PROMPT-MATCHED SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The objective of this work is to explore how to effectively and efficiently adapt pre-trained visual foundation models to downstream tasks, e.g., image semantic segmentation. Conventional methods usually fine-tuned the entire networks for each specific dataset, which will be burdensome to store massive parameters of these networks. Several recent works attempted to insert some extra trainable parameters into the frozen networks to learn visual prompts for parameter-efficient tuning. However, these works showed poor generality as they were designed specifically for Transformers. Moreover, using limited information in these schemes, they exhibited a poor capacity to learn effective prompts. To alleviate these issues, we propose a novel Inter-Stage Prompt-Matched Framework for generic and effective visual prompt tuning. Specifically, to ensure generality, we divide the pre-trained backbone with frozen parameters into multiple stages and perform prompt learning between different stages, which makes the proposed scheme applicable to various architectures of CNN and Transformer. For effective tuning, a lightweight Semantic-aware Prompt Matcher (SPM) is designed to progressively learn reasonable prompts with a recurrent mechanism, guided by the rich information of interim semantic maps. Working as a deep matched filter of representation learning, the proposed SPM can well transform the output of the previous stage into a desirable input for the next stage, thus achieving the better matching/stimulating for the pre-trained knowledge. Finally, we apply the proposed method to handle various semantic segmentation tasks. Extensive experiments on five benchmarks show that the proposed scheme can achieve a promising trade-off between parameter efficiency and performance effectiveness.

1 INTRODUCTION

Over the past decade, various Convolutional Neural Networks (CNN) and Transformer architectures have been proposed for visual understanding. In general, previous works usually pre-trained a foundation model on large-scale benchmarks such as ImageNet (Deng et al., 2009), and then fine-tuned the network's parameters on specific downstream datasets. In the literature, there were usually two different strategies for model fine-tuning. The first one is ***full-tuning*** that adjusts all parameters of the entire network. However, it usually requires large amounts of training data with rich annotations for effective representation learning. Moreover, this strategy requires storing a proprietary model with massive parameters for each task/dataset, which is expensive and unsustainable for many service platforms. As the second strategy, ***head-tuning*** freezes the parameters of the backbone network and only optimizes the model's head. Intuitively, all tasks share the same backbone and we only need to maintain an individual head for each task. Despite being parameter efficient, this strategy has limited capacity to learn discriminative representations and can not well exploit pre-trained knowledge to handle complex visual understanding. Overall, these strategies suffer from various issues, and we desire a more effective and efficient fine-tuning method for widespread downstream tasks.

Recently, ***prompt-tuning*** (Liu et al., 2021) has achieved considerable results in Natural Language Processing (NLP). Some textual prompts are inserted into the downstream input to better explore the knowledge of language models. To facilitate downstream visual tasks in computer vision, a few works (Jia et al., 2022; Chen et al., 2022) have attempted to apply visual prompt tuning to energize the prior knowledge in those parameter-frozen foundation models. Despite certain progress using only a small amount of extra parameters, these works suffer from the following problems. **First**, they were specially designed for Transformer, not generic to commonly-used CNN architectures.

For instance, VPT (Jia et al., 2022) learned visual prompts in the token space of each transformer layer, while AdaptFormer (Chen et al., 2022) replaced the original MLP block with a trainable bottleneck module. Moreover, these works significantly modified the original structure of foundation units[1], making them inapplicable on many existing high-speed inference devices, where foundation units and their parameters have been embedded (Ma et al., 2017; Wang et al., 2022). **Second**, these works performed prompt tuning with limited information in a black-box mapping manner. Specifically, they only utilized the final recognition loss to optimize prompted modules and had limited capacities to learn reasonable prompts. It is worth noting that some downstream tasks (such as semantic segmentation) are more challenging than image recognition, requiring richer information to perform elaborated inferences. Therefore, informative knowledge should be fully explored to generate effective visual prompts for parameter-efficient representation learning.

To address the above problems, we propose a novel Inter-Stage Prompt-Matched Framework, which can adapt those visual foundation models of different architectures to facilitate widespread downstream tasks, e.g., semantic segmentation under various scenarios. Specifically, instead of modifying each foundation unit with extra parameters, we partition the backbone network with frozen parameters into multiple stages and perform architecture-independent inter-stage prompt tuning for specific datasets. Furthermore, we introduce a lightweight module termed Semantic-aware Prompt Matcher (SPM) that incorporates rich information of interim semantic maps to learn reasonable visual prompts between any two stages in a progressive and recurrent manner. Working as a deep matched filter, our SPM can effectively transform the output representation of the previous stage into an appropriate input representation for the next stage, making it better to match/stimulate the pre-trained knowledge in the frozen backbone. Finally, to verify the generality of our method, we apply the proposed SPM to fine-tune various backbone networks such as ResNet (He et al., 2016) and Vision Transformer (Dosovitskiy et al., 2020) to handle semantic segmentation of natural, satellite, and medical images. Extensive experiments conducted on five benchmarks show the parameter efficiency and performance effectiveness of the proposed method. In particular, our method significantly outperforms *head-tuning* and is comparable to *full-tuning*, optimizing only a small number of parameters of our SPM and the head segmenter.

In summary, the contributions of our work are three-fold:

- A novel Inter-Stage Prompt-Matched Framework is proposed to learn task-relevant visual prompts between different stages of the pre-trained foundation model with frozen parameters. Without specifically modifying those foundation units, our method is universal for various network architectures of CNN and Transformer.
- A lightweight SPM is introduced to progressively learn reasonable visual prompts between two adjacent stages with a recurrent mechanism. Guided by the rich information of interim semantic maps, our SPM can well transform the output of the previous stage into an informative input for the next stage that can better energize the pre-trained knowledge.
- Extensive experiments on five benchmarks demonstrate that our method is performance-effective and parameter-efficient for fine-tuning foundation models to various semantic segmentation tasks.

2 Related Work

Prompt Tuning: In recent years, various large-scale NLP models such as BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), and Pangu-α (Zeng et al., 2021) have been developed by pre-training on huge datasets. With the emerging prompt tuning, these large-scale models achieved impressive transfer performance on myriads of downstream tasks such as translation (Tan et al., 2022), reading comprehension (Hu et al., 2022), question answering (Yang et al., 2022), etc. Inspired by the NLP prompt tuning paradigm, some computer vision researchers have attempted to facilitate visual understanding by fine-tuning vision-language models, where the textual prompts generated by text encoders were used to guide the representation learning of visual encoders (Radford et al., 2021; Jia et al., 2021). Despite achieving promising performance, these works relied heavily on the textual prompt design and can not be smoothly applied to various vision tasks. With this concern, a few recent works (Jia et al., 2022; Chen et al., 2022) introduced some extra trainable parameters

[1] In our work, those stacked base modules in foundation models are called foundation units, such as the transformer layer for Transformer and the residual module for Residual Networks.

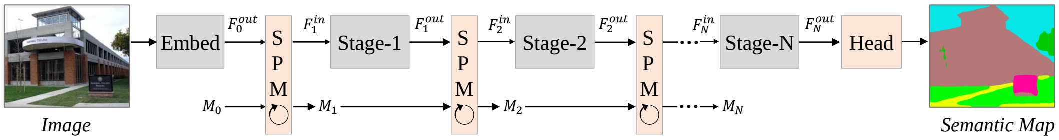

Figure 1: The architecture of the proposed Inter-Stage Prompt-Matched Framework for semantic segmentation prompt tuning. A lightweight Semantic-aware Prompt Matcher (SPM) is introduced to learn reasonable visual prompts recurrently between every two stages of the frozen backbone network. Guided by the rich information of interim semantic maps, the i-th SPM transforms the output feature F_{i-1}^{out} of the previous stage into a suitable input feature F_i^{in} for the next stage. M_i is the interim semantic map generated at the i-the SPM. The symbol ↻ denotes the recurrent prompt learning mechanism in our SPM. It is worth noting that only the parameters of our SPM and head segmenter are updated during the training phase.

to directly learn visual prompts from the input visual features or randomized noise. Nevertheless, they were specially designed for Transforms and may fail to generate effective prompts due to the scarcity of instructive information. Therefore in this work, we are committed to exploring general and effective strategies for visual prompt tuning.

Semantic Segmentation: As a typical pixel-wise prediction problem, semantic segmentation has been significantly promoted by deep neural networks. For instance, Long et al. (2015) proposed Fully Convolutional Networks that replaced fully-connected layers with convolutional layers to handle images of arbitrary sizes. Ronneberger et al. (2015) applied a convolutional encoder-decoder architecture with skip connections to generate semantic maps with high resolutions. Zhao et al. (2017) used a backbone network to extract the feature maps of input images and then introduced a pyramid pooling module to aggregate different sub-region representations for multiscale contextual modeling. DeepLab family (Chen et al., 2015; 2017) further applied dilated convolutions to enlarge the network receptive fields and introduced Conditional Random Fields to refine the final segmentation results. Recently, transformers (Vaswani et al., 2017) have also been applied to address this problem. One representative work is Segmenter (Strudel et al., 2021), which divided the image into local patches and fed their linear embeddings into Vision Transformer (Dosovitskiy et al., 2020) to capture global context at each layer for semantic segmentation. Despite progress, previous methods usually fine-tuned the parameters of the whole networks respectively for each specific task of semantic segmentation. It is burdensome to store massive parameters of these models, especially on some resource-constrained devices. Under this circumstance, we crave for a novel fine-tuning approach, where these tasks can share most of the parameters and achieve competitive results.

3 Methodology

3.1 Overall Architecture

In this work, we aim to utilize visual prompt learning to fine-tune the pre-trained foundation models for various downstream tasks, e.g., semantic segmentation. When developing our algorithm, we consider the following two questions: i) ***where to learn visual prompts*** and ii) ***how to learn reasonable prompts***? We argue that previous works (Jia et al., 2022; Chen et al., 2022) that performed customized prompt learning within each foundation unit are not applicable to different network architectures. We also observe that it is suboptimal to learn prompts with limited information of the final loss of the model's head. So we desire an architecture-general prompt tuning model that fully exploits rich information to learn effective visual prompts.

To this end, we propose a unified Inter-Stage Prompt-Matched Framework to effectively and efficiently fine-tune pre-trained foundation models to deal with downstream visual tasks. Here we take semantic segmentation as an example to illustrate the working process of our method. As shown in Figure 1, a semantic segmentation model usually consists of an universal backbone network pre-trained on a large-scale dataset and a customized head segmenter with random initialization. To reduce the number of tunable/stored parameters, we freeze the backbone network so that it can be shared by different tasks of semantic segmentation, while the head segmenter is optimized for each specific dataset. Inspired by previous NLP works (Liu et al., 2021), we apply prompt tuning to efficiently exploit the backbone pre-trained knowledge to facilitate the visual representation learning of downstream semantic segmentation. To make our method applicable to various network architectures, we propose to learn visual prompts between different stages of the frozen backbone, without modifying the original structures of foundation units, e.g., residual module or transformer layer. Specifically, based on its architecture, we partition the backbone network into N stages, each of which is composed of multiple foundation units. Notice that there may be some embedding layers before the first stage. For convenience, the output feature of the stage i is denoted as F_i^{out} (i=1,...,N), while the output feature of those embedding layers is denoted as F_0^{out}.

We then introduce a differentiable Semantic-aware Prompt Matcher (SPM) to learn visual prompts between two adjacent stages using a small number of parameters. As mentioned above, rich information is desired to perform prompt learning for various vision tasks including semantic segmentation. In this work, we find that interim semantic maps generated at intermediate layers can provide fine-grained prior information of object semantics distributions. Therefore, before the stage i, our SPM incorporates the output feature F_{i-1}^{out} and the interim semantic map M_{i-1} of the previous stage to progressively learn reasonable visual prompts with a recurrent mechanism, since it is difficult to directly generate desirable prompts in some complex scenarios. After multiple iterations, we can obtain a refined semantic map M_i and a suitable input F_i^{in} for the stage i. This process can be formulated as:

$$F_i^{in}, M_i = SPM(F_{i-1}^{out}, M_{i-1}, K), \tag{1}$$

where K denotes the number of recurrent iterations. Notice that the initial semantic map M_0 is generated from the statistic category probability. More specifically, each position on M_0 is set to the pixel probability vector of different categories in the training set of the downstream dataset. Intuitively, our SPM works as a representation-level matched filter that can better match input features with pre-trained knowledge. More details of the proposed SPM are described in Section 3.2.

As shown in Figure 1, our SPM is hierarchically inserted between different stages to learn semantic-aware visual prompts for downstream representation learning. Finally, the output feature F_N^{out} of the N-th stage is fed into the head segmenter to generate the high-quality semantic map M.

3.2 Semantic-aware Prompt Matcher

In this subsection, we introduce the details of the proposed SPM. The purpose of this module is to integrate rich information of interim semantic maps to learn reasonable visual prompts with a recurrent mechanism, so that the output feature of the previous stage can be transformed into a desirable input for the next stage. Specifically, before the stage i, our SPM first takes the feature F_{i-1}^{out} and the semantic map M_{i-1} to generate semantic-aware prompts. Similar to Recurrent Neural Networks (Lipton et al., 2015), the prompted feature and the refined semantic map are fed back into SPM for recurrent prompt learning. Thus the Eq.1 can be unfolded as follows:

$$\begin{aligned}
&F_i^0, M_i^0 = F_{i-1}^{out}, M_{i-1}, \\
&F_i^k, M_i^k = f(F_i^{k-1}, M_i^{k-1}, \theta) \quad \text{for } k \text{ in } \{1, ..., K\}, \\
&F_i^{in}, M_i = F_i^K, M_i^K,
\end{aligned} \tag{2}$$

where $f(\cdot)$ represents the SPM function with trainable parameters θ. F_i^k and M_i^k are the prompted feature and the refined semantic map at the k-th iteration of the i-the SPM. After K iterations, we can obtain the desirable input F_i^{in} for the next stage to better energize the pre-trained knowledge. Notice that θ are shared for all iterations for parameter efficiency.

Figure 2 shows the visual prompt tuning for the k-th iteration of the i-the SPM. We can see that our SPM consists of two parallel branches to refine the interim semantic map and generate the prompted feature. As mentioned in previous works (Zhu et al., 2019; He et al., 2019), long-range spatial context is crucial for semantic segmentation. Thus we develop a group Pyramid Dilation Convolution (PDC) that uses four dilated group convolutional layers to capture the multi-scale lang-range context. As shown in Figure 3, the input feature of PDC is divided into four sub-features along the channel dimension. The i-th sub-feature is fed into the i-th convolutional layer with a kernel size of 3×3 and a dilated rate of r. The outputs of all dilated layers are concatenated and fused using a 1×1 convolutional layer. The proposed PDC is integrated into both branches to generate the long-range contextualized features. The details of these branches are described as follows.

Interim Semantic Map Refinement: In this branch, we use the current feature to enhance the semantic map generated from the previous feature. Specifically, we first feed the concatenation of

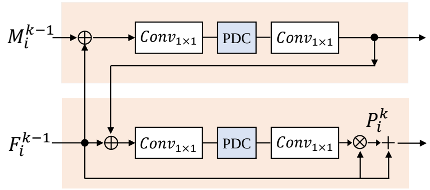

Figure 2: The architecture of Semantic-aware Prompt Matcher. $Conv_{1\times1}$ denotes a convolutional layer with a kernel size of 1×1 and PDC is our Pyramid Dilation Convolution. $\oplus$ represents the feature concatenation operation, while $\otimes$ denotes the element-wise multiplication.

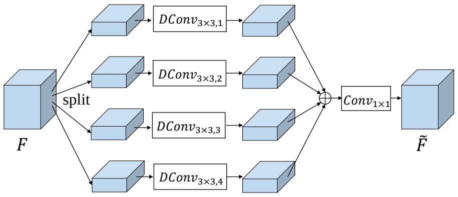

Figure 3: The architecture of Pyramid Dilation Convolution for long-range spatial context modeling. $DConv_{3\times3,r}$ denotes a dilated group convolutional layer with a kernel size of 3×3 and a dilated rate of r. The input feature F and output feature $\tilde{F}$ have the same dimension.

F_i^{k-1} and M_i^{k-1} into a 1×1 group convolutional layer to generate a compact feature $F_M^{i,k}$, which has a low channel number c to reduce the trainable parameters. We then apply the proposed PDC to obtain the long-range contextualized feature $\tilde{F}_M^{i,k}$, which is further fed into another 1×1 group convolutional layer to generate the refined semantic map M_i^k. This process can be formulated as:

$$F_M^{i,k} = Conv_{1\times1}(F_i^{k-1} \oplus M_i^{k-1}), \quad \tilde{F}_M^{i,k} = PDC(F_{i-1}^{m1}), \quad M_i^k = Softmax\{Conv_{1\times1}(\tilde{F}_M^{i,k})\}, \tag{3}$$

where $\oplus$ denotes the feature concatenation operation and the $Softmax$ layer is used to normalize the predicted scores of semantic categories at each location.

Semantic-aware Prompt Generation: We then incorporate the feature F_i^{k-1} and the refined map M_i^k to learn visual prompts progressively. Similar to the first branch, we feed F_i^{k-1} and M_i^k into a 1×1 convolutional layer and a PDC to obtain the features $F_P^{i,k}$ and $\tilde{F}_P^{i,k}$. Inspired by the attention mechanism, we utilize another 1×1 convolutional layer to generate a prompt weight $W_P^{i,k}$, which is further applied to multiply with F_i^{k-1} to generate the visual prompt map P_i^k. Finally, F_i^{k-1} and P_i^k are added to obtain the new prompted feature F_i^k. This process can be formulated as:

$$\begin{aligned} F_P^{i,k} &= Conv_{1\times1}(F_i^{k-1} \oplus M_i^k), \quad & \tilde{F}_P^{i,k} &= PDC(F_P^{i,k}), \\ W_P^{i,k} &= Conv_{1\times1}(\tilde{F}_P^{i,k}), & P_i^k &= F_i^{k-1} \otimes W_P^{i,k}, \quad F_i^k = F_i^{k-1} + P_i^k, \end{aligned} \tag{4}$$

where $\otimes$ denotes the element-wise multiplication operation.

3.3 Network Optimization

During training, the backbone network is frozen and we only update the parameters of our SPM and the head segmenter on specific datasets. We use the Cross-Entropy (CE) loss function to optimize our network. The total loss is defined as follows:

$$loss = CE(M, \hat{M}) + \sum_{i=1}^{N} \sum_{k=1}^{K} a_i * CE(M_i^k, \hat{M}), \tag{5}$$

where $\hat{M}$ is the ground-truth semantic map and M is our final semantic map predicted by the head segmenter. a_i is the loss weight of the interim semantic maps generated by the i-th SPM.

4 Experiment

4.1 Downstream Datasets

In this work, we conduct extensive experiments on five semantic segmentation datasets of various scenarios, including ADE20K (Zhou et al., 2017), Vaihingen[2], CHASE-DB1 (Fraz et al., 2012), STARE (Hoover et al., 2000) and HRF (Budai et al., 2013). The overview of these datasets is summarized in Table 1 and their details are described as follows. Some examples of these datasets are visualized in Figure 4.

[2] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

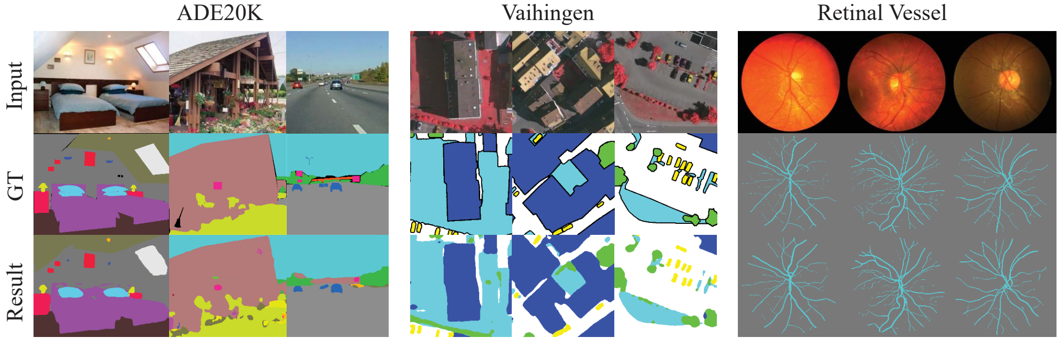

Figure 4: Visualization of various samples of semantic segmentation and the results of the proposed method. Using the same frozen backbone network, our method can generate high-quality semantic maps for natural, satellite, and medical image segmentation.

ADE20K: This is a large-scale scene parsing dataset with 150 object and stuff classes. The public ADE20K consists of 20,210 images for training and 2,000 images for validation. Those images have different resolutions and their objects suffers from great scale variations. This dataset is challenging due the high intra-class variance and low inter-class variance.

Table 1: Overview of five downstream benchmarks of semantic segmentation.

Dataset	#Train	#Test	#Category	Scene
ADE20K	20,210	2,000	150	natural image
Vaihingen	344	398	6	satellite image
CHASE-DB1	20	8	2	medical image
STARE	10	10		
HRF	15	30		

Vaihingen: It is a medium-scale dataset for satellite image segmentation. This dataset consists of six types of semantic object, e.g., buildings, streets, cars, etc. The training set contains 344 images, while the testing set has 398 images. All images has the same resolution 512 × 512.

CHASE-DB1, STARE, HRF: They are small-scale medical image segmentation datasets that aim to segment retinal vessels from background. These datasets consist of the eye fundus images of some healthy patients and some patients with representative eye diseases. These images have higher resolutions, e.g., 999 × 960 for CHASE-DB1, 700 × 605 for STARE, 3504 × 2336 for HRF.

4.2 Comparison with Common Fine-tuning Methods

Here we compare the proposed method with three common approaches, i.e., *full-tuning*, *head-tuning* and *learn-from-scratch*. The overall experimental settings are as follows:

- Backbone Network: In this subsection, all methods adopt ResNet-101 (He et al., 2016) as the backbone network. Except *learn-from-scratch*, the backbone of other methods was pre-trained with a auxiliary task of image recognition on ImageNet (Deng et al., 2009). The experiments of the Transformer backbone can be referred to Section 4.4.
- Head Segmenter: For ADE20K and Vaihingen, we use Pyramid Pooling Module (Zhao et al., 2017) as the head segmenter, as it has a great capacity to capture the scale variations of objects. For those medical datasets, we use the Progressive UPsampling head (Zheng et al., 2021), because retinal vessels are very thin and we require high-resolution segmentation results for medical diagnosis. The parameters of all heads are randomly initialized.
- Semantic-aware Prompt Matcher: Based on its architecture, the pre-trained ResNet-101 is divided into four stages. In our framework, the proposed SPM is inserted before each stage and the head segmenter to learn visual prompts hierarchically. The reduced channel $\hat{c}$ in our SPM is uniformly set to 256 and the group number of convolutional layers in PDC is set to 16. Therefore, our SPM introduces only a small number of extra parameters.
- Training Details: We apply MMSegmentation (Contributors, 2020) to implement our experiments on 4 Nvidia GeForce 3090 GPUs. All methods are trained for 80k iterations with a batch size of 16 for ADE20K and Vaihingen. For those medical datasets, we train the models for 8k iterations with a batch size of 4, due to the small amount of training data. Other hyper-parameters are completely set to the default settings in MMSegmentation.

Table 2: Performance on the ADE20K dataset for large-scale natural image segmentation, when the backbone network is ResNet-101 pre-trained on ImageNet.

Method	Trainable Parameters (M)			mIoU ↑
	Backbone	Prompt	Head	
Full-Tuning	42.41	0	25.54	43.96
Learn from Scratch	42.41	0	25.54	34.84
Head-Tuning	0	0	25.54	34.08
Ours	0	3.11	25.54	41.83

Table 3: Performance on the Vaihingen dataset for medium-scale satellite image segmentation, when the backbone network is ResNet-101 pre-trained on ImageNet.

Method	Trainable Parameters (M)			mIoU ↑
	Backbone	Prompt	Head	
Full-Tuning	42.41	0	25.43	73.96
Learn from Scratch	42.41	0	25.43	68.85
Head-Tuning	0	0	25.43	62.45
Ours	0	2.22	25.43	72.17

Large-scale Natural Image Segmentation: Table 2 shows the performance of different methods on the ADE20K dataset. We can observe that *full-tuning* obtains the best mIoU of 43.96%, much outperforming *learn-from-scratch* with the same number of trainable parameters. This is attributed to the fact that the pre-trained backbone contains rich prior knowledge. When the backbone is frozen, *head-tuning* gets a poor mIoU of 34.08%. In contrast, our method can achieve a promising mIoU of 41.83%, by using a small amount of 3.11M parameters to learn visual prompts between frozen stages. We can observe that the proposed method is significantly better than *head-tuning* and comparable to *full-tuning*. Figure 4 shows that our method can generate high-quality semantic maps for unconstrained natural scenarios.

Medium-scale Satellite Image Segmentation: Table 3 summarizes the results of all methods on the Vaihingen dataset. We can see that the compared results on this dataset are consistent with that on the ADE20K dataset. Specifically, *full-tuning* fine-tunes all the parameters to obtain the best performance with a mIoU of 73.96%. When freezing the backbone and using 2.22M trainable parameters for prompt learning, our method achieves a competitive mIoU of 72.17%, significantly outperforming *learn-from-scratch* and *head-tuning*. This shows the great potential of our method for medium-scale satellite image segmentation.

One-shot Medical Image Segmentation: Here we apply the proposed model to deal with one-shot medical semantic segmentation, as it is difficult to obtain massive medical images with pixel-wise annotations. Specifically, we conduct five one-shot experiments. In each experiment, we randomly select a sample to train the model and then validate it on the testing set. Table 4 shows the mean results and variances of different methods on three retinal segmentation datasets. By introducing few parameters for prompt tuning, our method outperforms *head-tuning* and *learn-from-scratch* consistently on all datasets. Moreover, our method is also comparable to *full-tuning* which requires training and storing a large number of parameters.

4.3 Ablation Studies

Effects of Different Pre-training Datasets: In this subsection, we explore the effect of pre-training data from different source domains. Besides the backbone pre-trained on ImageNet, we reimplement our experiments using the ResNet-101 backbone pre-trained on the COCO-Stuff-164K dataset (Caesar et al., 2018), a large-scale semantic segmentation benchmark. Table 5 summarizes the results of different methods based on the COCO pre-trained backbone. Compared with the results of ImageNet pre-training in Table 2, we observe that our method can achieve a better mIoU of 42.11% by using fewer prompt learning parameters (1.78M). This indicates that it is beneficial to choose models pre-trained on a proper source domain. However, many downstream tasks do not have large-scale datasets in practice, therefore it makes sense to adopt the foundation models pre-trained on ImageNet with good generalization.

Table 4: The foreground Dice Similariy Coefficient (Dice) of one-shot medical image semantic segmentation, when the backbone network is ResNet-101 pre-trained on ImageNet.

Method	Trainable Parameters (M)			Dice ↑		
	Backbone	Prompt	Head	CHASE-DB1	STARE	HRF
Full-Tuning	42.41	0	8.26	76.07±0.57	75.90±1.98	77.12±2.02
Learn from Scratch	42.41	0	8.26	73.20±1.11	73.01±0.84	71.93±2.04
Head-Tuning	0	0	8.26	59.35±0.55	54.14±1.82	71.80±1.98
Ours	0	2.54	8.26	74.83±0.80	74.20±1.50	75.91±2.75

Table 5: Performance on the ADE20K dataset when the backbone network was pre-trained on the COCO-Stuff-164k dataset.

Method	Trainable Parameters (M)			mIoU ↑
	Backbone	Prompt	Head	
Full-Tuning	42.41	0	25.54	43.47
Learn from Scratch	42.41	0	25.54	34.84
Head-Tuning	0	0	25.54	40.21
Ours	0	1.78	25.54	42.11

Effects of Different Prompted Stages:

We further explore the effects of inserting the proposed SPM into varied stages of the backbone. As mentioned above, ResNet-101 consists of 4 stages and we treated the head segmenter as the fifth stage. As shown in Table 6, our performance gradually increases as the number of prompted stages increases. Specifically, our method achieves competitive results by using five SPM on the ImageNet pre-trained model and three SPM on the model pre-trained with COCO-Stuff-164K. Thus, we can conclude that more prompted stages can lead to better results to a certain extent, while the good pre-trained models need fewer prompt learning modules.

Table 6: Performance of different prompted stages on the ADE20K dataset. The iteration number K of SPM is set to 1 in this experiment.

Prompted Stages	Prompted Parameters (M)	Pre-trained Dataset	
		ImageNet	COCO
1	0.48	34.43	40.89
1-2	1.07	35.11	41.74
1-3	1.78	36.39	**42.11**
1-4	2.36	38.54	41.74
1-5	3.11	**40.01**	42.10

Effects of Semantic-aware Prompt Learning:

In the proposed SPM, interim semantic maps are incorporated to learn visual prompts. We ablate this model design by removing the guided semantic information. Specifically, we set all loss weights a_i in Eq.5 to 0, making the network learn prompts in a black-box mapping manner. Table 7 (Row 1) shows that without the semantic-aware learning will lead to inferior results regardless of which pre-trained model is used. This experiment illustrates that the prior information of interim semantic maps is beneficial for effective prompt learning.

Table 7: Performance of our method with/without semantic-aware prompt learning and long-range spatial context modeling on the ADE20K dataset. The iteration number K of SPM is set to 1 in this experiment.

Method	Pre-trained Dataset	
	ImageNet	COCO
w/o semantic-aware	38.08	41.10
w/o long-range context	39.25	41.66
Ours	40.01	42.11

Effects of Long-range Spatial Context: We then explore the effectiveness of long-range spatial context modeling. To this end, we implement a variant of SPM that does not explicitly capture long-range context by setting the dilated rate of all convolutional layers in PDC to 1. As shown in Table 7 (Row 2), our performance drop to 39.25% for ImageNet pre-training and 41.66% for COCO pre-training, when the removing long-range context modeling. This indicates that the long-range spatial context is meaningful for semantic segmentation prompt learning.

Effects of Different Recurrent Iterations of SPM:

We further explore the recurrent prompt learning mechanism of our SPM. As shown in Table 8, our method obtains better results as the recurrent number K increases, when using the backbone pre-trained on ImageNet. This is because that downstream segmentation tasks have a large domain gap with the pre-trained knowledge from ImageNet, and we need more recurrent iterations to learn prompts progressively. When using the COCO pre-trained backbone that has learned rich knowledge of semantic segmentation, our SPM obtains the best results with only one iteration, and more iterations may bring certain performance degradation due to overfitting. The same phenomenon occurs on small/medium-scale datasets (*Please refer to Table 11 in our Appendix*). Therefore, we can draw the following conclusions. **i)** More iterations are required to learn prompts when the downstream task/dataset is quite different from the pre-training task/dataset; otherwise, fewer iterations are required. **ii)** Large-scale downstream datasets require more iterations of prompt tuning, while small/medium-scale datasets require fewer iterations. In the future, we would develop a more flexible mechanism to determine the recurrent iteration number for each task/dataset dynamically.

Table 8: Performance of different recurrent numbers of SPM on the ADE20K dataset.

Recurrent Number	Pre-trained Dataset	
	ImageNet	COCO
1	40.01	**42.11**
2	41.08	41.54
3	**41.83**	41.24

4.4 Apply to Transformer

As mentioned above, our method is also generic to the Transformer architecture. In this subsection, We apply the proposed method to fine-tune the large-scale Vision-Transformer (ViT-L) (Zheng et al., 2021), which consists of 24 transformer layers. Specifically, we divide these layers into N stages evenly and perform visual prompt learning before each stage. Meanwhile, we also explore the recurrent mechanism of SPM based on the backbone ViT-L. Table 10 summarizes the performance of five variants of our method, which are trained for 160k iterations with a batch size of 8. We can see that our method achieves a competitive mIoU 43.49% on the ADE20K dataset, when ViT-L is divided into three stages and each SPM performs twice prompt learning. Thus we adopt this setting for ViT-L prompt tuning, which only introduces 1.76M prompt parameters.

Moreover, we compare our method with some basic and advanced fine-tuning approaches. As shown in Table 10, our method outperforms *head-tuning* with a large margin and is better than VPT (Jia et al., 2022), which learns visual prompts in the token space of each transformer layer. Notice that BIAS (Cai et al., 2020) and VPT+BIAS were trained with a batch size of 16. When using the same batch size, our method (i.e., Ours*) can obtain a better result with fewer prompt parameters. These experiments show the promising potential of our method for large-scale Transformer fine-tuning.

Table 9: Ablation Studies of different stage numbers and SPM recurrent numbers of our method on ADK20K when the backbone is the ViT-L pre-trained on ImageNet.

#Stage	#Recurrent	#Prompt (M)	mIoU ↑
1	1	0.59	42.23
2	1	1.17	42.71
3	1	1.76	43.03
3	2	1.76	**43.49**
3	3	1.76	43.45

Table 10: Performance of different methods on ADE20K when the backbone is ViT-L pre-trained on ImageNet. Those methods with * are trained with a batch size of 16.

Method	Trainable Parameters (M)			mIoU ↑
	Backbone	Prompt	Head	
Full-Tuning	304.15	0	13.14	47.53
Head-Tuning	0	0	13.14	37.77
VPT*	0	0.29	13.14	42.11
BIAS*	0	0.32	13.14	43.40
VPT+BIAS*	0	2.65	13.14	44.04
Ours	0	1.76	13.14	43.49
Ours*	0	1.76	13.14	44.16

5 Conclusion

In this paper, we propose an universal Inter-Stage Prompt-Matched Framework to fine-tune foundation models of CNN/Transformer to handle various downstream tasks. A lightweight Semantic-aware Prompt Matcher is introduced to learn effective visual prompts progressively between different stages of the frozen backbone network, thereby better stimulating pre-trained knowledge and promoting downstream representation learning. We conduct extensive experiments on five datasets of semantic segmentation to verify the performance effectiveness and parameter efficiency of our method. In the future, we would apply the proposed method to more downstream tasks.

REFERENCES

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. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013, 2013.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

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.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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.

Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59 (9):2538–2548, 2012.

Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7519–7528, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 2225–2240, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Yufei Ma, Minkyu Kim, Yu Cao, Sarma Vrudhula, and Jae-sun Seo. End-to-end scalable fpga accelerator for deep residual networks. In *IEEE International Symposium on Circuits and Systems*, pp. 1–4. IEEE, 2017.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272, 2021.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. Msp: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6131–6142, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yang Wang, Yubin Qin, Dazheng Deng, Jingchuan Wei, Yang Zhou, Yuanqi Fan, Tianbao Chen, Hao Sun, Leibo Liu, Shaojun Wei, et al. A 28nm 27.5 tops/w approximate-computing-based transformer processor with asymptotic sparsity speculating and out-of-order computing. In *IEEE International Solid-State Circuits Conference*, volume 65, pp. 1–3. IEEE, 2022.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3081–3089, 2022.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *IEEE International Conference on Computer Vision*, pp. 593–602, 2019.

A Appendix

In Section 4.3 of the main text, we have explored the recurrent prompt learning mechanism of our SPM on the large-scale ADE20K dataset. Here we conduct more ablation studies of this mechanism on small/medium-scale datasets, i.e., Vaihingen, CHASE-DB1, STARE, and HRF. As shown in Table 11, our method obtains the best results with only one iteration on these small/medium-scale datasets, and more iterations may bring certain performance degradation due to overfitting. Considering the results in Tables 8, 9 and 11, we can draw the following conclusions:

- More iterations are required to learn prompts when the downstream task/dataset is quite different from the pre-training task/dataset; otherwise, fewer iterations are required.
- Large-scale downstream datasets require more iterations of prompt visual tuning, while small/medium-scale datasets require fewer iterations.

In the future, we would develop a more flexible recurrent mechanism that can dynamically determine the number of recurrent iterations of our SPM for each dataset and even for each image.

Table 11: Performance of different recurrent numbers of SPM on small/medium-scale datasets. The evaluation metric is mIoU for Vaihingen and Dice for other datasets.

Recurrent Number	Vaihingen	CHASE-DB1	STARE	HRF
1	72.17	74.83±0.80	74.20±1.50	75.91±2.75
2	70.87	72.29±0.97	70.28±1.23	74.07±3.67