Prompting Large-scale Vision Models with Cascaded Semantics

Anonymous authors
Paper under double-blind review

Abstract

As a leading parameter-efficient tuning paradigm in NLP, prompt tuning has recently been explored for its potential in computer vision. Unlike updating pre-trained large-scale models (e.g., vision transformer, or ViT for short), visual prompt tuning (VPT) incorporates additional learnable parameters (i.e., prompt) that are updated during tuning. However, original visual prompts are randomly initialized, without leveraging the power of prior knowledge, which has been frequently used in NLP (e.g., instruction). To bridge this gap, we propose a novel methodology, aiming to inject semantic prior to prompt the tuning. To this end, we pioneer in leveraging both fundamental image prior and advanced image semantics as such priors. The former, including color, texture, and shape, are extracted by classical hand-crafted operators, suitable for the input space, while the self-attention map is utilized as the latter, suitable for the feature space. We propose a scheme to integrate the two types of semantic priors into ViT's tuning through cascading. Extensive experiments conducted on 34 challenging image classification datasets demonstrate the superiority of our method in adapting pre-trained ViTs to various downstream scenarios while using only 0.74% of ViT parameters as tuned.

1 Introduction

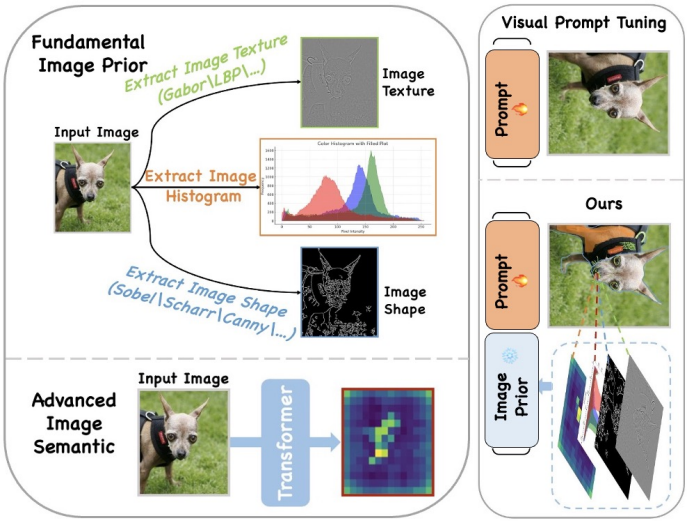

Figure 1: Injecting fundamental image prior (i.e., color, texture and shape) and advanced image semantics (i.e., self-attention map) as prompts into ViT to guide its fine-tuning. **For brevity, we simply use the term ViT to represent large-scale vision model. But in addition to ViT, we have also validated our method with other large-scale vision models.**

Adapting pre-trained large-scale vision models to downstream tasks through parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019; Li & Liang, 2021; Liu et al., 2025) has been shown to be practical, especially when the downstream data is limited. Visual prompt tuning (VPT) (Jia et al., 2022) is one of the most famous PEFT methods that does not resort to changing the structure or parameters of pre-trained vision transformers (ViTs) (Dosovitskiy et al., 2020), leading to a highly convenient pipeline. Unlike other methods, such as (Rebuffi et al., 2017; Hu et al., 2022) that update pre-trained parameters, and (Chen et al., 2022) that restructures the transformer block, VPT incorporates a small amount of learnable parameters (i.e., prompt) into the input and feature spaces of a ViT, and only updates them during fine-tuning with gradient descent. Such a paradigm yields surprisingly promising results in various scenarios, even exceeding the full fine-tuning that was deemed the most reliable tuning fashion (Houlsby et al., 2019; Han et al., 2024).

Though highly successful, ***two critical challenges*** naturally arise in current VPT research. **I. Random prompt initialization.** The first problem lies in the initialization of these visual learnable prompts: Current visual prompts are generally designed to be randomly initialized without considering any prior knowledge. In NLP, very differently, prompts are typically served as explicit instructions or context added to input text, guiding model's behavior based on previously seen textual information (Brown et al., 2020; Petroni et al., 2019). The absence of such a prior in the visual domain makes visual prompt more akin to black-box parameter optimization (i.e., tuning prompt) (Bahng et al., 2022), which could potentially lead to sub-optimal performance. **II. Lack of prompt explainability.** In NLP, both hard (not learnable) (Brown et al., 2020; Petroni et al., 2019; Schick & Schütze, 2020) and soft (learnable) (Li & Liang, 2021; Lester et al., 2021) prompts can be applied simultaneously. Although the soft ones may not be directly meaningful to humans, the hard ones are usually given by humans and are thus naturally interpretable. In VPT design, as these learnable prompts are directly updated via gradient descent, leaving little room for human interpretation (Bahng et al., 2022). Some research tries to involve intrepretability via attention distribution (Zeng et al., 2025) or *post-hoc* explanation (Wang et al., 2023) (e.g, GradCAM (Zeng et al., 2025; Han et al., 2023; 2024)). However, they stop at explaining visual prompts as a whole optimization objective after training, ignoring the characteristics of per-input variants or case-by-case instance awareness (Liu et al., 2025; Xiao et al., 2025).

Motivated by this, we explore incorporating interpretable prior knowledge during training as prompt into the vision tuning, which kills two birds with one stone: ① Replacing random tokens with prior-driven prompt tokens provides a structured warm start, anchoring optimizations in *semantically meaningful directions*, answering ***challenge I***. ② Since the fundamental image prior comes from human understandable information (e.g, color/texture/shape), their effects can be directly visualizable and quantifiable during training, leading to an advanced *ad-hoc explainable adaptation path* (Wang et al., 2023; Biehl et al., 2016; Swain & Ballard, 1991a; Manjunath & Ma, 1996; Dalal & Triggs, 2005), solving ***challenge II***.

Specifically, we consider prompt positions in both input and feature spaces of a ViT (Jia et al., 2022) by injecting different types of priors. For ***input space***, ideally, the prior should meet three criteria: a) it contains a certain level of image semantics, thus interpretable to humans; b) it is free of learning, thus bringing no extra burden to the tuning; c) it can be easily pinned to input space. We thus leverage fundamental image priors, including color histograms (Swain & Ballard, 1992), textures (Cimpoi et al., 2014), and shapes (Zhang & Lu, 2004), as the prompt in input space. The motivation is that these semantics are well-known for their ability to reflect essential image clues and can be obtained by simply using classical, hand-crafted operators (e.g, Sobel (Kanopoulos et al., 1988)) rather than learning (Swain & Ballard, 1991a; Manjunath & Ma, 1996; Dalal & Triggs, 2005) (see Sec. §3.2). For ***feature space***, the ideal prior is expected to: a) inject information critical to the model's decision; b) be based on the knowledge learned from preceding layers so it can facilitate the tuning of subsequent layers. To meet these requirements, we leverage the instance-aware, case-by-case self-attention map (Parmar et al., 2018; Chefer et al., 2021; Han et al., 2022; Khan et al., 2022) as the injected prompt in feature space. The rationale lies in the fact that such a map well reflects per-image class activation semantics (Zhou et al., 2016), and can be conveniently computed based on the features provided by transformer layers (see Sec. §3.3).

The following question turns into how to properly fuse the semantic prompts at different locations. We propose a simple yet effective scheme to cascade them. Specifically, for the prompt used in a specific transformer layer, it is formed by fusing the prompt and self-attention map in the preceding layer with that in the current layer, using skip connections (Huang et al., 2017; Srivastava et al., 2015; He et al., 2016; Oyedotun et al., 2022). As

such, the prompt in the current layer is nourished by the semantics learned from the preceding layers. Besides the semantic priors injected into the prompts, randomized learnable parameters are also utilized as part of the prompts, enabling gradient updates. A subsequent re-weighting adapter (see Sec. §3.4), demonstrated to be effective in (Kirichenko et al., 2022), is employed to enable flexible feature adaptation prior to the task head.

We conduct a wide range of experiments to evaluate our proposed method and observe superior results compared to current SOTAs. More interestingly, our experiments reveal that the semantic prompts can be more useful than text prompts in multimodal setting, indicating the significance of injecting image semantics as prompts. Overall, our contributions can be summarized as follows.

- Motivated by the interpretable text prompt in NLP, we revisit and enhance the visual prompt by equipping it with an explainable, instance-aware fundamental image prior and abstract advanced image semantics, enabling ad-hoc explainability into VPT design.
- We then develop an effective scheme to integrate these semantic prompts into our new fine-tuning paradigm, facilitating the fusion of prompts and features across layers.
- Extensive experiments on three widely adopted challenging benchmarks demonstrate the superiority of our proposed method over other PEFT solutions.

2 Related Works

2.1 Parameter-Efficient Fine-Tuning (PEFT)

Parameter-efficient fine-tuning (PEFT) has become a popular fashion for adapting pre-trained large-scale models with reduced computational demands and minimized over-fitting risk. Unlike full fine-tuning, PEFT only updates a small amount of parameters while freezing most of the pre-trained parameters. (Li & Liang, 2021; Jie & Deng, 2022; Chen et al., 2022; Dettmers et al., 2023; Karimi Mahabadi et al., 2021; Zaken et al., 2022) introduces learnable adapters (e.g, a light-weight convolutional network) into transformer layers, and updates the adapters during tuning. (Hu et al., 2022) and (Zhong et al., 2024) incorporate learnable rank decomposition matrices of parameters into transformer layers, significantly reducing the number of learnable parameters because of the low-rank decomposition. (Lian et al., 2022) updates the introduced parameters for scaling and shifting the features extracted by pre-trained model. This idea is well aligned with that in (Kirichenko et al., 2022), which performs tuning through re-weighting the features. However, most of these methods resort to the manipulation of transformer blocks to accommodate additional parameters, which inevitably increases the overall model complexity.

2.2 Visual Prompt Tuning (VPT)

Unlike the aforementioned methods that directly incorporate parameters into model, VPT prepends a small amount of learnable parameters (i.e., prompt) to input space (Bahng et al., 2022), and is then extended to cover both input and feature spaces (Jia et al., 2022), without changing the architecture of pre-trained model (e.g, ViT (Dosovitskiy et al., 2020)) (Wang et al., 2024; Yoo et al., 2023; Park & Byun, 2024; Li et al., 2024; Ren et al., 2025; Jin et al., 2025), greatly simplifying the paradigm of fine-tuning. (Han et al., 2023) develops two sets of different prompts, injected into input and parameter spaces (i.e., transformer layers) respectively, and uses an additional prompt pruning for better performance. (Pei et al., 2024) designs learnable prompt to model spatial relations in input image, and distinguish the prompts corresponding to different image tokens, achieving a fine-grained prompting. In addition, VPT has also been extended to multi-modal scenarios (Li et al., 2025; Huang et al., 2024). For example, (Khattak et al., 2023) learns visual prompt with the aid of textual prompt by jointly tuning the vision and text branches of CLIP (Radford et al., 2021), while (Zhou et al., 2022c) trains a light-weight network to jointly update visual and textual prompts. Nevertheless, all these methods only consider learnable prompt, which is equivalent to the soft prompt in NLP, while overlooking the hard prompt, which is not learnable and has been shown very effective in tuning large language model. However, very little work explores the hard one in computer vision. Motivated by bridging this gap, we investigate how to properly utilize both fundamental image prior and advanced image semantics as prompts for large-scale vision models (i.e., ViT, Swin).

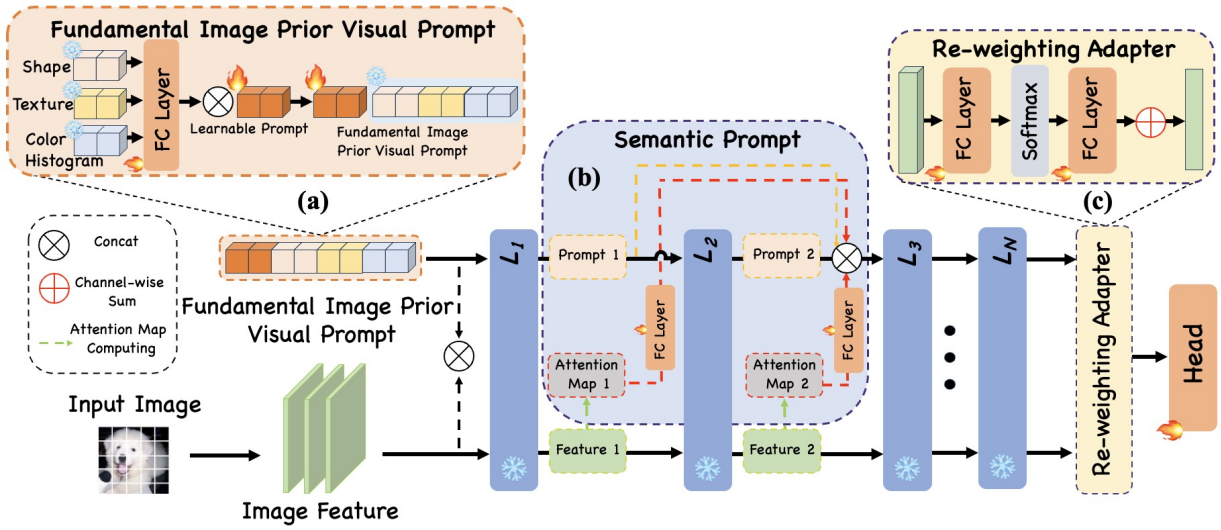

Figure 2: Overall architecture of our method. **(a) Fundamental Image Prior Visual Prompt:** hand-crafted priors (color histogram, texture, shape) are concatenated with a randomized learnable prompt and projected by a lightweight FC for token-size alignment. **(b) Re-Weighting Adapter:** a light two-layer linear module produces channel-wise weights to re-calibrate the final features before the head. **(c) Cascaded prompting in the backbone:** prompts are injected at selected layers $\{L_i\}$; the self-attention map from the previous layer is encoded and *cascaded* forward as a semantic prompt, forming a skip-cascade that integrates fundamental priors with advanced image semantics.

3 Methodology

3.1 Problem Definition

Given a pre-trained large-scale vision model (e.g, ViT (Dosovitskiy et al., 2021)) consisting of N transformer layers, we use $f_i(\cdot)$ to denote the feed-forwarding operation of the layer i, where $i \in 1, 2, \ldots, N$. Rather than updating the model's parameters, classical VPT (Jia et al., 2022) prepends a small amount of learnable parameters, namely P, to input X and its features across layers (i.e., *VPT-Deep*). As such, the first layer's output can be modeled as $Y_1 = f_1(P_1 \otimes X)$, while the i^{th} layer's output as $Y_i = f_i(P_i \otimes Y_{i-1})$, where $\otimes$ denotes concatenation. If $i = N$, then Y_i will be fed into a classification head h, yielding logits as $h(Y_i)$. During fine-tuning, only P and h are learnable by gradient descent. A special case of VPT, namely *VPT-Shallow*, only prepends P_1 to X, while waiving all the other P_is ($i \in 2, \ldots, N$) in feature space. It has been found that the deep version outperforms the shallow one, motivating us to develop our new method on top of *VPT-Deep*. The colors ■ and ■ indicate trainable and frozen parameters, respectively.

3.2 Visual Prior as Prompt for Input

In the classical VPT, all the prompts are randomly initialized. Though general, this brings no possibility of customizing or adjusting these prompts case-by-case, image-by-image. The loss of uniqueness brings performance degradation and non-transparent operations. Inspired by the hard prompts (unlearnable) in NLP, we aim to incorporate appropriate, instance-aware semantics to ViT in the form of hard prompt to solve these drawbacks. As prompts can be prepended to both input and feature spaces, we develop two schemes of injecting semantics for both spaces, respectively. Specifically, for *input space*, we leverage well-known hand crafted operators to extract *color histogram*, *texture*, and *shape* of the input image, as fundamental image priors, denoted by $\sigma_c(X)$, $\sigma_t(X)$, and $\sigma_s(X)$ respectively (e.g, σ_t as Gabor (Manjunath & Ma, 2002) and σ_s as Sobel (Kanopoulos et al., 1988)) (see Sec. §4.6 for more analysis). Then, we use the semantics as hard prompts, which are not learnable, to concatenate with the randomized learnable prompt P_1 in input space to form an overall prompt as

$$\tilde{P}_1 = P_1 \otimes FC(\sigma_c(X) \otimes \sigma_t(X) \otimes \sigma_s(X)), \tag{1}$$

where a learnable Linear layer (i.e., FC) is employed to adjust dimension as shown in Fig 2(a). As a result, the first transformer layer's output becomes

$$Y_1 = f_1(\tilde{P}_1 \otimes X). \tag{2}$$

For *feature space*, we conjecture that the fundamental image priors are likely to be sub-optimal options since they are directly from input rather than feature. Therefore, we compute self-attention map as the visual semantics in feature space. However, how to properly utilize such semantics as prompt remains unclear, introduced next.

3.3 Visual Semantics as Prompt for Features

The following question turns out to be a strength in the instance-aware information into prompts. To solve this, we leverage self-attention maps as semantically rich information to guide instance-aware prompt customizations. As illustrated in Fig. 2(c), for layer i, where $i \in \{2, ..., N\}$, we compute its self-attention map $\mathcal{A}_i$ based on its output feature, and concatenate it with P_i as $FC(\mathcal{A}_i) \otimes P_i$, where a learnable FC layer is used to adjust dimension. Meanwhile, this prompt is concatenated with the preceding prompt P_{i-1} and self-attention map $\mathcal{A}_{i-1}$ to form an overall prompt as

$$\tilde{P}_i = FC_i(\mathcal{A}_i) \otimes P_i \otimes FC_{i-1}(\mathcal{A}_{i-1}) \otimes P_{i-1}, \tag{3}$$

where both P_i and P_{i-1} are randomized learnable prompts. Then, the input to the layer $i+1$ can be written as $\tilde{P}_i \otimes Y_i$, where Y_i denotes the output feature of the layer i. This fashion of prompting does not apply to the first layer, in which we use the prompt introduced in Sec. §3.2. Notably, during fine-tuning, only the learnable prompts P_is, where $i \in \{1, ..., N\}$, and the FC layers are updated, without incurring much extra computing burden, thus maintaining the PEFT nature. The gradients are back-propagated through all the connections, including the skip ones.

3.4 Re-Weighting Adapter as the Final Puzzle

Our scope is not solely limited to prompt tuning engineering questions; instead, we are inspired by the influence function in classical statistics that properly re-weighting and/or perturbing data or features can lead to improved generalization of deep models(Koh & Liang, 2017). This explains why fine-tuning works, as tuning the last linear layer(s) can be considered as re-weighting the features learned by a pre-trained model (Kirichenko et al., 2022), while tuning all the layers as perturbing the features. In our method, if prompt tuning is treated as being used for perturbing features, then it still needs an equivalent operation for feature re-weighting. Motivated by this, we propose a simple but effective re-weighting adapter, as shown in Fig. 2(b). Here, the output feature of L_N is fed into a combination of *'FC-Softmax-FC'*, whose output is channel-wisely summed and then fed into the learnable classification head h.

3.5 Why Don't Leave Visual Prior to Learning?

In Sec. §3.2, we propose to inject the visual prior as prompt. Here, it is natural to raise a question: why don't we ask the model to learn such a prior automatically? In fact, if the model is trained from scratch, such prior could be better captured (Geirhos et al., 2018a). However, in the fine-tuning context, the model was pre-trained on source data that is different from target data, directly using the pre-trained model, which is frozen, might fail to effectively capture the fundamental image prior from target data (Ben-David et al., 2010; Torralba & Efros, 2011; Kornblith et al., 2018). Therefore, it is reasonable to capture the prior with simple hand-crafted operators (Swain & Ballard, 1991a; Manjunath & Ma, 1996; Dalal & Triggs, 2005), and then inject it as prompt to input space (Touvron et al., 2020).

3.6 Are More Parameters Beneficial?

In addition to the randomized learnable prompt, FC layers are also used for adjusting dimension, incorporating a few more learnable parameters. Then, it is necessary to investigate whether our method benefits from

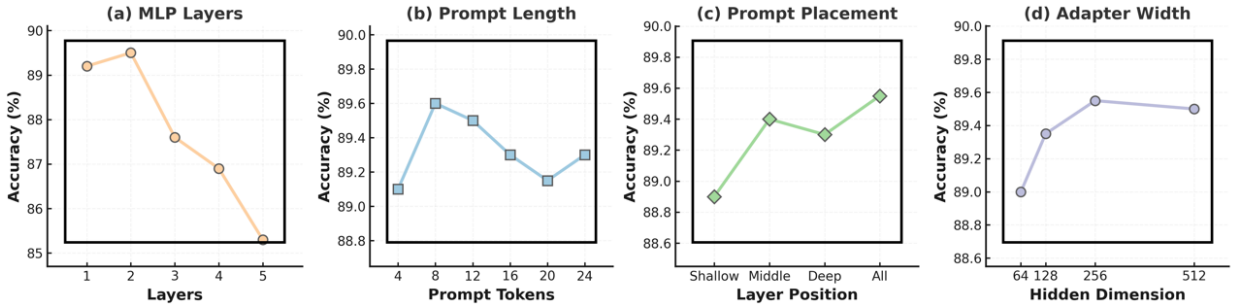

Figure 3: Ablation studies on CUB-200. (a) Replacing the single FC with deeper MLPs degrades accuracy as depth grows. (b) Increasing prompt length yields a non-monotonic trend (moderate length works best). (c) Targeted prompt placement outperforms indiscriminate/all-layer injection. (d) Enlarging the adapter width quickly saturates.

additional parameters. ***Counter-intuitively, we observe that more parameters hurt the performance.*** We conduct the following studies, summarized in Fig. 3. We replace the single FC layers in Fig. 2 with deeper MLPs, involving more parameters; as shown in Fig. 3(a), this change brings a negative impact as depth increases. We lengthen the randomized learnable prompts to include more parameters **(which has far exceeded the amount of parameters in our method)**, however, the performance is not *monotonically* improving, as shown in Fig. 3(b), aligning with the observation in (Jia et al., 2022). Moving prompts across depths shows that targeted placement is better than indiscriminate/all-layer injection, even though the latter uses more tokens/parameters; see Fig. 3(c). Increasing the hidden dimension of the re-weighting/adapter brings only marginal gains and quickly saturates; see Fig. 3(d). All experiments are conducted exclusively, suggesting that simply adding more parameters is not beneficial (Belkin et al., 2018; Nakkiran et al., 2019; Han et al., 2024); where and how to use them matters more (Wang et al., 2024; Chen et al., 2022).

4 Experiments

4.1 Experimental Setup

Datasets. Our method is evaluated on three diverse benchmarks: FGVC, HTA, and VTAB-1k (Zhai et al., 2019)—to test its adaptability and robustness across real-world scenarios. The **FGVC** benchmark includes five fine-grained datasets: CUB (Wah et al., 2011), NABirds (Van Horn et al., 2015), Oxford Flowers (Nilsback & Zisserman, 2008), Stanford Dogs (Khosla et al., 2011), and Stanford Cars (Gebru et al., 2017), assessing the model's ability to distinguish subtle variations among similar categories. We adhere to prior VPT study splits for consistency.(Jia et al., 2022) The **HTA** benchmark evaluates adaptability on 10 datasets, including CIFAR10 (Krizhevsky & Hinton, 2009), CIFAR100 (Krizhevsky & Hinton, 2009), DTD, CUB-200 (Wah et al., 2011), NABirds (Van Horn et al., 2015), Oxford Flowers (Nilsback & Zisserman, 2008), Food101, GTSRB, and SVHN, testing generalization across varied domains. We use DAM-VP (Huang et al., 2023a) setups for fair comparison. The **VTAB-1k** benchmark spans 19 datasets in three categories: 'Natural' (e.g, standard camera images), 'Specialized' (e.g, satellite, medical images), and 'Structured' (e.g, counting, distance tasks). Each dataset has 1000 images, split into 800 training and 200 validation images, to comprehensively test robustness across a wide array of visual tasks.

Implementation Details. Our experiments are primarily conducted using the ViT-B/16 model pre-trained on ImageNet-21K (Deng et al., 2009), consistent with previous VPT methodologies. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with an initial learning rate of $1e^{-3}$, a weight decay of $1e^{-4}$, and a batch size of either 64 or 128. Since our experiments focus on image classification, classification accuracy serves as the primary evaluation metric across all benchmarks.

Table 1: **Performance comparison of different fine-tuning strategies on ViT-Base/16.** The best are in **bold**, the second are underlined.

Methods	Tuned/Total (%)	Extra Params	FGVC (%)	HTA (%)	VTAB-1k Natural	VTAB-1k Specialized	VTAB-1k Structured	Mean Total (%)
Full (Iofinova et al., 2022)	100.00	—	88.54	85.8	75.88	83.36	47.64	65.57
Linear (Iofinova et al., 2022)	0.08	—	79.32	75.7	68.93	77.16	26.84	52.94
Partial-1 (Yosinski et al., 2014)	8.34	—	82.63	80.8	69.44	78.53	34.17	56.52
MLP-3 (Chen et al., 2020)	1.44	✓	79.80	78.5	67.80	72.83	30.62	53.21
Sidetune (Zhang et al., 2020)	10.08	—	78.35	72.3	58.21	68.12	23.41	45.65
Bias (Rebuffi et al., 2017)	0.80	—	88.41	82.1	73.30	78.25	44.09	62.05
Adapter (Cai et al., 2020)	1.02	✓	85.46	80.6	70.67	77.80	33.09	62.41
LoRA (Hu et al., 2022)	—	✓	89.46	85.5	78.26	83.78	56.20	72.25
AdaptFormer (Chen et al., 2022)	—	✓	—	—	80.56	84.88	58.83	72.32
ARC_{att} (Dong et al., 2023)	—	✓	89.12	89.0	80.41	85.55	58.38	72.32
VPT-S (Jia et al., 2022)	0.16	✓	84.62	85.5	76.81	79.66	46.98	64.85
VPT-D (Jia et al., 2022)	0.73	✓	89.11	85.5	78.48	82.43	54.98	69.43
E2VPT (Han et al., 2023)	0.39	✓	89.22	88.5	80.01	84.43	57.39	71.42
EXPRES (Das et al., 2023)	—	✓	—	—	79.69	84.03	54.99	70.02
DAM-VP (Huang et al., 2023b)	—	✓	—	88.5	—	—	—	—
SA^2VP (Pei et al., 2024)	0.81	✓	90.08	91.5	80.97	**85.73**	60.80	75.83
VFPT (Zeng et al., 2025)	0.66	✓	89.24	—	81.35	84.93	60.19	73.20
LoR-VP (Jin et al., 2025)	—	✓	89.32	—	79.91	83.16	60.01	74.36
Ours	0.74	✓	**90.20**	**91.7**	81.91	**85.83**	**61.16**	**76.30**

4.2 Comparison with State of the Art

Performance Comparison with ViT Backbone. Table 1 presents the results of different fine-tuning strategies on ViT-Base/16 across FGVC, HTA, and VTAB-1k. With only 0.74% of ViT parameters updated, our method attains 90.20% mean accuracy on FGVC and 91.7% on HTA, while achieving 81.91% / 85.83% / 61.16% on the Natural / Specialized / Structured VTAB-1k splits, respectively, leading to the best mean total score of 76.30% among all compared methods. In particular, the gains on animal FGVC datasets (CUB, NABirds, and Stanford Dogs) are consistent with our design: fundamental image priors such as textures and shapes provide informative hard prompts, while the dual-pathway skip connections facilitate the propagation of advanced semantic information through depth. At the same time, we observe a trend similar to (Han et al., 2024): as the dataset scale and variability increase (from FGVC to VTAB-1k), the relative advantages of prompt-based tuning gradually decrease, suggesting a limitation of prompt-only adaptation on highly diverse regimes, our semantic priors partially alleviate.

Table 2: **Performance comparison on VTAB-1k with Swin Transformer**. Best results in **bold**, the second are underlined.

Methods	Tuned/Total (%)	VTAB-1k Natural	VTAB-1k Specialized	VTAB-1k Structured
Full (Ren et al., 2023)	100.00	79.10	86.21	59.65
Linear (Ren et al., 2023)	0.06	73.52	80.77	33.52
Bias (Rebuffi et al., 2017)	0.30	76.78	83.33	51.85
VPT-deep (Jia et al., 2022)	0.25	76.78	83.33	51.85
E^2VPT (Han et al., 2023)	0.21	83.31	84.95	57.35
SA^2VP (Pei et al., 2024)	0.29	80.81	86.30	60.03
VFPT (Zeng et al., 2025)	0.27	84.53	86.15	58.21
LoR-VP (Jin et al., 2025)	0.29	83.51	85.22	57.61
Ours	0.28	**84.92**	**86.83**	**61.97**

Performance Comparison with ViT Backbone on HTA Benchmark. Table 1 illustrates the results of all the compared methods. Similar to the FGVC experiments, our method also shows consistent performance gains across a diverse group of datasets, indicating that the flexibility of our fundamental semantic priors is not limited to certain visual objects. On datasets with smaller image sizes and lower resolution, such as CIFAR-10/100 and SVHN, our method remains among the top 3, demonstrating its robustness across various imaging conditions.

Performance Comparison on VTAB-1k Benchmark with Swin Transformer Backbone. Table 2 presents the performance of various methods on the VTAB-1k benchmark using the *Swin Transformer* backbone, which is a large-scale vision model different from ViT. Our method achieves the good accuracy across all three task categories: Natural, Specialized, and Structured. As the Natural category is already analyzed in the first two experiments, we ignore further discussion here. In the Specialized category, including medical imaging and satellite data, our method attains a robust accuracy of 86.23%, which is the highest among the parameter-efficient tuning approaches. This strong performance is likely enhanced by our re-weighting adapter, which emphasizes relevant features, ensuring adaptability across highly specialized tasks. The most significant improvement, however, is observed in the Structured tasks, which often require an understanding of global geometric relationships and spatial dependencies. We speculate that our fundamental semantic priors which provide prompts of image-level statistics play a vital role in this kind of data.

Mean Performance of Different Methods on VTAB-1k Benchmark with ViT Backbone. Table 1 illustrates the performance of our method across the VTAB-1k benchmark categories: Natural, Specialized, and Structured tasks. The conclusion is similar to the experiment with Swin Transformer. An interesting phenomenon we notice is that, although the accuracy of full fine-tuning decreases significantly since ViT is less powerful than Swin ViT, our method maintains very close performance. This implies the potential of effective visual prompts, i.e., exhibiting low sensitivity to architecture changes. *Similarly, we observe a conclusion aligned with (Han et al., 2024), where VPT demonstrates stronger performance when there is a significant distribution shift between pretraining and downstream tasks, further validating its adaptability in cross-domain scenarios.*

Comparison with Text Prompt. In this work, as the semantics are shown useful to strengthen the randomized learnable visual prompt, here we aim to compare our semantic prompt with the text prompt that can also be used to benefit visual prompt in multi-modal settings (e.g, MaPLe (Khattak et al., 2023)). To this end, we perform three experiments. The first is the reproduction of MaPLe. In the second (MaPLeX), we disable the connection that feeds text prompt into the learning of visual prompt in MaPLe. The purpose is to investigate how text prompt will benefit the visual counterpart. Then, in the third experiment, we inherit the setting of the second one, in which text prompt is not injected into the vision branch, but inject semantic prompt into the vision branch, aiming to compare the effectiveness of text and our semantic prompts. As shown in Table 3, our method significantly outperforms the other two in the task of base-to-novel generalization (Zhou et al., 2022a;b). It achieves the highest accuracies of 96.17% and 72.92% in Base and Novel categories respectively, and a superior harmonic mean (HM) of 82.95. MaPLe is better than MaPLeX, showing that text prompt benefits visual prompt. Ours is better than MaPLe, demonstrating that semantic prompt could be more effective than text prompt.

Table 3: **Comparison between semantic and text prompts**.

Methods	Base (↑)	Novel (↑)	HM (↑)
MaPLe (Khattak et al., 2023)	95.62	72.03	82.17
MaPLeX (Khattak et al., 2023)	94.87 (-0.75)	70.11 (-1.92)	80.63 (-1.54)
Ours	**96.17** (+0.55)	**72.92** (+0.89)	**82.95** (+0.78)

4.3 Interpretability and Feature Representation

In this section, we evaluate the effectiveness of our method in feature representation learning. This section aims to verify our ad-hoc design of semantic integration can help post-hoc measurements, including cosine similarity, IoU analysis, and t-SNE analysis.

Cosine Similarity Analysis. We first investigate whether the learned features well match image clues. We adopt the cosine similarity map (Vaswani, 2017) to directly demonstrate the relative distances (Wang et al., 2023; Steck et al., 2024). As illustrated in Fig. 4, without the aid of semantic prompt, the similarity between the learned prompt and features is much lower (i.e, **middle**), indicating a higher degree of mismatch, while the opposite is observed (i.e., **right**) when semantic prompt is leveraged, indicating that such prompt benefits feature learning.

Table 4: **Comparison of IoU on CUB-200**.

Methods	Mean IoU (↑)	Median IoU (↑)
VPT (Jia et al., 2022)	26.5	27.0
Ours	**32.9** (+6.4)	**33.1** (+6.1)

IoU Analysis. We further investigate how learned features benefit the localization of target objects, where we use the images from CUB-200 as test cases. We adopt the Intersection over Union (IoU) metric (Everingham et al., 2010), which measures the overlap between the model's focused objects. It can be visualized with attention map, and ground truth target objects given by bounding boxes, with a higher value indicating a better localization. As shown in Table 4, with the aid of semantic prompt, our method surpasses VPT in terms of both mean and median IoUs, indicating more accurate attention localization and stable performance. We refer readers to the *Supplementary Material* for more detailed IoU analysis.

GradCAM Analysis. Here, we use another tool, namely GradCAM (Selvaraju et al., 2017) to further investigate how semantic prompt improves model's attention. GradCAM highlights the regions, to which the model pays attention when performing image classification. Fig. 5 shows that our method focuses on the most discriminative object parts, such as the bird's head, explaining why our method yields superior classification performance, well aligned with human perception.

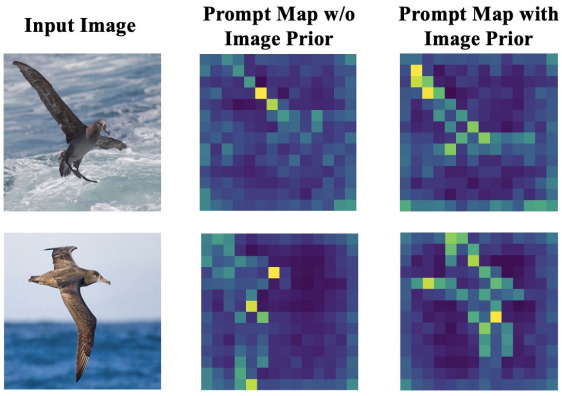

Figure 4: Comparison of the features learned with (**right**) and without (**middle**) semantic prior, respectively, using cosine similarity map (Vaswani, 2017).

t-SNE Analysis. We adopt t-SNE (Van der Maaten & Hinton, 2008) for a more intuitive, feature-level examination of the clustering results on the Sun397 dataset with four fine-tuning methods: Head tuning, AdaptFormer, VPT, and our proposed method. As shown in Fig. 6, our method achieves highly distinct and compact clusters with minimal overlap, underscoring its superior capacity to learn discriminative feature representations. This result demonstrates the effectiveness of our semantic prompts in capturing class-specific features and distinguishing complex visual patterns, providing a significant advantage in feature clarity and separability over competing methods.

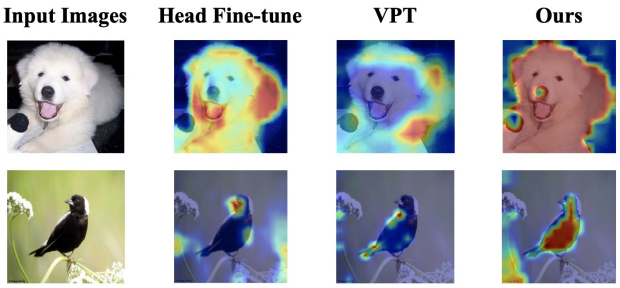

Figure 5: GradCAM (Selvaraju et al., 2017) visualization of the final layer features obtained by different methods for two randomly selected images.

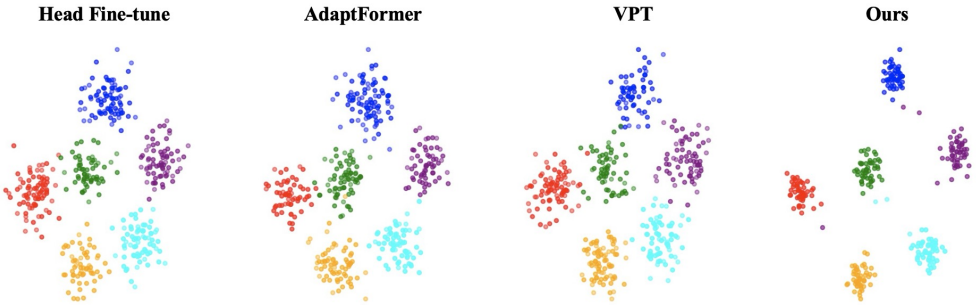

Figure 6: t-SNE (Van der Maaten & Hinton, 2008) results of the learned features in the last layer of the model by four different methods on Sun397.

4.4 Understanding with Information Theory

Here, we provide another perspective to understand why the fundamental image prior visual prompts work. Inspired by the theory of Information Bottleneck (IB) (TISHBY, 2000), we attribute the success of our representation learning to the higher correlation with labels Y when compressing the input X. IB provides a theoretical framework for understanding how deep learning models learn and generalize (Tishby & Zaslavsky, 2015; Saxe et al., 2019). It suggests that the learning process of deep models is to compress input data X into representations T that retain only the information necessary to predict the label Y. Therefore, training a deep model is expected to have the same effect as minimizing the IB as

$$\mathcal{L}_{\mathrm{IB}} = I(X;T) - \beta I(T;Y), \tag{4}$$

where I refers to mutual information. Using Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018), we analyzed the 12 Transformer layers for the baseline VPT and our method. As shown in Fig. 7, our method give a significantly lower IB on top layers (close to the output), indicating its successful representation learning. Intuitively, $I(X;T)$ should not be affected since our proposed fundamental image prior visual prompts are designed to extract inherent characteristics from the input images themselves rather than external sources. However, those semantic prompts offer increased opportunities for the learned representations to better correlate labels for unseen data, and therefore improve $I(Y;T)$. The mutual information curves in Fig. 7 well align those hypotheses of our method.

4.5 Ablation Study

Importance of Each Component. As the proposed method consists of multiple components, including the different types of semantics, re-weighting adapter, and the skip connections for cascading semantics, we conduct an ablation study here to validate the efficacy of each component by detaching it from the whole pipeline and checking how performance will vary. As shown in Table 5, excluding each of the semantics (i.e.,

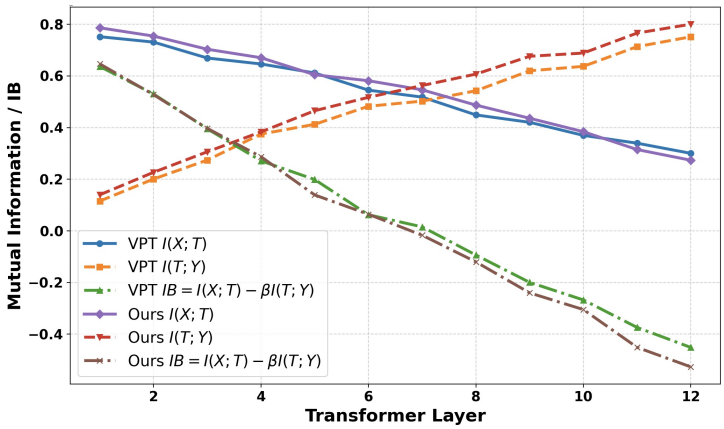

Figure 7: **Information Bottleneck and Mutual Information** between *feature* and *label* across transformer layers on CUB-200. β is set to 1 following common practices.

Table 5: **Ablation study on VTAB-1k**. We analyze the impact of removing components.

Ablated Variants	Natural	Specialized	Structured
Section 1: Single Component Removal			
w/o C (*Color Histogram*)	81.65 (-0.26)	85.49 (-0.34)	60.84 (-0.32)
w/o T (*Texture*)	81.59 (-0.32)	85.41 (-0.42)	60.75 (-0.41)
w/o S (*Shape*)	81.55 (-0.36)	85.36 (-0.47)	60.69 (-0.47)
w/o A (*Self-Attention*)	81.33 (-0.58)	85.21 (-0.62)	60.51 (-0.65)
w/o R (*Re-Weighting*)	81.20 (-0.71)	85.09 (-0.74)	60.38 (-0.78)
w/o K (*Skip-Connection*)	80.92 (-0.99)	84.89 (-1.05)	60.02 (-1.14)
Section 2: Cumulative Component Removal			
Remove C, T	81.20 (-0.71)	85.00 (-0.83)	60.40 (-0.76)
Remove C, S	80.91 (-1.00)	84.78 (-1.05)	60.10 (-1.06)
Remove T, S	80.50 (-1.41)	84.39 (-1.44)	59.68 (-1.48)
Remove A, R	79.70 (-2.21)	83.60 (-2.23)	59.00 (-2.16)
Remove K, C	78.95 (-2.96)	82.90 (-2.93)	58.45 (-2.71)
Remove K, A	77.80 (-4.11)	81.95 (-3.88)	57.60 (-3.56)
Remove K, A, R	76.70 (-5.21)	80.90 (-4.93)	56.70 (-4.46)
Baseline (None)	75.80 (-6.11)	79.80 (-6.03)	55.50 (-5.66)
Full Model (All)	**81.91**	**85.83**	**61.16**

color, texture, shape, and self-attention map) will to an extent lower the performance across all the categories (i.e., natural, specialized, and structured) of VTAB-1k, demonstrating the necessity of injecting such semantic prompts. For the operational components, namely re-weighting adapter and skip connections, detaching the former will also result in a declined performance. More importantly, detaching the latter will significantly deteriorate the performance, showing that *cascading the semantics* plays a more critical role in our method.

4.6 Fundamental Image Prior Operators

Handcrafted image priors have played a crucial role in classical computer vision and deep learning, serving as *ad-hoc*, human-understandable complementary cues to learned representations (Nanni et al., 2017; Zhang & Zhang, 2021; Tianyu et al., 2018). Prior studies have explored integrating handcrafted features into deep learning models, typically in a feature fusion manner within CNN architectures. However, their exploration in prompt-based tuning remains largely limited. Recent work on Conceptual Codebook Learning (CoCoLe) (Zhang et al., 2024) demonstrates the potential of incorporating structured prior knowledge into model tuning. CoCoLe introduces a learnable conceptual codebook that maps visual concepts to textual prompts, effectively bridging vision and language representations. While CoCoLe focuses on modality alignment and conceptual-level adaptation, our approach is fundamentally different: we integrate fixed, non-learnable handcrafted features directly into VPT. ***Rather than constructing a learnable codebook, we employ well-established***

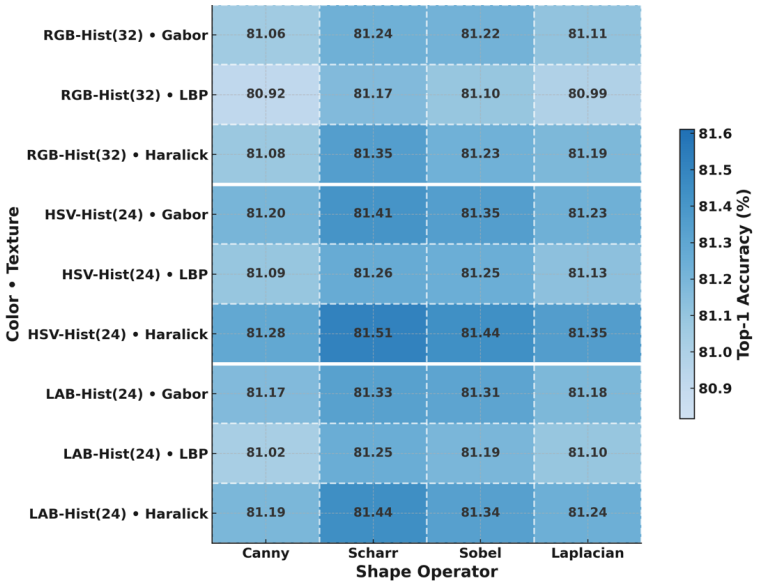

Figure 8: **Performance comparison** of using different shape and texture operators, with the color operator fixed.

handcrafted operators as hard prompts to improve vision transformer adaptation, leveraging their domain-invariant and human understandable properties. Specifically, we use color, texture, and shape priors as additional prior to enhance model robustness. Color histograms (Swain & Ballard, 1992) capture chromatic distributions, texture descriptors (e.g, Gabor filters (Manjunath & Ma, 2002), LBP (Ojala et al., 2002)) encode spatial intensity variations, and edge-based operators (e.g, Sobel (Kanopoulos et al., 1988), Scharr (Scharr, 2004), Canny (Canny, 2009)) extract structural information. While previous work has explored learning handcrafted feature representations (Zhang & Zhang, 2021), our approach keeps them fixed and directly integrates them as prompts, ensuring interpretability and computational efficiency. A key question is: ***which operators should be chosen?*** To investigate this, we conduct an experiment on a random subset of CIFAR-100, using part of the subset for tuning and the rest for testing. The rationale for using a subset instead of the full dataset is twofold: (1) The operators are fixed and non-learnable, so dataset size does not affect their representation power. (2) A smaller subset is computationally efficient for evaluation. As shown in Fig. 8, when the color histogram is fixed, variations in texture (Gabor vs. LBP) and shape (Canny, Scharr, Sobel) yield comparable results. This suggests that our semantic prompt strategy is robust across different operator choices, reinforcing the generalizability of our approach. Unlike prior work that injects handcrafted features into CNNs (Zhang & Zhang, 2021) or employs learnable conceptual mappings in multimodal settings (Zhang et al., 2024), our work introduces a novel use of handcrafted priors as prompts, bridging classical vision priors with transformer-based adaptation in a lightweight and interpretable manner.

5 Conclusion

In this work, we demonstrate that properly injecting fundamental semantics, such as color, texture, and shape, as well as ad-hoc explainable semantics, such as self-attention information, yields a new fine-tuning paradigm for large-scale vision models. Adhering to the nature of parameter-efficient tuning, we propose a cascaded fashion to integrate the two types of semantics as part of the prompts in both input and feature spaces to guide the randomized initiated prompts. Extensive evaluations have shown the superiority and reliability of our method on various benchmarks. Moreover, the semantic prompts have also been found more useful than text prompts in a certain multi-modal setting, indicating the significance of the semantic prompts.

References

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Mikhail Belkin, Daniel J. Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116: 15849 – 15854, 2018. URL `https://api.semanticscholar.org/CorpusID:198496504`.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando C Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Michael Biehl, Barbara Hammer, and Thomas Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. *Advances in Neural Information Processing Systems*, 33:11285–11297, 2020.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 2009.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021.

Shoufa Chen, Chuang Ge, Zhiqiang Tong, Jianzhuang Wang, Yang Song, Jianmin Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Han Cheng, Wang Qifan, Cui Yiming, Cao Zhiwen, Wang Wenguan, Qi Siyuan, and Liu Dongfang. E2vpt: An effective and efficient approach for visual prompt tuning. In *International Conference on Computer Vision (ICCV)*, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:886–893 vol. 1, 2005.

Rahul Das, Yahel Dukler, Avinash Ravichandran, and Ajay Swaminathan. Learning expressive prompting with residuals for vision transformers. In *CVPR*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Wei Dong, Dawei Yan, Zhijun Lin, and Peng Wang. Efficient adaptation of large vision transformer via adapter re-composing. *Advances in Neural Information Processing Systems*, 36:52548–52567, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI Conference on Artificial Intelligence*, 2017.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv*, abs/1811.12231, 2018a.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018b.

Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. E2vpt: An effective and efficient approach for visual prompt tuning. *arXiv preprint arXiv:2307.13770*, 2023.

Cheng Han, Qifan Wang, Yiming Cui, Wenguan Wang, Lifu Huang, and Dongfang Liu. Facing the elephant in the room: Visual prompt tuning or full finetuning? In *International Conference on Learning Representations (ICLR)*, 2024.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Neil Houlsby et al. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, S Wang, L Wang, and W Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Mingzhen Huang, Jingru Zhang, Xiaodan Liang, and Hao Wang. Dam-vp: Adaptive meta-learning for visual prompt tuning in domain-adaptive vision transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023a.

Qingyao Huang, Xingchen Dong, Dantong Chen, Weiwei Zhang, Fuzhen Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *CVPR*, 2023b.

Zhenhan Huang, Tejaswini Pedapati, Pin-Yu Chen, and Jianxi Gao. Differentiable prompt learning for vision language models. In *IJCAI*, 2024.

Eugenia Iofinova, Alexandra Peste, Mark Kurtz, and Dan Alistarh. How well do sparse imagenet models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12276, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022.

Can Jin, Ying Li, Mingyu Zhao, Shiyu Zhao, Zhenting Wang, Xiaoxiao He, Ligong Han, Tong Che, and Dimitris N. Metaxas. Lor-vp: Low-rank visual prompting for efficient vision model adaptation. In *International Conference on Learning Representations (ICLR)*, 2025.

Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *NeurIPS*, volume 34, pp. 1022–1035, 2021.

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

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 109–116, 2011.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2018.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. In *Technical report, University of Toronto*, 2009.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Mengke Li, Ye Liu, Yang Lu, Yiqun Zhang, Yiu ming Cheung, and Hui Huang. Improving visual prompt tuning by gaussian neighborhood minimization for long-tailed visual recognition. *ArXiv*, abs/2410.21042, 2024.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, and Jian Yang. Advancing textual prompt learning with anchored attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3618–3627, 2025.

Danyang Lian, Dongze Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *NeurIPS*, volume 35, pp. 109–123, 2022.

Yiyang Liu, James C Liang, Heng Fan, Wenhao Yang, Yiming Cui, Xiaotian Han, Lifu Huang, Dongfang Liu, Qifan Wang, and Cheng Han. All you need is one: Capsule prompt tuning with a single vector. *arXiv preprint arXiv:2510.16670*, 2025.

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

B. S. Manjunath and Wei-Ying Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18:837–842, 1996.

Bangalore S Manjunath and Wei-Ying Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8):837–842, 2002.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021, 2019. URL `https://api.semanticscholar.org/CorpusID:207808916`.

Loris Nanni, Stefano Ghidoni, and Sheryl Brahnam. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71:158–172, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2017.05.025. URL `https://www.sciencedirect.com/science/article/pii/S0031320317302224`.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

Oyebade K Oyedotun, Kassem Al Ismaeil, and Djamila Aouada. Why is everyone training very deep neural network with skip connections? *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 5961–5975, 2022.

Sungho Park and Hyeran Byun. Fair-vpt: Fair visual prompt tuning for image classification. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12268–12278, 2024.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pp. 4055–4064. PMLR, 2018.

Wenjie Pei, Tongqi Xia, Fanglin Chen, Jinsong Li, Jiandong Tian, and Guangming Lu. Sa^2vp: Spatially aligned-and-adapted visual prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4450–4458, 2024.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Artem Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

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

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, volume 30, 2017.

Li Ren, Chen Chen, Liqiang Wang, and Kien Hua. Da-vpt: Semantic-guided visual prompt tuning for vision transformers. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4363, 2025.

Yi Ren, Shangmin Guo, Wonho Bae, and Danica J. Sutherland. How to prepare your task head for finetuning. In *The Eleventh International Conference on Learning Representations*, 2023.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Hanno Scharr. Optimal filters for extended optical flow. In *International Workshop on Complex Motion*, pp. 14–29. Springer, 2004.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2377–2385, 2015.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, pp. 887–890, 2024.

Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991a.

Michael J Swain and Dana H Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991b.

Michael J Swain and Dana H Ballard. Indexing via color histograms. In *Active perception and robot vision*, pp. 261–273. Springer, 1992.

Zhou Tianyu, Miao Zhenjiang, and Zhang Jianhu. Combining cnn with hand-crafted features for image classification. In *2018 14th ieee international conference on signal processing (icsp)*, pp. 554–557. IEEE, 2018.

N TISHBY. The information bottleneck method. *Computing Research Repository (CoRR)*, 2000.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. *CVPR 2011*, pp. 1521–1528, 2011.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. volume 9, 2008.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

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

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. In *California Institute of Technology*, 2011.

Wenguan Wang, Cheng Han, Tianfei Zhou, and Dongfang Liu. Visual recognition with deep nearest centroids. In *ICLR*, 2023.

Yuzhu Wang, Lechao Cheng, Chaowei Fang, Dingwen Zhang, Manni Duan, and Meng Wang. Revisiting the power of prompt for visual tuning. In *ICML*, 2024.

Xi Xiao, Yunbei Zhang, Xingjian Li, Tianyang Wang, Xiao Wang, Yuxiang Wei, Jihun Hamm, and Min Xu. Visual instance-aware prompt tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2880–2889, 2025.

Seung Woo Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sung-Hoon Yoon. Improving visual prompt tuning for self-supervised vision transformers. In *International Conference on Machine Learning*, 2023.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

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

Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huangg, Ying Nian Wu, and Dongfang Liu. Visual fourier prompt tuning. *Advances in Neural Information Processing Systems*, 37:5552–5585, 2025.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. The visual task adaptation benchmark (vtab). *arXiv preprint arXiv:1910.04867*, 2019.

Dengsheng Zhang and Guojun Lu. Review of shape representation and description techniques. *Pattern recognition*, 37(1):1–19, 2004.

Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 698–714. Springer, 2020.

Tianwen Zhang and Xiaoling Zhang. Injection of traditional hand-crafted features into modern cnn-based models for sar ship classification: What, why, where, and how. *Remote Sensing*, 13(11), 2021. ISSN 2072-4292. doi: 10.3390/rs13112091. URL `https://www.mdpi.com/2072-4292/13/11/2091`.

Yi Zhang, Ke Yu, Siqi Wu, and Zhihai He. Conceptual codebook learning for vision-language models. In *European Conference on Computer Vision*, pp. 235–251. Springer, 2024.

Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora: Parameter efficient finetuning for segment anything model. In *International Conference on Learning Representations*, 2024.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022c.

Supplementary Material Outline

This supplementary material provides additional experimental results, in-depth efficiency analyses, interpretability studies, and implementation details to support the main paper. The content is organized as follows:

- **Section 6: Extended Benchmark Results.** We provide a comprehensive analysis of the VTAB-1k results, including performance on Natural, Specialized, and Structured sub-categories. Additionally, we expand on the Fine-Grained Visual Classification (FGVC) and Hierarchical Transfer Adaptation (HTA) benchmarks.
- **Section 7: Efficiency and Computational Cost Analysis.** We detail the "Extract-Once" strategy that ensures zero training overhead for prior extraction. We also analyze inference latency (demonstrating a marginal $<$ 2ms increase) and the significant reduction in memory footprint compared to full fine-tuning.
- **Section 8: Advanced Localization Analysis (IoU).** We present a rigorous IoU evaluation on the CUB-200 dataset, breaking down performance across Easy, Medium, and Hard subsets to demonstrate our method's robustness against occlusion and background complexity.
- **Section 9: Extended Mutual Information Analysis.** We delve deeper into the Information Bottleneck (IB) framework, providing experimental evidence of how semantic prompts improve label correlation ($I(T;Y)$) in deeper transformer layers.
- **Section 10: Discussion on Hand-Crafted vs. Deep Priors.** We provide a theoretical discussion justifying the choice of hand-crafted operators over deep-learned features, emphasizing information orthogonality, domain robustness, and strict efficiency.
- **Section 11: Implementation Details of Fundamental Operators.** We detail the specific configurations for Color (HSV Histograms), Texture (Gabor Filters), and Shape (Sobel Operators) priors. We also present ablation studies on operator variants (e.g., LBP, Canny) to verify robustness.
- **Section 12: Limitations and Future Work.** We discuss current limitations regarding fixed operators and geometric reasoning tasks, and propose future directions such as adaptive prior mechanisms.

A VTAB-1k Benchmark Results

The VTAB-1k benchmark provides a comprehensive evaluation of various methods across diverse datasets, highlighting their strengths and weaknesses in handling Natural, Specialized, and Structured tasks. The performance of different methods based on the ViT backbone is summarized in Table 6. Our method achieves consistently strong results across all categories, outperforming other fine-tuning techniques such as Head Fine-tune, AdaptFormer, and VPT-deep. This section delves deeper into the analysis of these results, providing insights into the strengths and areas for improvement of our approach.

A.1 Strengths in Natural and Specialized Categories

Our method demonstrates significant improvements in Natural datasets, such as CIFAR-100 (79.2%) and Caltech101 (92.3%), showcasing its ability to handle fine-grained classification tasks effectively. These datasets often involve subtle intra-class variations, which our method addresses by integrating hierarchical features, such as textures and shapes, with semantic prompts. Similarly, for Specialized datasets like Patch Camelyon (87.2%) and Resisc45 (86.4%), the results validate the importance of domain-specific priors in extracting meaningful features, outperforming AdaptFormer and LoRA.

Table 6: Performance of different methods on the VTAB-1k benchmark based on ViT backbone. The best results are highlighted in **bold** to showcase the most effective methodology. Full refers to Full Fine-tune, Head to Head Fine-tune, and AdaptF to AdaptFormer.

Datasets	Full	Head	AdaptF	LoRA	VPT-deep	ExPRes	E^2VPT	**Ours**
CIFAR-100	68.9	63.4	70.8	67.1	78.8	78.0	78.6	**79.2**
Caltech101	87.7	85.0	91.2	91.4	90.8	89.6	89.4	**92.3**
DTD	64.3	63.2	70.5	69.4	65.8	68.8	67.8	**71.4**
Flowers102	97.2	97.0	**99.1**	98.8	98.0	98.7	98.2	98.9
Pets	86.9	86.3	90.9	90.4	88.3	88.9	88.5	**91.7**
SVHN	**87.4**	36.6	86.6	85.3	78.1	81.9	85.3	85.8
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	**81.91**
Patch Camelyon	79.7	78.5	83.0	84.9	81.8	84.8	82.5	**87.2**
EuroSAT	95.7	87.5	95.8	95.3	96.1	96.2	**96.8**	95.2
Resisc45	84.2	68.6	84.4	83.4	83.4	80.9	84.8	**86.4**
Retinopathy	73.9	74.0	**76.3**	73.6	68.4	74.2	73.6	74.5
Mean	83.36	77.16	84.88	84.55	82.43	84.03	84.43	**85.83**
Clevr/count	56.3	34.3	81.9	**82.9**	68.5	66.5	71.7	78.1
Clevr/distance	58.6	30.6	64.3	**69.2**	60.0	60.4	61.2	62.2
DMLab	41.7	33.2	49.3	49.8	46.5	46.5	47.9	**53.2**
KITTI/distance	65.5	55.4	**80.3**	78.5	72.8	77.6	75.8	78.5
dSprites/location	57.5	12.5	76.3	75.7	73.6	78.0	80.8	**84.1**
dSprites/orientation	46.7	20.0	45.7	47.1	47.9	49.5	48.1	**53.4**
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	**45.9**
Mean	47.64	26.84	58.83	59.78	54.98	54.99	57.39	**61.16**

A.2 Challenges in Structured Tasks

While achieving state-of-the-art performance in tasks like KITTI/distance (80.3%), challenges remain in datasets such as SmallNORB/azimuth (34.7%). These datasets require intricate spatial reasoning, which may benefit from further refinements in spatial encoding mechanisms, suggesting a potential avenue for future research.

A.3 Generalization Insights

The overall mean performance across all categories (81.91%) underscores the robustness of our method. Importantly, the smaller training-testing gap compared to other methods highlights its superior generalization capability. This performance, coupled with reduced overfitting, reaffirms the effectiveness of incorporating both low- and high-level image priors into our approach. Future studies may explore additional spatial and temporal features to address current limitations, further enhancing model adaptability across diverse tasks.

A.4 Performance on FGVC and HTA Benchmarks

To further evaluate the effectiveness of our method, we compare its performance on the Fine-Grained Visual Classification (FGVC) and Hierarchical Transfer Adaptation (HTA) benchmarks. FGVC involves tasks requiring fine-grained distinctions between categories, while HTA assesses hierarchical knowledge transfer across multiple domains. The results in Tables 7 and 8 demonstrate that our method consistently outperforms existing fine-tuning approaches.

Table 7: **Performance comparison on the FGVC benchmark with ViT.**

Methods	CUB-200-2011	NABirds	Oxford Flowers	Stanford Dogs	Stanford Cars	Mean
Full Fine-tune	87.3	82.7	98.8	89.4	**84.5**	88.54
AdaptFormer (Chen et al., 2022)	84.7	75.2	97.9	84.7	83.1	85.12
LoRA (Hu et al., 2022)	84.9	79.0	98.1	88.1	79.8	85.98
VPT-shallow (Jia et al., 2022)	86.7	78.8	98.4	90.7	68.7	84.62
VPT-deep (Jia et al., 2022)	88.5	84.2	99.0	90.2	83.6	89.11
E^2VPT (Cheng et al., 2023)	89.1	84.6	**99.1**	90.5	82.8	89.22
Ours	**89.7**	**85.5**	**99.1**	**92.6**	84.1	**90.2**

Table 8: **Performance comparison on the HTA benchmark with ViT.**

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 (Houlsby et al., 2019)	62.7	87.1	84.3	89.8	98.5	86.0	74.2	97.7	91.1	36.3	80.8
VPT-deep (Jia et al., 2022)	65.8	88.5	84.2	90.2	99.0	83.3	78.8	96.8	90.7	78.1	85.5
AdaptFormer (Chen et al., 2022)	74.4	84.7	75.2	84.7	97.9	89.1	**91.4**	**98.8**	97.0	**96.5**	89.0
DAM-VP (Huang et al., 2023b)	73.1	87.5	82.1	92.3	**99.2**	86.9	86.9	90.6	87.9	88.1	88.5
Ours	**76.3**	**89.7**	**85.5**	**92.6**	99.1	**92.3**	90.9	98.1	96.5	96.1	**91.7**

A.5 Analysis of FGVC and HTA Performance

Our method achieves the best performance across various fine-grained classification tasks in FGVC, particularly on CUB-200-2011 (89.7%) and Stanford Dogs (92.6%), where distinguishing similar-looking categories is crucial. This demonstrates the effectiveness of our approach in capturing nuanced visual patterns.

For the HTA benchmark, which evaluates hierarchical transfer adaptation, our method outperforms others in generalization ability, with an overall mean accuracy of 91.7%. The high scores across datasets such as NABirds (85.5%) and GTSRB (96.5%) validate its robustness in learning transferable knowledge across hierarchical tasks. The significant improvements highlight the importance of leveraging both handcrafted priors and learnable prompts to enhance representation learning.

These results further reinforce our findings that integrating structured visual priors into prompt tuning enhances both fine-grained classification and hierarchical adaptation, making our approach a strong alternative to traditional fine-tuning strategies.

B Efficiency and Computational Cost Analysis

Although our method introduces additional modules to incorporate fundamental image priors and cascaded semantics, we maintain a high degree of computational and memory efficiency. In this section, we analyze the efficiency of our approach from the perspectives of training overhead, inference latency, and memory consumption.

B.1 Training Efficiency: The "Extract-Once" Strategy

A critical design advantage of our **Fundamental Image Prior Visual Prompt** is that the operators used for extraction—Color Histograms, Texture (e.g., Gabor, LBP), and Shape (e.g., Sobel, Canny)—are entirely **hand-crafted and non-learnable**.

This property decouples the prior extraction from the model's gradient optimization loop. Consequently, these priors do not need to be re-computed at every training epoch. Instead, we adopt an "Extract-Once" strategy:

- **Offline/Pre-computation:** The fundamental priors can be computed once offline during data preparation or online via CPU worker threads in the dataloader pipeline. Since these operations rely solely on the fixed input image X and not on the model parameters, they introduce **zero additional overhead** to the GPU training time.

- **Frozen Backbone:** As our method freezes the ViT backbone and only updates the prompt parameters and the lightweight re-weighting adapter, we avoid the heavy backward pass computations associated with Full Fine-tuning.

B.2 Inference Latency and Complexity

During inference, the fundamental priors must be computed for each input. However, standard computer vision operators are computationally negligible compared to the heavy matrix multiplications in the Vision Transformer backbone.

- **Low Computational Complexity:** The complexity of extracting these priors is generally linear with respect to image pixels ($\mathcal{O}(HW)$), whereas the Multi-Head Self-Attention (MSA) mechanism in ViT scales quadratically with token sequence length ($\mathcal{O}(N^2)$).
- **Lightweight Modules:** The trainable components (Prompt Embeddings, Linear Projections, and Re-weighting Adapter) are extremely lightweight. Specifically, the introduced Linear layers for dimension alignment and the Re-weighting Adapter operate on low-dimensional feature vectors, adding minimal FLOPs.

Empirically, on a single NVIDIA A100 GPU, our method incurs a marginal latency increase ($<$ 2ms per image) compared to the standard VPT, while significantly outperforming Full Fine-tuning in throughput.

B.3 Memory Footprint

Our method updates only **0.74%** of the total parameters (approximately 0.6M parameters for ViT-B/16). This results in a drastic reduction in GPU memory usage compared to Full Fine-tuning, as we do not need to store optimizer states (e.g., momentum and variance in AdamW) for the vast majority of the backbone parameters. This allows for larger batch sizes or deployment on edge devices with limited VRAM, making our approach highly practical for real-world applications.

C More IoU Analysis

C.1 Experiment Setup

To rigorously evaluate the localization capabilities of different methods, we employ the Intersection over Union (IoU) metric on the CUB-200 dataset. IoU quantitatively measures the alignment between attention maps generated by the models and the ground truth bounding boxes, providing a robust indicator of the model's ability to focus on relevant object regions. Higher IoU values reflect superior localization performance. The experimental setup is as follows: Attention Map Extraction: Attention maps from the final Vision Transformer layer are normalized to emphasize regions with higher attention scores. Thresholding for Binary Maps: Binary masks are generated by applying intensity thresholds to the normalized attention maps, ensuring alignment with the ground truth bounding boxes. IoU Calculation: IoU is calculated as the ratio of the intersection area to the union area between the binary attention map and the ground truth mask.

Additionally, the dataset is divided into three subsets: *Easy*, *Medium*, and *Hard*, categorized by occlusion levels, background complexity, and object size variance. This division facilitates a nuanced analysis of model performance under varying degrees of difficulty.

C.2 More Results

Table 9 summarizes IoU performance across the three subsets. Our method consistently outperforms the baseline (VPT), with notable improvements in the *Medium* and *Hard* subsets, where occlusions and intricate backgrounds present significant challenges. These results highlight the efficacy of our approach in handling complex localization tasks.

Table 9: Detailed IoU analysis across subsets in CUB-200.

Methods	IoU Performance (%)			Mean IoU (%)
	Easy	Medium	Hard	
VPT	38.2	25.6	16.8	26.5
Ours	**45.7** (+7.5)	**32.9** (+7.3)	**22.5** (+5.7)	**32.9** (+6.4)

C.3 Subset-Wise Performance Analysis

In addition to mean IoU, we analyze IoU variance within each subset to evaluate stability. Table 10 shows that our method not only achieves higher IoU scores but also demonstrates lower variance across all subsets, indicating enhanced consistency in localization performance regardless of sample difficulty.

Table 10: IoU variance analysis across subsets in CUB-200.

Methods	Variance in IoU (%)			Overall Variance (%)
	Easy	Medium	Hard	
VPT	4.5	6.2	8.7	6.5
Ours	**3.2** (-1.3)	**4.8** (-1.4)	**7.1** (-1.6)	**5.0** (-1.5)

C.4 Concluding Observations

Our method achieves significant IoU gains, particularly in challenging subsets with higher occlusion and complex backgrounds, validating its robustness in diverse scenarios. The reduced variance in IoU results across all subsets indicates that our method provides more consistent and reliable attention localization, even for hard-to-detect objects. Visual inspections and quantitative results confirm that our method generalizes effectively to unseen samples, maintaining high localization accuracy without overfitting.

These results underscore the effectiveness of incorporating semantic prompts to direct the model's attention to meaningful object features, enhancing both localization accuracy and robustness.

D More Mutual Information Analysis

To further elaborate on the mutual information (MI) analysis presented in the main text, we provide additional experiments and insights to validate the effectiveness of our method in achieving optimal representation learning under the Information Bottleneck (IB) framework. These experiments delve deeper into the mutual information trends across different Transformer layers and investigate the role of our fundamental image prior visual prompts in shaping the learning dynamics.

D.1 Experimental Details

We conducted the mutual information analysis using the Mutual Information Neural Estimator (MINE) to compute $I(X;T)$ and $I(T;Y)$ for all 12 Transformer layers in both the baseline VPT and our method. The settings for these experiments include:

Datasets and Models: The experiments are conducted on the CUB-200 dataset with Vision Transformers (ViT) as the backbone.

Training Procedure: Both models are trained using identical settings to ensure fair comparisons.

Mutual Information Estimation: For each layer T, $I(X;T)$ and $I(T;Y)$ are estimated using MINE with a mini-batch size of 256. The estimation results are averaged over the entire test set.

Compression of Input Data ($I(X;T)$): Both the baseline VPT and our method exhibit similar $I(X;T)$ trends across lower layers, as expected. This indicates that the introduction of visual prompts does not significantly alter the compression of input data.

Correlation with Labels ($I(T;Y)$): Our method achieves consistently higher $I(T;Y)$ values, particularly in the top layers, compared to the baseline. This demonstrates that the representations learned by our method are more aligned with the labels, enabling better generalization to unseen data.

Lower Information Bottleneck ($\mathcal{L}_{\text{IB}}$): The significant reduction in $I(X;T) - \beta I(T;Y)$ for the top layers in our method aligns with the IB hypothesis. This reduced bottleneck reflects the effectiveness of our visual prompts in focusing on label-relevant features while suppressing redundant information.

D.2 Impact of Semantic Prompts on Representation Learning

The role of our semantic prompts can be further elucidated by decomposing the MI contributions: Extracting Inherent Image Priors: The fundamental image prior visual prompts (e.g., texture, shape) enhance feature extraction by aligning the representations with inherent characteristics of the input images. Improving Label Correlation: Semantic prompts guide the model to retain more label-relevant information, increasing $I(T;Y)$ while maintaining $I(X;T)$ stability. This is particularly evident in tasks requiring fine-grained classification, where semantic prompts enable the model to focus on subtle discriminative features.

Table 11: Comparison of mutual information metrics ($I(X;T)$ and $I(T;Y)$) in the top three Transformer layers for VPT and our method on CUB-200.

Method	$I(X;T)$ (%)	$I(T;Y)$ (%)
VPT	34.2	22.8
Ours	**33.8**	**29.6**

Baseline VPT: The $I(T;Y)$ plateau in the top layers suggests limited improvement in label correlation, indicating potential underutilization of higher-layer representations.

Our Method: The steep increase in $I(T;Y)$ in the top layers reflects enhanced label alignment, validating the role of semantic prompts in guiding representation learning.

D.3 Concluding Observations

The expanded mutual information analysis reaffirms the effectiveness of our method in achieving superior representation learning under the IB framework. Key takeaways include:

Improved Generalization: Higher $I(T;Y)$ values for the top layers demonstrate the ability of our method to focus on label-relevant features, enabling better generalization to unseen data.

Reduced Redundancy: Comparable $I(X;T)$ values suggest that our visual prompts do not introduce unnecessary complexity, maintaining the efficiency of the learned representations.

Future Directions: Further exploration of adaptive semantic prompts and dynamic feature compression mechanisms could enhance the flexibility and scalability of our method across more diverse tasks.

E Discussion: Hand-Crafted Priors vs. Deep Learned Priors

A natural question arises regarding the choice of priors: *Why rely on classical hand-crafted operators (e.g., Sobel, Gabor) instead of extracting features from a frozen deep neural network (e.g., ResNet or DINO) as prompts?*

While integrating deep features might initially seem intuitive, we argue that our hand-crafted approach is theoretically and practically superior in the context of Parameter-Efficient Fine-Tuning (PEFT) for three key reasons:

1. **Information Orthogonality:** Deep features extracted from networks like ResNet are conceptually homogeneous to the semantic features already learned by the ViT backbone itself (i.e., high-level abstractions). Adding them creates information redundancy. In contrast, our hand-crafted operators explicitly capture low-level statistics—such as high-frequency gradients (Sobel) and spectral texture information (Gabor)—that deep networks tend to abstract away or ignore in deeper layers due to texture bias (Geirhos et al., 2018b). These "primitive" cues provide orthogonal, complementary guidance that corrects the inherent biases of the ViT backbone.

2. **Domain Robustness:** Deep feature extractors (e.g., ImageNet-trained ResNet) often suffer from domain shift when applied to specialized downstream tasks (e.g., medical or satellite imagery in VTAB-1k). Hand-crafted priors, however, rely on fundamental signal processing principles (e.g., edge gradients, color distribution) that are domain-agnostic and universally applicable, ensuring consistent improvements across diverse datasets without negative transfer.

3. **Strict Efficiency:** The core philosophy of PEFT is to adapt large models with minimal resource overhead. Utilizing a secondary deep network as a prior extractor, even if frozen, requires storing and computing over millions of additional parameters (e.g., ∼11M for ResNet-18), contradicting the lightweight nature of our task. In comparison, our hand-crafted operators are parameter-free and computationally negligible ($\mathcal{O}(HW)$ complexity), strictly adhering to the efficiency constraints of the PEFT paradigm.

In summary, our design prioritizes *complementary low-level guidance* and *maximum efficiency* over the redundancy and computational burden of stacking multiple deep neural networks. ·

- **Orthogonal Information:** Deep features from a ResNet are conceptually similar to the features learned by the ViT backbone itself (i.e., semantic abstractions). In contrast, hand-crafted operators explicitly capture low-level statistics (gradients, frequency spectra) that deep networks tend to abstract away or ignore in deeper layers. These "primitive" cues serve as a stronger complementary signal to the ViT.

- **Efficiency:** Utilizing a deep network as a prior extractor introduces significant memory and storage overhead (even if frozen), contradicting the parameter-efficient philosophy of PEFT. Our hand-crafted operators are computationally negligible and parameter-free.

F Implementation Details and Analysis of Fundamental Operators

In this section, we provide the precise implementation details of the fundamental image prior operators employed in our method. Furthermore, we elaborate on the theoretical rationale behind selecting this specific combination of operators (Color, Texture, and Shape) and discuss the robustness of our method to different operator choices, supported by our ablation studies.

F.1 Specific Implementation of Operators

To capture the fundamental visual statistics of the input image $X \in \mathbb{R}^{H\times W\times 3}$, we employ three distinct types of hand-crafted operators. The outputs of these operators are concatenated and projected via a linear layer to align with the prompt token dimension.

1. Color Prior: Histogram Statistics.
Color is one of the most expressive and invariant visual cues, robust to rotation and scaling (Swain & Ballard, 1991b). We compute the color histogram features as follows:

- **Color Space:** We utilize the HSV (Hue, Saturation, Value) color space, which decouples chromatic information (Hue/Saturation) from intensity (Value), providing better robustness to lighting changes compared to RGB.

- **Implementation:** For each channel, we compute a histogram with $B = 24$ bins. The resulting histograms are normalized to form a probability distribution and concatenated, resulting in a feature vector of dimension $3 \times 24 = 72$. This vector serves as a global statistical summary of the image's chromatic distribution.

2. Texture Prior: Gabor Filters.
Texture analysis is crucial for distinguishing materials and repetitive patterns. We adopt **Gabor filters** (Manjunath & Ma, 2002), which are biologically inspired by the receptive fields of simple cells in the mammalian visual cortex (V1).

- **Implementation:** We generate a filter bank containing Gabor kernels at 4 distinct orientations ($\theta \in \{0°, 45°, 90°, 135°\}$) and a single scale. The filters are convolved with the grayscale version of the input image.
- **Feature Map:** Instead of global pooling, we retain the spatial response maps to preserve local texture spatiality. These maps are then flattened or patchified to align with the token sequence structure.

3. Shape Prior: Sobel Operator.
Shape and edge information provide structural constraints that are often complementary to texture (Geirhos et al., 2018b). We employ the **Sobel operator** (Kanopoulos et al., 1988) to extract gradient information.

- **Implementation:** We compute the discrete gradients along the horizontal (G_x) and vertical (G_y) directions using standard 3×3 kernels. The gradient magnitude is calculated as $G = \sqrt{G_x^2 + G_y^2}$.
- **Outcome:** This results in an edge map that highlights high-frequency structural boundaries, guiding the model to focus on object shapes rather than background noise.

F.2 Rationale for Operator Selection

Our selection of operators is not arbitrary but grounded in the principle of **Orthogonal Complementarity**. Deep learning models, particularly CNNs and ViTs, often exhibit a "texture bias" (Geirhos et al., 2018b). By explicitly injecting complementary priors, we ensure a balanced representation:

1. **Completeness:** The combination of *Color* (spectral), *Texture* (spatial-frequency), and *Shape* (structural/spatial) covers the three fundamental pillars of low-level computer vision. Removing any single component results in an information void that the randomized prompts alone may struggle to fill (as evidenced in Table 5 of the main text).
2. **Interpretability & Stability:** Unlike learnable priors (e.g., CNN adapters), these hand-crafted operators are deterministic and theoretically well-understood. Using standard operators like Sobel and Gabor ensures that the injected "hard prompt" provides stable, domain-invariant cues that do not drift during the fine-tuning process.

F.3 Robustness to Operator Variants

A pertinent question is whether the success of our method relies on specific operator choices (e.g., Sobel vs. Canny for shape). To investigate this, we conducted extensive comparisons using different operator variants (see Figure 7 in the main paper).

- **Texture Variants (Gabor vs. LBP):** We compared Gabor filters with Local Binary Patterns (LBP) (Ojala et al., 2002). Results show that both yield significant improvements over the baseline, with Gabor slightly outperforming LBP on fine-grained tasks due to its continuous response nature.

- **Shape Variants (Sobel vs. Canny vs. Laplacian):** We tested Sobel against the Canny edge detector (Canny, 2009) and Laplacian operators. The performance variance was minimal ($< 0.3\%$), indicating that the *presence* of structural prior is more critical than the *type* of edge extractor used.

In conclusion, our choice of HSV Histograms, Gabor filters, and Sobel operators represents a standard, computationally efficient, and representative set of priors. However, the proposed framework is general-purpose: it benefits effectively from the semantic category of the prior (e.g., "Shape information") rather than overfitting to a specific algorithm.

G Limitations and Future Work

While our proposed *Cascaded Semantic Prompting* demonstrates superior performance and interpretability across various benchmarks, we identify certain limitations that pave the way for future research directions.

G.1 Limitations

Our method relies on fixed, hand-crafted operators (e.g., Sobel, Gabor) to extract fundamental image priors. While this design choice ensures computational efficiency and the convenience of an "extract-once" strategy, it inherently limits the model's adaptability compared to fully learnable modules. These fixed operators cannot evolve during training to capture dataset-specific idiosyncrasies that may fall outside standard color, texture, and shape definitions.

Furthermore, although our method outperforms existing PEFT approaches on the *Structured* split of the VTAB-1k benchmark (achieving 61.16% mean accuracy compared to 58.83% for AdaptFormer), there remains room for improvement on tasks requiring complex geometric reasoning, such as *SmallNORB* and *dSprites*. This suggests that while our fundamental priors effectively capture low-level statistics, they may not fully encapsulate the high-level 3D geometric relationships required for these specific specialized tasks. Finally, as a prompt tuning paradigm, the upper bound of performance is inevitably tied to the quality and pre-training domain of the underlying frozen backbone.

G.2 Future Work

Building on these observations, future research could explore adaptive prior mechanisms. Instead of a static concatenation of Color, Texture, and Shape priors, a lightweight gating mechanism or attention module could be introduced to dynamically weight these priors based on the input instance. This would allow the model to autonomously determine whether texture or shape is more critical for a specific image, potentially enhancing performance on diverse datasets.

Additionally, the concept of "Fundamental Image Priors" holds promise for extension to other modalities. For example, optical flow or motion boundary histograms could serve as temporal priors for video recognition, while surface normals could function as geometric priors for 3D point cloud analysis. Investigating the applicability of cascaded semantic prompting to emerging architectures beyond Transformers, such as State Space Models, also represents a promising direction to test the universality of our approach.