts and prefixed to the sequence representation $\mathbf{E}$. The formed sequence is fed into the pre-trained vision model $\mathcal{F}_M$ to produce the prediction $\mathbf{O}$ for the downstream task:

$$\mathbf{O} = \mathcal{F}_M([\mathbf{P}, \mathbf{E}]). \tag{1}$$

Note that all the parameters in $\mathcal{F}_M$ are frozen and only $\mathbf{P}$ and a classification head are learnable, which substantially reduces the tuning overhead.

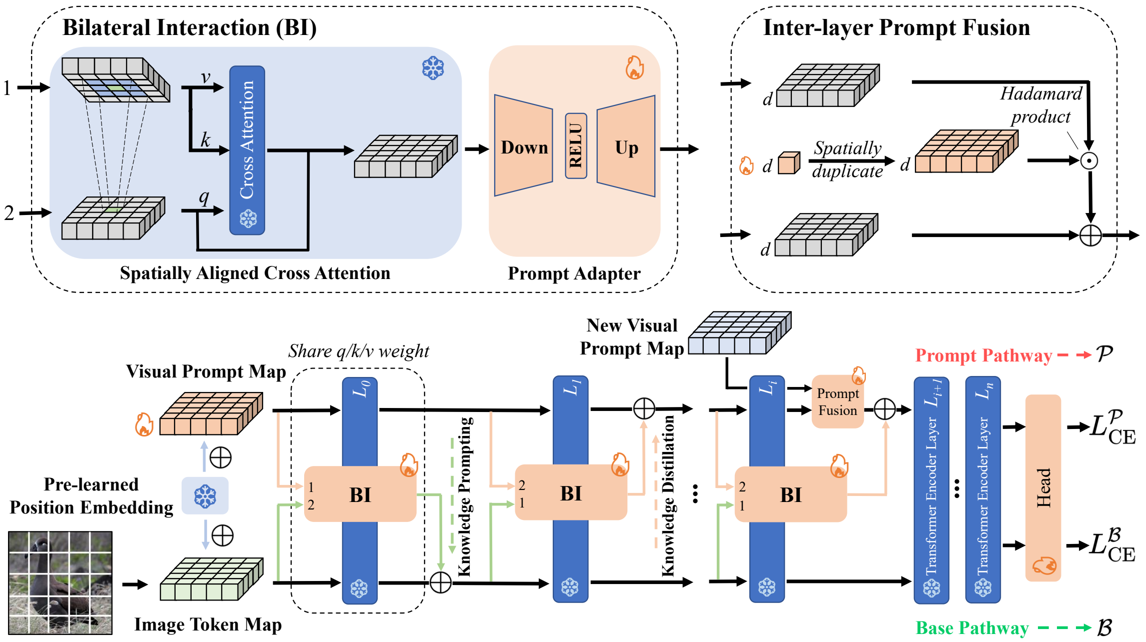

Figure 2: Our SA^2VP follows siamese dual-pathway architecture which consists of Prompt pathway for learning visual prompts and Base pathway for image encoding. Bilateral interaction is performed between two pathways via the designed spatially aligned cross attention. The inter-layer prompt fusion scheme allows for learning multiple prompt maps in different layers.

Intuitively, the relevant knowledge to the downstream task is distilled from the pre-trained model into the learned visual prompt tokens during training. As a result, the pre-trained model can be adapted to the downstream task by incorporating the distilled knowledge into modeling inference. While such sequential learning paradigm of visual prompt has shown great promise, we investigate two limitations of it to unleash more potential of visual prompt tuning:

1. The learned visual prompts in sequential structure cannot capture the spatial relations in the input image which are important for image encoding.
2. Since all visual prompts play the same role of prompting for all image tokens, such learning paradigm cannot perform individual prompting for different image tokens in a fine-grained manner.

Spatially Aligned-and-Adapted Visual Prompt

Siamese Dual-Pathway Architecture. To tackle aforementioned two limitations of sequential modeling paradigm, we design the Spatially Aligned-and-Adapted Visual Prompt (SA^2VP) which is illustrated in Figure 2. It consists of two pathways: Prompt pathway $\mathcal{P}$ for learning the visual prompts and Base pathway $\mathcal{B}$ for learning image features based on the knowledge prompting from visual prompts. Both pathways employ the same pre-trained vision model with frozen parameters as the backbone. Herein we take ViT (Dosovitskiy et al. 2020) as an instantiation of the pre-trained vision model while our model is readily applied to other classical vision models like Swin Transformer (Liu et al. 2021b) or ResNet (He et al. 2016). Thus the proposed SA^2VP follows siamese architecture with bilateral interactions between two pathways. Prompt pathway distills the relevant knowledge to the downstream task via interaction with Base pathway while Base pathway acquires the prompted knowledge by attending to Prompt pathway.

Spatial-Structure-Preserved Prompt Map. The essential design of the proposed SA^2VP lies in the spatially aligned 2D structure of the visual prompt tokens, which is in contrast to the sequential prompts of the sequential modeling paradigm. As shown in Figure 2, our SA^2VP learns a 2D prompt token map which has equal size (or scaled) as the image token map. Meanwhile, the pre-learned positional embeddings of the pre-trained model are added to each token of the prompt map, thus the spatial structure can be well preserved by the prompt map. As a result, SA^2VP is able to align the prompt token map with the image token map spatially, which yields two key benefits corresponding to the two limitations of the sequential modeling paradigm. First, each prompt token is responsible for prompting knowledge for the spatially corresponding image tokens specifically, which enables our SA^2VP to perform fine-grained prompting. Second, benefiting from the spatial alignment between the prompt token map and the image token map, the learned prompt token map can well preserve the spatial structure of the image and thereby models the spatial relations contained in the input image. This is potentially beneficial for more effective image encoding.

Bilateral Interaction by Spatially-Aligned bidirectional Cross Attention. Prompt pathway and Base pathway perform bilateral interaction with each other in the same depth of layers via the designed spatially-aligned cross attention operation in bidirectional manner. To be specific, the prompt tokens of Prompt pathway serve as the query of cross attention to distill the relevant knowledge to the downstream task from Base pathway. Conversely, the image tokens of Base pathway query Prompt pathway to obtain the prompting knowledge. As illustrated in Figure 2, two directions of interactions are performed asynchronously.

Different from the typical cross attention that performs global interactions between a query token and all value tokens for each attention operation, during the proposed spatially-aligned cross attention, the query token only attends to the value tokens in the spatially corresponding location. Taking the cross attention for the interaction from Base pathway to Prompt pathway as an example, a token located at $\langle x, y\rangle$ in the l-th layer of Base pathway serves as the query $\mathbf{q}^{\mathcal{B},l}_{\langle x,y\rangle}$ and attends to the tokens in the spatially corresponding locations in the l-th layer of Prompt pathway which is a $c \times c$ squared window $\mathbf{R}^{\mathcal{P},l}$ centered at $\langle x, y\rangle$:

$$\begin{aligned}&\mathbf{q} = \mathbf{q}^{\mathcal{B},l}_{\langle x,y\rangle}\mathbf{W}^q,\ \mathbf{K} = \mathbf{R}^{\mathcal{P},l}\mathbf{W}^K,\ \mathbf{V} = \mathbf{R}^{\mathcal{P},l}\mathbf{W}^V,\\ &\mathbf{o}^{\mathcal{B},l}_{\langle x,y\rangle} = \mathrm{softmax}(\frac{\mathbf{q}\mathbf{K}^\top}{\sqrt{d_s}})\mathbf{V} + \mathbf{q}^{\mathcal{B},l}_{\langle x,y\rangle}.\end{aligned} \tag{2}$$

Herein, $\mathbf{R}^{\mathcal{P},l}$ denotes the set of tokens in the squared window of size $c\times c$ centered at $\langle x, y\rangle$ in the l-th layer of Prompt pathway while d_s is a scaling factor. Residual connection is employed after the cross attention. Note that all the involved parameters in Equation 2, including $\mathbf{W}^q$, $\mathbf{W}^K$ and $\mathbf{W}^V$, are reused from the self-attention operation in the l-th layer of the pre-trained vision model, which eliminates the learning overhead. c is a hyper-parameter to balance between the local specificity and the global smoothness for prompting. $c = 1$ corresponds to the most fine-grained prompting.

Prompt Adapter. To facilitate the feature adaptation of prompted knowledge from the pre-trained backbone model to the downstream task, we additionally insert a lightweight prompt adapter after the cross attention operation. As shown in Figure 2, the prompt adapter consists of two MLP layers with a ReLU layer and a layer-normalization layer in between. Thus, the prompted feature at $\langle x, y\rangle$ is adapted by:

$$\mathbf{o}^{\mathcal{B},l}_{\langle x,y\rangle} = \mathrm{ReLU}(\mathrm{LN}(\mathbf{o}^{\mathcal{B},l}_{\langle x,y\rangle}) \cdot W_{\text{down}}) \cdot W_{\text{up}}, \tag{3}$$

where the feature dimension is first shrunk from d to t ($t \ll d$) by the learnable matrix $W_{down} \in R^{d\times t}$ and then expanded back to d by $W_{up} \in R^{t\times d}$. The adapted prompt feature $\mathbf{o}^{\mathcal{B},l}_{\langle x,y\rangle}$ is fused with the original output features $\mathbf{f}^{\mathcal{B},l}_{\langle x,y\rangle}$ in the l-th layer of the backbone by weighted addition:

$$\mathbf{f}^{\mathcal{B},l}_{\langle x,y\rangle} = \mathbf{f}^{\mathcal{B},l}_{\langle x,y\rangle} + \gamma\mathbf{o}^{\mathcal{B},l}_{\langle x,y\rangle}, \tag{4}$$

where γ is the balancing weight tuned on a validation set.

Note that our prompt adapter focuses on adapting the prompted knowledge, which slightly differs from the classical design of adapters (Houlsby et al. 2019; Pfeiffer et al. 2020; Rücklé et al. 2020; Karimi Mahabadi, Henderson, and Ruder 2021; Rebuffi, Bilen, and Vedaldi 2017) performing feature adaptation over each layer of the backbone network.

Inter-layer Prompt Fusion. Similar to VPT, we can either learn the visual prompts only in the first layer termed as 'shallow mode' or learn prompts at multiple ViT layers named 'deep mode'. Generally, more layers of visual prompts in the deep mode have more learning capacity than the shallow mode. Besides, the visual prompts at different layers can potentially distill and prompt different levels of knowledge corresponding to the features in different depth of the backbone network. In our deep mode, we learn three prompt layers at Layer-0, 4, 8 of the backbone, respectively, which are set empirically and correspond to three different depths of the backbone.

Inspired by the powerful capability of feature adaptation of SSF (Lian et al. 2022), which performs affine transformation per each feature dimension, we perform similar transformation for the newly learned prompt token map before fused into Prompt pathway to bridge the feature gap between them. As shown in Figure 2, for a newly learned prompt token map $\mathbf{P}^l \in \mathbb{R}^{h^l\times w^l\times d}$ in the l-th layer of the backbone, we learn a parameterized scaling vector $\mathbf{e} \in \mathbb{R}^d$ and duplicate it spatially to have equal spatial size ($h^l \times w^l$) with $\mathbf{P}^l$, denoted as $\mathbf{E} \in \mathbb{R}^{h^l\times w^l\times d}$. Then we perform Hadmard product between $\mathbf{E}$ with $\mathbf{P}^l$ for feature scaling and conduct inter-layer prompt fusion by adding the scaled result to the feature map $\mathbf{O}^{\mathcal{P},l}$ in the l-th layer of Prompt pathway:

$$\mathbf{O}^{\mathcal{P},l} = \mathbf{E} \odot \mathbf{P}^l + \mathbf{O}^{\mathcal{P},l}. \tag{5}$$

Prompt Learning. Following the routine experimental setting (Jia et al. 2022), we use image classification as both the pre-training and downstream tasks whilst using different source data. A straightforward way to optimize our *SA*^{2}*VP* is to employ Cross-Entropy loss for supervised learning. The output features from Base pathway are finally used for classification during inference, thus we perform Cross-Entropy loss primarily on Base pathway. Besides, to explicitly enhance the capability of the visual prompts on learning the relevant knowledge to the downstream task, we also conduct Cross-Entropy loss on Prompt pathway as an auxiliary loss to supervise our model. Thus, our model is supervised by:

$$L_{all} = \lambda L^{\mathcal{B}}_{\mathrm{CE}}(y_r, y^*) + (1-\lambda)L^{\mathcal{P}}_{\mathrm{CE}}(y_p, y^*), \tag{6}$$

where y_p and y_r are the prediction by Prompt pathway and Based pathway respectively while y^* is the corresponding groundtruth. λ is a hyper-parameter to balance two losses.

Experiments

Experimental Setup

Datasets. We conduct experiments on three challenging benchmarks across diverse scenes: FGVC, HTA and VTAB-1k (Zhai et al. 2019). FGVC benchmark contains 5 image datasets, including CUB (Wah et al. 2011), NABirds (Van Horn et al. 2015), Oxford Flowers (Nilsback and Zisserman 2008), Stanford Dogs (Khosla et al. 2011) and Stanford Cars (Gebru et al. 2017). We follow VPT to split data for training and test. The head tuning adaptation benchmark (HTA) (Huang et al. 2023) comprises 10 datasets including CIFAR10 (Krizhevsky, Hinton et al. 2009), CIFAR100 (Krizhevsky, Hinton et al. 2009), DTD (Cimpoi et al. 2014), CUB-200 (Wah et al. 2011),

Methods	CUB-200-2011	NABirds	Oxford Flowers	Stanford Dogs	Stanford Cars	**Mean**	Tuned/Total(%)
Full Fine-tune	87.3	82.7	98.8	89.4	**84.5**	88.54	100
Head Fine-tune	85.3	75.9	97.9	86.2	51.3	79.32	0.12
AdaptFormer	84.7	75.2	97.9	84.7	83.1	85.12	0.44
LoRA	84.9	79.0	98.1	88.1	79.8	85.98	0.55
VPT-shallow	86.7	78.8	98.4	90.7	68.7	84.62	0.31
VPT-deep	88.5	84.2	99.0	90.2	83.6	89.11	0.98
E^{2}VPT	**89.1**	84.6	99.1	90.5	82.8	89.22	0.65
SA^{2}VP (Ours)	**89.1**	**85.8**	**99.3**	**92.1**	84.1	**90.08**	0.99

Table 1: Performance of different methods on FGVC benchmark based on ViT backbone.

Methods	DTD	CUB-200	NABirds	Dogs	Flowers	Food-101	CIFAR-100	CIFAR-10	GTSRB	SVHN	**Mean**
Full Fine-tune	64.3	87.3	82.7	89.4	98.8	84.9	68.9	97.4	**97.1**	87.4	85.8
Head Fine-tune	63.2	85.3	75.9	86.2	97.9	84.4	63.4	96.3	68.0	36.6	75.7
Adapter	62.7	87.1	84.3	89.8	98.5	86.0	74.2	97.7	91.1	36.3	80.8
VPT-deep	65.8	88.5	84.2	90.2	99.0	83.3	78.8	96.8	90.7	78.1	85.5
AdaptFormer	74.4	84.7	75.2	84.7	97.9	89.1	**91.4**	**98.8**	97.0	**96.5**	89.0
DAM-VP	73.1	87.5	82.1	**92.3**	**99.2**	86.9	88.1	97.3	90.6	87.9	88.5
SA^{2}VP (Ours)	**75.6**	**89.0**	**86.0**	92.2	**99.2**	**90.1**	91.3	98.6	96.3	96.4	**91.5**

Table 2: Performance of different methods on HTA benchmark based on ViT backbone.

NABirds (Van Horn et al. 2015), Stanford-Dogs(Khosla et al. 2011), Oxford-Flowers (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014), GTSRB (Stallkamp et al. 2012) and SVHN (Netzer et al. 2011). Using default train-val-test data split, we follow the experimental configuration of DAM-VP (Huang et al. 2023) to have a fair comparison. VTAB-1k benchmark consists of 19 datasets belonging to three groups of tasks: 1) 'Natural' group that captures natural objects by standard cameras, 2) 'Specialized' group taking photos utilizing specialized equipments like satellite or medical device, and 3) 'Structured' group performing geometric comprehension such as counting and distance perception. Each dataset in VTAB-1k contains 1000 training images, among which 800 images are for training and the left is for validation (Zhai et al. 2019).

Implementation Details. Following VPT, we mainly conduct experiments using ViT-B/16 pre-trained on ImageNet-21K (Deng et al. 2009) as the backbone. Besides, we also evaluate our model instantiated with the pre-trained Swin Transformer (Liu et al. 2021b) on ImageNet-21K. During training, all the parameters of the backbone are frozen while only the prompt map and the classification head are learnable. AdamW (Loshchilov and Hutter 2017) is used for optimization with the initial learning rate $1e^{-3}$, weight decay $1e^{-4}$ and the batch size 64 or 128. Since all experiments are performed for image classification on all benchmarks, we use classification accuracy as the evaluation metric.

Comparison with State-of-the-art Methods

We compare our *SA^{2}VP* with four types of fine-tuning methods: 1) standard fine-tuning methods which either fine-tune all parameters of the backbone, termed as 'Full Fine-Tune', or only fine-tune the linear classification head dubbed 'Head Fine-Tune'; 2) Adapter tuning method, with the representative 'AdaptFormer' (Chen et al. 2022) in our comparison; 3) visual prompt tuning methods, including 'VPT', 'E^{2}VPT' (Han et al. 2023), 'EXPRES' (Das et al. 2023) and 'DAM-VP' (Huang et al. 2023). Note that 'VPT', 'E^{2}VPT' and 'EXPRES' all follow the sequential modeling paradigm of visual prompt; 4) LoRA (Hu et al. 2022), which tunes partial parameters of the pre-trained model by learning the variation of the tunable parameters directly.

Quantitative Evaluation with 'ViT' backbone. The experimental results on **FVGC** benchmark in Table 1 show that our *SA^{2}VP* achieves the best *mean* performance. In particular, our *SA^{2}VP* outperforms both VPT and E^{2}VPT which follow the sequential modeling paradigm. Besides, our model also outperforms AdaptFormer by a large margin. Table 2 shows that our *SA^{2}VP* outperforms all other methods by a large margin in terms of *mean* metric on **HTA** benchmark. Notably, our model performs consistently well on all sub-datasets, which reveals the strong robustness of our model. The results on **VTAB-1k** in Table 3 show that our *SA^{2}VP* achieves the best *mean* performance on all three groups of datasets. Particularly, 'Structured' group is much more challenging than the other two groups since comprehending the structure information like counting or distance perception is generally more difficult than the classification of object categories. Our model surpasses other methods substantially on 'Structured' group. Further, our model outperforms all the methods following sequential modeling paradigm of visual prompts, including VPT, EXPRES and E^{2}VPT, which reveals the advantages of our model.

Quantitative Evaluation with 'Swin Transformer' backbone. Table 4 shows the comparison between our model and other methods when instantiating the backbone with Swin Transformer. The results show that our model achieves the best performance on 'Specialized' and 'Structured' groups

	Datasets	Full Fine-tune	Head Fine-tune	AdaptFormer	LoRA	VPT-deep	EXPRES	E^{2}VPT	SA^{2}VP (Ours)
Natural	CIFAR-100	68.9	63.4	70.8	67.1	**78.8**	78.0	78.6	73.0
	Caltech101	87.7	85.0	91.2	91.4	90.8	89.6	89.4	**91.9**
	DTD	64.3	63.2	**70.5**	69.4	65.8	68.8	67.8	**70.5**
	Flowers102	97.2	97.0	**99.1**	98.8	98.0	98.7	98.2	**99.1**
	Pets	86.9	86.3	**90.9**	90.4	88.3	88.9	88.5	90.8
	SVHN	**87.4**	36.6	86.6	85.3	78.1	81.9	85.3	84.7
	Sun397	38.8	51.0	54.8	54.0	49.6	51.9	52.3	**56.8**
	Mean	75.88	68.93	80.56	79.49	78.48	79.69	80.01	**80.97**
	Tuned/Total(%)	100	0.10	–	–	0.23	–	0.19	0.37
Specialized	Patch Camelyon	79.7	78.5	83.0	84.9	81.8	84.8	82.5	**86.0**
	EuroSAT	95.7	87.5	95.8	95.3	96.1	96.2	**96.8**	95.9
	Resisc45	84.2	68.6	84.4	84.4	83.4	80.9	84.8	**85.8**
	Retinopathy	73.9	74.0	**76.3**	73.6	68.4	74.2	73.6	75.2
	Mean	83.36	77.16	84.88	84.55	82.43	84.03	84.43	**85.73**
	Tuned/Total(%)	100	0.01	–	–	0.57	–	0.17	0.45
Structured	Clevr/count	56.3	34.3	81.9	**82.9**	68.5	66.5	71.7	76.6
	Clevr/distance	58.6	30.6	64.3	**69.2**	60.0	60.4	61.2	61.8
	DMLab	41.7	33.2	49.3	49.8	46.5	46.5	47.9	**50.8**
	KITTI/distance	65.5	55.4	**80.3**	78.5	72.8	77.6	75.8	79.9
	dSprites/location	57.5	12.5	76.3	75.7	73.6	78.0	80.8	**84.5**
	dSprites/orientation	46.7	20.0	45.7	47.1	47.9	49.5	48.1	**52.8**
	SmallNORB/azimuth	25.7	9.6	31.7	31.0	32.9	26.1	31.7	**34.7**
	SmallNORB/elevation	29.1	19.2	41.1	44.0	37.8	35.3	41.9	**45.3**
	Mean	47.64	26.84	58.83	59.78	54.98	54.99	57.39	**60.80**
	Tuned/Total(%)	100	0.01	–	–	1.14	–	0.51	0.79

Table 3: Performance of different methods on VTAB-1k benchmark based on ViT backbone.

Methods	Tuned/Total	VTAB-1k		
		Natural	Specialized	Structured
Full Fine-tune	100.00%	79.10	86.21	59.65
Head Fine-tune	0.06%	73.52	80.77	33.52
VPT-deep	0.25%	76.78	83.33	51.85
E^{2}VPT	0.21%	**83.31**	84.95	57.35
SA^{2}VP (Ours)	0.29%	80.81	**86.30**	**60.03**

Table 4: Performance of different methods on VTAB-1k benchmark based on Swin Transformer backbone.

Ablated Variants	VTAB-1k		
	Natural	Specialized	Structured
w/o Spatial alignment	79.37	83.65	56.98
w/o Prompt adapter	79.86	82.63	59.30
w/o Deep prompting	80.10	85.13	59.16
w/o Supervision on Prompt	80.37	85.73	57.36
Intact *SA^{2}VP*	80.97	85.73	60.80

Table 5: Ablation study of various core technical components of our *SA^{2}VP* on VTAB-1k. 'w/o' denotes ablation.

while ranking second on 'Natural'. It is worth noting that our model outperforms VPT by a large margin on all three groups with comparable ratio of tuning parameters.

Qualitative Evaluation. To evaluate the effect of prompt tuning on the feature learning for classification, we further conduct two sets of qualitative evaluation. First, we visualize the t-SNE (Van der Maaten and Hinton 2008) map of the extracted features in the last layer of the backbone (ViT) for different methods on randomly selected sub-datasets from VTAB-1k benchmark (one for each group). Figure 3 clearly show that our *SA^{2}VP* has more precise clustering w.r.t. different classes than other methods, which reveals the higher quality of learned features by our model. Subsequently, we visualize the GradCAM (Selvaraju et al. 2017) map of features in the last layer of the backbone for these methods, which indicates the attention area of learned features by each method. The results in Figure 4 show that our *SA^{2}VP* is more focused on the target object than other methods.

Ablation Study

Effect of spatially aligned prompting. To validate the effect of the spatially aligned prompting, we perform global cross attention instead of the spatially aligned cross attention during bilateral interaction, resulting the global variant which is essentially similar to VPT. The large performance gap between it and the intact model in Table 5 demonstrates the effectiveness of the spatially aligned prompting.

To obtain more insight into the effect of the spatial alignment, we further visualize the similarity relations between the learned prompt map and the image token map. We randomly select a visually salient token on the object from the image feature map and calculate the Cosine similarities be-

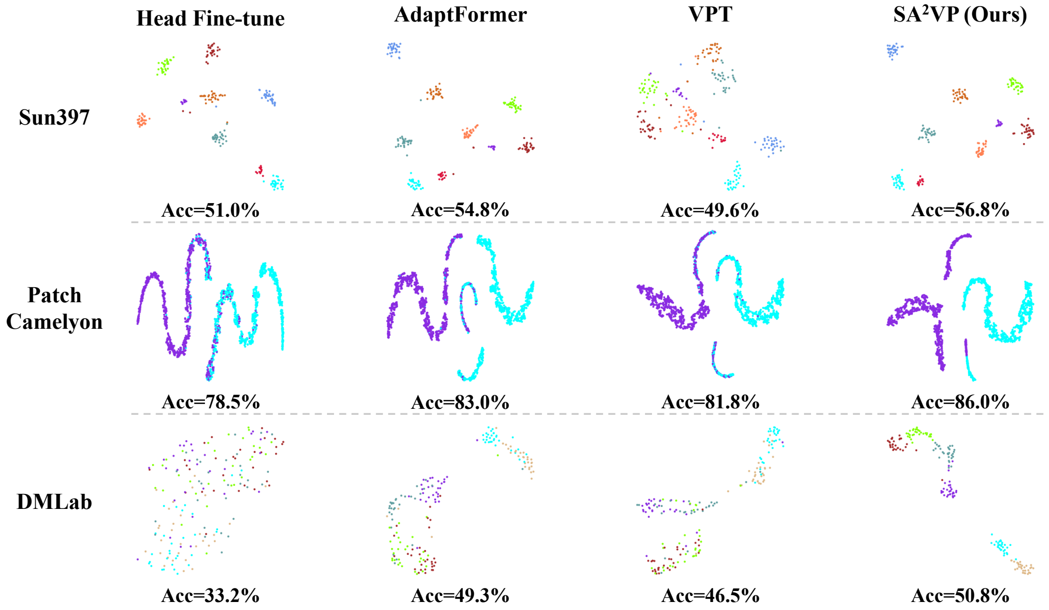

Figure 3: t-SNE maps of learned features in the last layer of the backbone by four different methods on three datasets randomly selected from three groups of VTAB-1k benchmark.

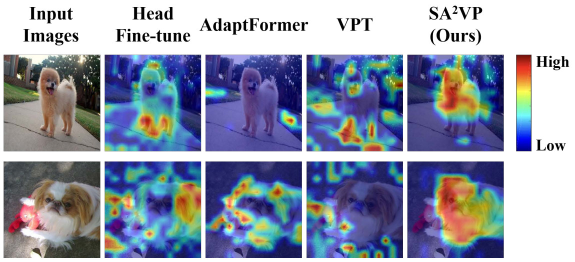

Figure 4: GradCAM maps of features in the last layer of the backbone by different methods on two randomly selected samples. Our model is more focused on the target objects.

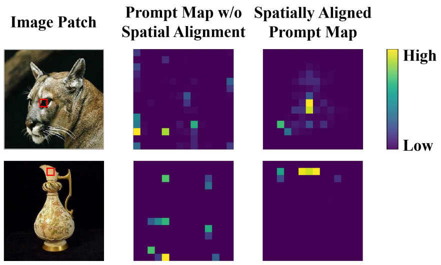

Figure 5: Cosine similarity maps between a visually salient token in the image token map and all learned prompt tokens. Our model can indeed preserve such spatial alignment.

tween the salient token and all tokens in the prompt map. We compare the result of our model with that of the global variant. Figure 5 indicate that our model can indeed preserve the spatial alignment between the learned prompt map and the image feature map, which is potentially beneficial to our model to capture spatial relations in the visual prompts.

Effect of Prompt adapter. Table 5 show that the designed Prompt adapter is beneficial to the performance of our SA^2VP on all three groups of datasets, which indicates its effectiveness. It is reasonable since the prompt tuning and the adapter play different role to improve the performance. The prompts distill the relevant knowledge from the pre-trained model while the prompt adapter further bridges the feature gap between the pre-trained model and the downstream task.

Deep prompting mode vs shallow prompting mode. We also compare the performance of 'deep' and 'shallow' prompting modes in Table 5, which reveals that the shallow mode performs slightly worse than the deep mode on 'Natural' and 'Specialized' groups. Nevertheless, the deep mode shows larger superiority than the shallow mode on the more challenging 'Structured' group.

Effect of supervision on Prompt pathway. Finally, we investigate the effect of supervision on Prompt pathway. Table 5 shows that such design is effective in two groups, especially on the more challenging 'Structured' group.

Conclusion

In this work we have presented the Spatially Aligned-and-Adapted Visual Prompt model (SA^2VP). In contrast to the sequential modeling paradigm for visual prompting that learns an unordered sequence of prompts, our SA^2VP learns a 2D prompt map aligning with the image feature map spatially. Each prompt token is designated to prompt knowledge only for the spatially corresponding image tokens. As a result, our SA^2VP is able to conduct individual prompting for different image tokens in a fine-grained manner. Besides, it can capture spatial relations in the input image.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant NO. 62006060, U2013210, 62372133, 62176077), in part by the Guangdong Basic and Applied Basic Research Foundation under Grant (Grant NO. 2022A1515010306), in part by Shenzhen Fundamental Research Program (Grant NO. JCYJ20220818102415032), in part by the Shenzhen Key Technical Project (NO. 2022N001, 2020N046), in part by the Guangdong International Science and Technology Cooperation Project (NO. 20220505), and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (NO. 2022B1212010005).

References

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*.

Bar, A.; Gandelsman, Y.; Darrell, T.; Globerson, A.; and Efros, A. 2022. Visual prompting via image inpainting. *NeurIPS*, 35: 25005–25017.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *ECCV*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *CVPR*.

Das, R.; Dukler, Y.; Ravichandran, A.; and Swaminathan, A. 2023. Learning Expressive Prompting With Residuals for Vision Transformers. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

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*.

Gebru, T.; Krause, J.; Wang, Y.; Chen, D.; Deng, J.; and Fei-Fei, L. 2017. Fine-grained car detection for visual census estimation. In *AAAI*.

Han, C.; Wang, Q.; Cui, Y.; Cao, Z.; Wang, W.; Qi, S.; and Liu, D. 2023. Eˆ 2VPT: An Effective and Efficient Approach for Visual Prompt Tuning. *arXiv preprint arXiv:2307.13770*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

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 *ICML*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Huang, Q.; Dong, X.; Chen, D.; Zhang, W.; Wang, F.; Hua, G.; and Yu, N. 2023. Diversity-Aware Meta Visual Prompting. In *CVPR*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *ECCV*.

Karimi Mahabadi, R.; Henderson, J.; and Ruder, S. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *NeurIPS*, 34: 1022–1035.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR workshop*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2021a. GPT understands, too. *arXiv preprint arXiv:2103.10385*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; and Gurevych, I. 2020. AdapterFusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Rebuffi, S.-A.; Bilen, H.; and Vedaldi, A. 2017. Learning multiple visual domains with residual adapters. *NeurIPS*, 30.

Rücklé, A.; Geigle, G.; Glockner, M.; Beck, T.; Pfeiffer, J.; Reimers, N.; and Gurevych, I. 2020. Adapterdrop: On the efficiency of adapters in transformers. *arXiv preprint arXiv:2010.11918*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*.

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*.