Learning Expressive Prompting With Residuals for Vision Transformers

Rajshekhar Das[1,2*], Yonatan Dukler[2], Avinash Ravichandran[2**], Ashwin Swaminathan[2]
Carnegie Mellon University[1] AWS AI Labs[2]
rajshekd@andrew.cmu.edu, dukler@amazon.com, swashwin@amazon.com

Abstract

Prompt learning is an efficient approach to adapt transformers by inserting learnable set of parameters into the input and intermediate representations of a pre-trained model. In this work, we present Expressive Prompts with Residuals (EXPRES) which modifies the prompt learning paradigm specifically for effective adaptation of vision transformers (ViT). Our method constructs downstream representations via learnable "output" tokens (shallow prompts), that are akin to the learned class tokens of the ViT. Further for better steering of the downstream representation processed by the frozen transformer, we introduce residual learnable tokens that are added to the output of various computations. We apply EXPRES for image classification and few-shot semantic segmentation, and show our method is capable of achieving state of the art prompt tuning on 3/3 categories of the VTAB benchmark. In addition to strong performance, we observe that our approach is an order of magnitude more prompt efficient than existing visual prompting baselines. We analytically show the computational benefits of our approach over weight space adaptation techniques like finetuning. Lastly we systematically corroborate the architectural design of our method via a series of ablation experiments.

1. Introduction

Scaling up of neural nets in the past few years has steadily improved performance on wide variety of downstream visual tasks. However, model adaptation is often necessary to achieve the best performance in downstream tasks like fine-grained recognition [89], semantic segmentation [8] or object recognition [34]. While traditional techniques like full-model finetuning have become the de-facto approach to adaptation, they are not well suited for many scenarios. For example, finetuning is susceptible to catastrophic forgetting [37] as it modifies model parameters without the knowledge of future domains, and potentially losing prior knowledge of current adaptation. Moreover, finetuning all of the model parameters of a large vision model with just a few training examples can lead to poor generalization. This is in contrast to human intelligence that is capable of solving wide variety of downstream tasks with extremely few exemplars.

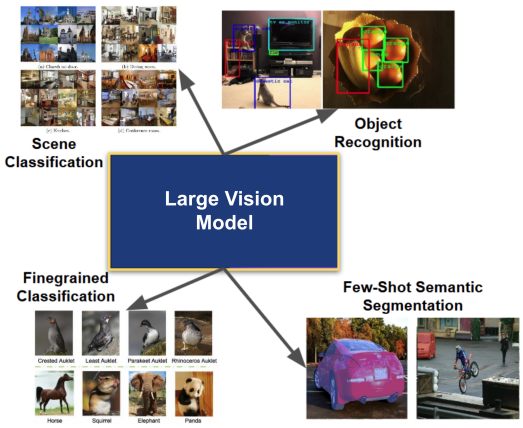

Figure 1. Adapting large vision models is crucial to solving downstream tasks with wide variety of semantics (*e.g.*, image classification, semantic segmentation *etc.*) as well as dataset sizes (few-shot, low-shot, full-shot). In this work, we propose a novel adaptation technique for large vision models that is capable of achieving the desired goal.

Motivated by the need for better adaptation, parameter efficient techniques like partial-finetuning or adapters [64, 101] have been developed to constructively adapt large models without significant parameter overhead. While serving as effective alternatives to finetuning, most parameter efficient techniques have been designed with convolutional architectures in mind. In light of recent works [15] that demonstrate that Vision Transformers are more suitable for scaling up than CNNs, designing adaptation techniques that exploit the Transformer architecture can be extremely useful. To that end, visual prompt tuning (VPT) [32] has been proposed as a way to constructively adapt transformers by introducing learnable tokens at every layer that interact with the patch and class tokens and are optimized together with a classi-

*Work conducted while interning at AWS AI Labs.
**Work conducted while at AWS AI Labs.

fier head. While being effective in practice, VPT allows only partial interactions between prompts and the remaining tokens, thus, leveraging only a part of the prompt capacity. Moreover, it often requires a large number of inserted prompts to achieve optimal performance but that significantly increases the computation costs due to the quadratic computational complexity of the self-attention layer.

In this work, we explore an alternate design to prompting motivated by the potential for greater prompt capacity. We propose ExPRes, an **ex**pressive **p**rompt tuning method with **res**idual tokens that inherits the strengths of parameter efficient adaptation while significantly improving downstream performance. Our prompt design is inspired by the two key observations - propagation of prompts by multilayered interaction with other tokens is crucial for strong capacity and learnable residual tokens can modulate the propagated prompts to favour task-specific relations (unlike in [32]). We first propagate *shallow* prompts through the encoder that are average pooled at the last layer to yield semantic image-level representations. Shallow prompts by themselves have limited capacity since they cannot specifically modulate token-token relations at higher layers. Therefore to harness the prompts, we add residual tokens to propagated prompts at various layerwise computations of the Transformer encoder including LayerNorm, self-attention and multi-head projection to facilitate layerwise modulation without increasing the number of prompts per layer. This results in enhanced prompt capacity at almost no additional computational cost.

We empirically validate the effectiveness of our method on a variety of downstream tasks including fine-grained recognition and semantic segmentation. Our use of additional learnable parameters in the form of residual and shallow prompts allows the retention of prior knowledge in the form of frozen encoder weights while being extremely parameter efficient (prompts are $\leq 1\%$ of the total parameters). Thus, our method is highly suited for real world adaptation that requires information retention at low memory and computational overheads. Additionally, we show that in most cases we require fewer prompts than VPT to achieve the same or better performance, making it more suitable for limited data settings. Our main contributions can be summarized as follows:

- We propose a novel prompting technique: EXPRES, that uses a combination of shallow and deep residual prompts to facilitate constructive adaptation to downstream tasks with limited labelled datasets.
- Our method significantly outperforms full-finetuning based adaptation by 4.6% on VTAB-1k. Moreover, our method outperforms state-of-the-art prompting approach [32] on the same benchmarks with significantly fewer prompts, suggesting that prompt design is crucial to extracting more capacity at a given parameter/computational budget.
- To the best of our knowledge, we are the first to demonstrate the effectiveness of prompting for diverse applications such as few-shot semantic segmentation. Our method outperforms strong adaptation baselines by 25% and achieves competitive performance with respect to language-assisted segmentation [40] despite training on significantly less data with dense annotations.

2. Related work

Large Vision Models: With the advent of Transformer models [80] and adoption to various computer vision tasks, including image classification [15, 51], object detection [7, 44], semantic and panoptic segmentation [76,85,102], video understanding [18, 25, 87] and few-shot learning [13], the scale of vision models have increased by orders of magnitude. Typically trained using large labelled data, either unimodal like ImageNet-21K [65] or multimodal, these models demonstrate superior performance on wide variety of visual tasks. Given their superior performance and much larger scale compared to ConvNets, the question of adaptating such models efficiently becomes crucial. Motivated by the need, our work is primarily focussed on adaptation of Vision Transformers.

Transfer Learning has been extensively studied for vision tasks in the context of ConvNets [105] and many techniques have been introduced including side tuning [101], residual adapter [63], bias tuning [6], *etc*. However, Transformer specific adaptation for visual tasks has received relatively less attention. At the same time, in the NLP domain, the dominance of large-scale pre-trained Transformer-based Large Language Models (LLM) [4, 11, 61], has paved way for many approaches [26, 29, 31] that efficiently fine-tune LLMs for different downstream NLP tasks [82, 83]. In this work we compare with the most representative methods for fair benchmarking. For example, Adapters [30] insert extra lightweight modules inside each Transformer layer. One adapter module generally consists of a linear down-projection, followed by a nonlinear activation function, and a linear up-projection, together with a residual connection [58, 59]. Instead of inserting new modules, [6] proposed to update the bias term and freeze the rest of backbone parameters when fine-tuning ConvNets. BitFit [3] applied this technique to Transformers and verified its effectiveness on LLM tuning. Through our experiments that we demonstrate that our method, EXPRES provides a more effective way of adapting Transformers compared to prior approaches.

Prompting: An alternative to traditional adaptation methods is prompting [48] - originally proposed as a way of

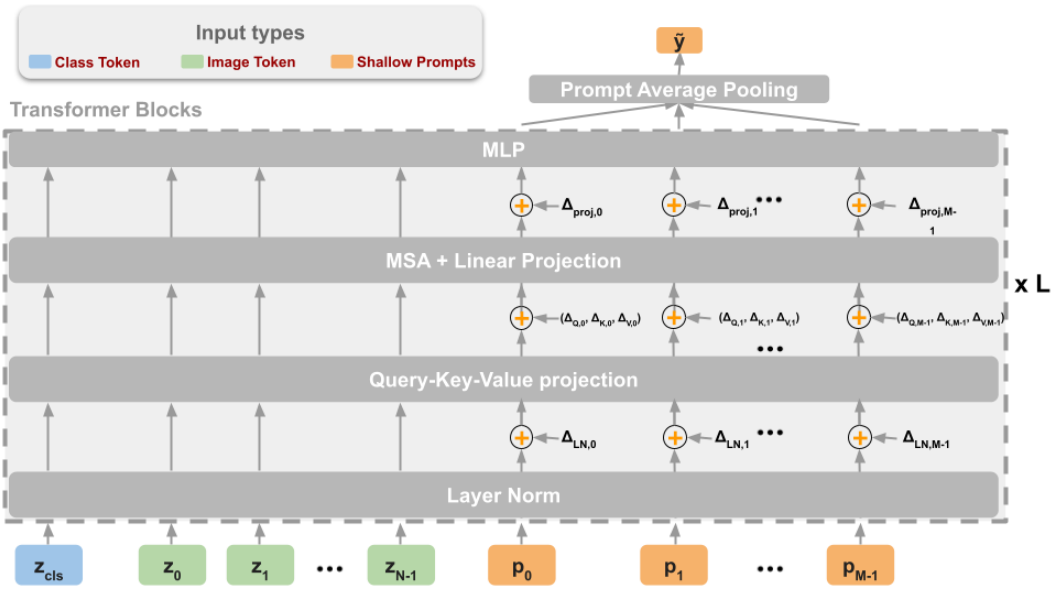

Figure 2. **EXPRES Architecture in detail:** EXPRES optimizes two types of prompts, shallow prompts (*e.g.*, p_i) and residual prompts (*e.g.*, $\nabla_{LN,i}$), to construct task specific representation without updating the pretrained encoder weights. Each residual prompt is a learnable vector that is added to the output of various computations such as Layer Norm, Query-Key-Value projections, and linear projection after the MSA operation.

prepending language instruction to the input text so that a pre-trained LLM can "understand" the task. Through trial and error selection of appropriate prompts, GPT-3 shows strong generalization to downstream transfer learning tasks even in the few-shot or zero-shot settings [4]. This was followed up by other works on better prompt construction [33, 71]. Recent works [39, 43, 50] propose to treat the prompts as task-specific continuous vectors and directly optimize them via gradients during fine-tuning. Such approaches, named "Prompt Tuning" achieve performance comparable to finetuning but with 1000× less parameters in some cases. Following the success in LLMs, prompt have also been adopted for vision-language models [22, 35, 60, 96, 103]. Nonetheless, all the above methods prompt the text encoders and hence are tied to language as input. However, many realistic visual tasks such as dense prediction may not be well aligned with the language modality. Thus, it becomes imperative to develop prompting approaches that can work in the visual modality. To that end, recent work on visual prompting [2, 10, 32, 66, 88] provides encouraging results. In particular, [32] demonstrate that even in the visual domain, adaptation based on continuous prompting can outperform finetuning, especially when the training datasets are small. Our work however shows that current prompting methods do not fully exploit the capacity of prompting for a vision transformer. Through a principled approach to prompting, we derive a more effective prompting technique that achieves state-of-the-art performance at much smaller computational overhead.

Few-Shot Classification: Few-shot classification has received a lot of attention in recent years. While a variety of approaches [75] have been proposed, the most successful ones seek to transfer *positive knowledge* either by finetuning [1, 12, 79] or meta-learning [17, 19, 21, 24, 38, 62, 74, 77, 81]. Finetuning based few-shot learners can be viewed as specialists that perform well on the target domain [9, 27, 79], but suffer from catastrophic forgetting [70] on the base domain. Meta-learning approaches, on the other hand, can be seen as generalists that enjoy complete immunity against forgetting but at the cost of somewhat lower performance in the target domain. In this work, we propose EXPRES as a constructive adaptation technique that is immune to forgetting but at the same time benefits from task specific parameter tuning, thus, leveraging the best of both worlds. One of our key contribution is to demonstrate the transferability of Transformers from classification to dense prediction tasks (semantic segmentation) when adapted with EXPRES.

Few-Shot Semantic Segmentation: This task was originally proposed in [68]. Most works after that follow the metric learning paradigm [14] with various novelties from improved support-query matching [47, 72, 95] to better optimization [46, 104], memory modules [90, 93], graph neural networks [84, 92, 99], and more [28, 42, 45, 52, 53, 78, 98, 106]. Some methods generate representative support prototypes with attention mechanism [20, 100], adaptive prototype learning [41, 57, 73], or various prototype generation

techniques [49, 54, 86, 94]. In contrast, our approach directly adapts a model pretrained on image classification to few-shot semantic segmentation without intermediate pretraining on densely-annotated datasets.

3. Approach

3.1. Preliminaries

Consider an input space $\mathcal{X}$ and a categorical label set $\mathcal{Y}$ where each class is represented via a one-hot encoding. A representation $\mathcal{R} \subset \mathbb{R}^d$ of the input is defined by the composition of an augmentation function $\mathcal{A} : \mathcal{X} \rightarrow \mathcal{X}$ and an encoder model $E_\theta : \mathcal{X} \rightarrow \mathcal{R}$, parameterized by θ. The augmentation function is a composition of standard image transformations such as random cropping, horizontal flipping *etc*. A general recognition task $\mathcal{T}$ can be defined in terms of a finite training set, $\{(x_i, y_i)\}_{i \in \mathcal{D}_{\text{train}}}$ and N_c categories such that, $y_i \in [N_c]$. The goal, is to leverage a pretrained encoder, E_{θ^*} and the training set to obtain a classifier for the task $\mathcal{T}$.

3.2. Overview of Vision Transformers (ViT)

In a typical ViT, the image, $x \in \mathbb{R}^{h \times w \times 3}$ is uniformly divided into N fixed-sized patches, each of which are projected to a d-dimensional embedding and are added a positional embedding, resulting in a patch token $z_n \in \mathbb{R}^d$. Additionally, a class token $z_{\text{cls}} \in \mathbb{R}^d$ is concatenated to the sequence of the patch tokens to form the input, $Z^0 \in \mathbb{R}^{(N+1) \times d}$. Starting from layer $l = 0$, the incoming activations at each layer, z^{l-1} are first normalized using LayerNorm (LN), and then processed by a multi-headed self-attention block (MSA) followed by an MLP block. At the last layer, the resulting class token is normalized to yield the final representation. The overall computations in a ViT encoder can be summarised as

$$
\begin{aligned}
Z^0 &= [z_{\text{cls}}, z_0, \ldots, z_{N-1}] \\
H^{l-1} &= \text{MSA}(\text{LN}(Z^{l-1})) + Z^{l-1} \quad l = 0, ..., L-1 \\
Z^l &= \text{MLP}(\text{LN}(H^{l-1})) + H^{l-1} \quad l = 0, ..., L-1 \\
y &= LN(Z^L_{\text{cls}})
\end{aligned}
$$

where L represents the number of encoder layers. The multi-headed self-attention mechanism (MSA) abstracts patchwise representation by aggregating the right context at each layer. The context aggregation is facilitated by softmax attention that relies on patch-to-patch similarities and its parameters are governed by the pre-training task objective. During adaptation, however, the representations for downstream task might benefit from aggregating a slightly different context at each layer. Our prompt tuning approach facilitates task specific modulation of this context by augmenting and learning layerwise patch tokens.

3.3. Expressive Prompt Tuning

At the input layer of the ViT, we introduce prompt tokens, $P^0 \in \mathbb{R}^{M \times d}$ which are M parameterized vectors of dimension d concatenated to the input token sequence, Z^0, of the ViT. Input-level prompts, also referred to as *shallow* prompts, are propagated through the encoder together with the class and patch tokens such that at every layer, each token interacts with every other token through the self-attention layers. The propagated prompts at the last layer are average-pooled to obtain the final representation. The insertion and propagation of the prompts described above can be expressed as

$$
\begin{aligned}
\tilde{Z}^0 &= [Z^0 || P^0] \\
\tilde{H}^{l-1} &= \text{MSA}(\text{LN}(\tilde{Z}^{l-1})) + \tilde{Z}^{l-1} \quad l = 0, ..., L-1 \\
\tilde{Z}^l &= \text{MLP}(\text{LN}(\tilde{H}^{l-1})) + \tilde{H}^{l-1} \quad l = 0, ..., L-1 \\
Z^L, P^L &= \texttt{chunk}(\tilde{Z}^l) \\
\tilde{y} &= \text{AvgPool}(P^L) \\
y &= LN(\tilde{y})
\end{aligned} \tag{1}
$$

we use $||$ to denote the concatenation along the sequence axis and `chunk` to denote the splitting of the propagated sequence of length $N+M+1$ into the $N+1$ tokens and M prompts. Shallow prompts are capable of modeling some of the desired token relations for a task. However, they have limited capacity due to the inability to alter for specific per-layer interactions with the class and image tokens.

To enhance the prompt capacity, we introduce layer-wise residual prompts, $\mathbf{\Delta}^l \in \mathbb{R}^{M \times d}$ that are added to the propagated prompts at various computations within the MSA block for an intermediate layer l. This includes the output of attention LayerNorm, query-key-value projections and linear multi-head projection. We summarise the MSA computations at layer l with N_h heads and input Z

$$
\begin{aligned}
Z' &= LN(Z) \\
Q_h &= Z' W^Q_h;\ K_h = Z' W^K_h;\ V_h = Z' W^V_h \\
\text{g}_h &= \texttt{Att}(Q_h, K_h, V_h) \quad i = 1, ..., N_h \\
O &= [\text{g}_1 || \ldots || \text{g}_{N_h}] W^{\text{proj}}
\end{aligned} \tag{2}
$$

above h is used to represent the head index of W^Q, W^K, W^V and W^{proj} denotes the projection matrix of the MSA block. The residually prompted computations can then be expressed as

$$
\begin{aligned}
Z' &= LN(Z) + [\bar{\mathbf{0}} || \mathbf{\Delta_{LN}}] \\
\tilde{Q}_h &= Q_h + [\bar{\mathbf{0}} || \mathbf{\Delta_Q}] \\
\tilde{K}_h &= K_h + [\bar{\mathbf{0}} || \mathbf{\Delta_K}] \\
\tilde{V}_h &= V_h + [\bar{\mathbf{0}} || \mathbf{\Delta_V}] \\
\tilde{O} &= O + [\bar{\mathbf{0}} || \mathbf{\Delta_{proj}}]
\end{aligned}
$$

where the concatenation of $(N+1) \times d$ dimensional zero matrix, $\bar{\mathbf{0}}$ with the residual-prompts, signifies that the residuals $\mathbf{\Delta}$ are added only at the propagated prompt positions and not to the image patch or class token positions. The above residual computations occur at every layer with independently learned residual prompts, and we drop the layer indices in the equations for brevity. The overall EXPRES architecture is visualized in Figure 2.

3.4. Interpreting Residual Prompting

We take a closer look at residual prompting for self-attention and interpret it's functionality. The `Att` operation in Eq. (2) facilitates weighted aggregation over all "value" tokens where the weights are computed using the "query" and a "key" tokens as $w_{ij} \propto \exp\left(\frac{q_i^T k_j}{\sqrt{d/N_h}}\right)$. When prompted with residual tokens, the expression for the weights that aggregate information for a patch token q_i can be factored as,

$$\begin{aligned}\tilde{w}_{ij} &\propto \exp\left(\frac{q_i^T(k_j + \mathbf{\Delta}_{K,j})}{\sqrt{d/N_h}}\right) \\ &= \exp\left(\frac{q_i^T k_i}{\sqrt{d/N_h}}\right)\exp\left(\frac{q_i^T \mathbf{\Delta}_{K,j}}{\sqrt{d/N_h}}\right) \\ &= w_{ij} * \alpha_{ij}.\end{aligned} \tag{3}$$

Based on Eq. (3), residual prompts facilitate task-specific reweighting of the attention weights independently at every layer, allowing the modulation of context aggregated per patch token. Such layerwise modulation can lead to better adaptation of the final representation compared to shallow prompting that restricts the modulation to the input layer. Moreover, the two-way interaction between prompts and patch tokens leads to greater flexibility than other forms of multilayered prompting [32, 67] that allow for only partial interaction. For instance, [32] restrict the prompts to act only as keys and never as queries.

Residual prompt based attention reweighting is also interesting from parameter efficiency perspective as it circumvents the need for updating the attention weight matrices (done in finetuning approaches) to achieve the same goal of task specific adaptation. Specifically, each layer of a ViT consists of three attention weight matrices, each with $d \times d$ dimensions, resulting in $\mathcal{O}(d^2)$ learnable parameters. In contrast, each prompt is d dimensional so, only $\mathcal{O}(d)$ parameters need to be adapted even with M prompts, where $M << d$. We empirically validate that the high capacity as well as parameter efficiency of our prompting approach is crucial for achieving good adaptation performance in limited labelled data settings.

3.5. Learning the Prompts

In this work, we are mainly interested in two types of downstream tasks - image classification and semantic segmentation. To train the prompts for classification, we optimize a standard cross entropy loss with respect to the representation, y in Eq. (1) and the corresponding ground-truth label, $y*$. In the case of semantic segmentation, we adapt the model as well as the objective to perform dense predictions. At the final layer of the encoder, we extract the keys corresponding to the patch tokens and pass them through the classifier. The sequence outputs are then reshaped into a 2d map and resized to original image resolution using bilinear interpolation, resulting in a pixel wise prediction. Finally, to optimize the prompts with the classifier head, we use a dense cross entropy loss as follows

$$\{p_m^*\}, \{\mathbf{\Delta}^*\} = \underset{\{p_m\},\{\mathbf{\Delta}\},C}{\arg\min} \sum_{j \in I_{\text{ctxt}}} \sum_{h,w} L_{\text{CE}}(y_{jhw}, y^*_{jhw})$$

where, h, w are used to index the spatial positions at the resolution, $H \times W$ of the input image.

4. Experiments

We validate the effectiveness of EXPRES on a variety of benchmarks consisting of wide variety of tasks and dataset sizes. We also analyse the importance of various model components and design decisions.

Datasets: To evaluate EXPRES, we use two different benchmarks, VTAB-1k [97] and FGVC [32]. The **VTAB-1k** benchmark consists of 19 different visual classification tasks categorized under three groups: *Natural* - tasks with natural images captured with standard cameras; *Specialized* - tasks with images captured under specialized settings (medical and satellite imagery); and *Structured* - tasks that requires understanding scene geometry, like object distance. Each task-specific dataset contains 1000 training examples with varying number of samples per class, depending on the number of classes. For validation purposes and hyperparameter selection, we use a $800-200$ split of the training set and then train on all 1000 examples for final results, which are based on evaluation on the entire test set. The **FGVC** benchmark consists of the finegrained datasets including CUB [89], Oxford Flowers [55], Stanford Dogs [36] and Stanford Cars [23]. In conjunction with VTAB-1k, we use the FGVC datasets to conduct key ablation studies for EXPRES. A random $90-10$ split of each dataset is used for hyperparameter selection. For few shot segmentation, we use the standard **PASCAL** $-\mathbf{5^i}$ [69] benchmark that was created from PASCAL VOC 2012 [16] with extra mask annotations for 20 object classes, evenly divided into 4 folds: $\{5^i : i \in \{0,1,2,3\}\}$. Following prior works for evaluation scheme, we randomly sample 1000 episodes per fold and report the average performance over all episodes

	natural								specialized					structured								
	CIFAR-100	**Caltech101**	**DTD**	**Flowers102**	**Pets**	**SVHN**	**Sun397**	**Mean**	**Patch Camelyon**	**EuroSAT**	**Resisc45**	**Retinopathy**	**Mean**	**Clevr/count**	**Clevr/distance**	**DMLab**	**KITTI/distance**	**dSprites/location**	**dSprites/orientation**	**SmallNORB/azimuth**	**SmallNORB/elevation**	**Mean**
Linear	63.4	85.0	63.2	97.0	86.3	36.6	51.0	68.93	78.5	87.5	68.6	74.0	77.16	34.3	30.6	33.2	55.4	12.5	20.0	9.6	19.2	26.84
MLP-2	63.2	84.8	60.5	97.6	85.9	34.1	47.8	67.70	74.3	88.8	67.1	73.2	75.86	45.2	31.6	31.8	55.7	30.9	24.6	16.6	23.3	32.47
MLP-3	63.8	84.7	62.3	97.4	84.7	32.5	49.2	67.80	77.0	88.0	70.2	56.1	72.83	47.8	32.8	32.3	58.1	12.9	21.2	15.2	24.8	30.62
MLP-5	59.3	84.4	59.9	96.1	84.4	30.9	46.8	65.98	73.7	87.2	64.8	71.5	74.31	50.8	32.3	31.5	56.4	7.5	20.8	14.4	20.4	29.23
MLP-9	53.1	80.5	53.9	95.1	82.6	24.4	43.7	61.90	78.5	83.0	60.2	72.3	73.49	47.5	27.9	28.9	54.0	6.2	17.7	10.8	16.2	26.15
Sidetune [101]	60.7	60.8	53.6	95.5	66.7	34.9	35.3	58.21	58.5	87.7	65.2	61.0	68.12	27.6	22.6	31.3	51.7	8.2	14.4	9.8	21.8	23.41
Biastune [6]	72.8	87.0	59.2	97.5	85.3	59.9	51.4	73.30	78.7	91.6	72.9	69.8	78.25	61.5	55.6	32.4	55.9	66.6	40.0	15.7	25.1	44.09
Adapter-256	74.1	86.1	63.2	97.7	87.0	34.6	50.8	70.50	76.3	88.0	73.1	70.5	76.98	45.7	37.4	31.2	53.2	30.3	25.4	13.8	22.1	32.39
Adapter-64	74.2	85.8	62.7	97.6	87.2	36.3	50.9	70.65	76.3	87.5	73.7	70.9	77.10	42.9	39.9	30.4	54.5	31.9	25.6	13.5	21.4	32.51
Adapter-8	74.2	85.7	62.7	97.8	87.2	36.4	50.7	70.67	76.9	89.2	73.5	71.6	77.80	45.2	41.8	31.1	56.4	30.4	24.6	13.2	22.0	33.09
Partial-1	66.8	85.9	62.5	97.3	85.5	37.6	50.6	69.44	78.6	89.8	72.5	73.3	78.53	41.5	34.3	33.9	61.0	31.3	32.8	16.3	22.4	34.17
FT-all	68.9	87.7	64.3	97.2	86.9	**87.4**	38.8	75.88	79.7	95.7	**84.2**	73.9	83.36	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1	47.64
VPT-shallow [32]	77.7	86.9	62.6	97.5	87.3	74.5	51.2	76.81	78.2	92.0	75.6	72.9	79.66	50.5	58.6	40.5	67.1	68.7	36.1	20.2	34.1	46.98
VPT-deep [32]	**78.8**	**90.8**	65.8	98.0	88.3	78.1	49.6	78.48	81.8	96.1	83.4	68.4	82.43	**68.5**	60.0	**46.5**	72.8	73.6	47.9	**32.9**	**37.8**	54.98
EXPRES (*ours*)	78.0	89.6	**68.8**	**98.7**	**88.9**	81.9	**51.9**	**79.7**	**84.8**	**96.2**	80.9	**74.2**	**84.0**	66.5	**60.4**	**46.5**	**77.6**	**78.0**	**49.5**	26.1	35.3	**55.0**

Table 1. **VTAB-1k benchmark:** Per task adaptation results with ViT-B/16 model pretrained on ImageNet-21k..

Method	Im-Pre	Dense-Pre	Backbone	5^0	5^1	5^2	5^3
SPNET [91]	In-1k	✓	ResNet-101	23.8	17.0	14.1	18.3
ZS3Net [5]	In-1k	✓	ResNet-101	40.8	39.4	39.3	33.6
LSEG [40]	In-1k	✓	ViT-L/16	**61.3**	**63.6**	43.1	41.0
Linear	In-21k	-	ViT-B/16	5.1	43.9	28.1	29.1
FT-all	In-21k	-	ViT-B/16	18.3	31.9	27.9	23.2
Biastune	In-21k	-	ViT-B/16	2.4	39.2	13.9	20.1
VPT-deep [32]	In-21k	-	ViT-B/16	33.4	56.2	49.8	47.7
EXPRES (***Ours***)	In-21k	-	ViT-B/16	41.8	60.2	**52.4**	**51.4**

Table 2. **Five-Shot Semantic Segmentation on PASCAL -5^i:** Per fold adaptation results with ViT-B/16 model pretrained on ImageNet-21k.

of the corresponding fold. However, unlike prior works, we tune our EXPRES model on few-shot training examples and cross-validate the hyperparameters (*e.g.* number of prompts, M) on a validation set with fold-exclusive categories. In all our experiments, we use a reasonably small budget for hyperparameters (see supplementary) following recent studies [56] that highlight the likelihood of overoptimistic results in limited-labelled data settings due to excessive hyperparameter tuning on large validation sets.

Implementation Details: In all our experiments, we use a fixed encoder, ViT-B/16 pretrained on ImageNet-22K [65]. This model is effective on wide variety of tasks and allows direct comparison with prior works. For each downstream task, we train for a total 100 epochs with an initial warmup of 10 epochs. We use AdamW as our default optimizer with a suitable learning rate and fixed weight decay of $1e{-}4$. For VTAB-1k and FGVC experiments, we use a fixed batch size of 64. For segmentation experiments, since overall training set sizes are extremely small (≤ 10 images in total, we use the entire training set per batch. We use input image resolution of 224×224 for classification tasks and 384×384 for semantic segmentation tasks as dense prediction usually benefits from larger resolution. Since, ViT-B/16 is pretrained on standard 224×224 resolution images, we use interpolated positional embeddings to accommodate larger resolutions in the case of segmentation. Finally, we use standard data augmentations for classification benchmarks *i.e.*, `Resize` $\rightarrow$ `Random-Crop` $\rightarrow$ `Horizontal-Flip` during training and `Resize` $\rightarrow$ `Center-Crop` during evaluation while for segmentation we only use `Resize`.

Evaluation Metrics For classification experiments, we use accuracy as our performance metric. For evaluating segmentation masks, we use mean intersection over union (mIoU) that averages over the intersection over union curves per class.

4.1. Main Results

We compare EXPRES with a number of commonly used adaptation techniques on VTAB-1k. The adaptation methods can be categorized as *head-oriented*, *backbone-oriented*

and *prompt-based*. Under the first category, *Linear* only optimizes a linear classifier for the downstream task while *MLP-k* uses a k-layer multilayer perceptron (MLP) as the classifier head. As an example of the second category, *Sidetune* [101] uses features that are linearly interpolated between pretrained features and features from a "side" network trained on downstream data. While *Biastune* [6] adapts only the bias terms of an otherwise frozen backbone. *Adapter-d* [30, 58, 59] introduces lightweight MLP modules inserted between Transformer layers. *Partial-k* finetunes the last *k* layers while keeping the rest of the backbone frozen, finally *FT-all* finetunes all the layers. *VPT-shallow* [32] additionally optimizes a linear classifier with the learnable input prompts propagated through the encoder. *VPT-deep* [32] introduces more capacity by replacing the propagated prompts with a new set of learnable prompts at each layer.

VTAB-1k Results (Table 1): Across all three splits of VTAB-1k, our method significantly outperforms the best *head-oriented* techniques: $+11\%$ (*natural*), $+7\%$ (*specialized*) and $+23\%$ (*structured*). Similar trends hold even when comparing with the more powerful class of *backbone-oriented* techniques. While *FT-all* has been widely adopted as an effective technique for most adaptation scenarios, our method consistently outperforms it by a significant margin of $+4\%$ (*natural*), $+1\%$ (*specialized*) and $+7\%$ (*structured*). Most interestingly our method even outperforms other powerful prompting techniques like *VPT-deep* on 12 (out of 19) datasets by a margin of about $+1\%$ (*natural*), $+2\%$ (*specialized*) and $+0.02\%$ (*structured*). The performance gains are particularly impressive when considered from the perspective of computation vs performance tradeoff. Compared to *VPT-deep* that uses 53 prompts for *natural* and 108 prompts for structures, our method only requires 10 and 29 prompts for the respective splits. A more detailed summary is provided in the supplementary.

4.2. Few-Shot Semantic Segmentation Without Dense Pretraining

We test the efficacy of EXPRES on a novel adaptation setup where the backbone, pretrained on classification tasks, is directly adapted for few-shot semantic segmentation task without additional training on large densely-annotated datasets. This is in contrast with most works that follow a two-stage pretraining procedure *i.e.*, training on ImageNet-1k with image level labels followed by meta-learning on densely annotated datasets constructed from the train folds of PASCAL -5^i. In contrast, we only perform the first stage pretraining on a sufficiently diverse dataset (ImageNet-21k in our case) with no meta-learning stage. During evaluation, each few-shot episode randomly samples a target image with a segmentation mask that assigns a label 1 to the pixels corresponding to one of the object categories in that image and a label 0 to remaining pixels (background). From the same object category, a set of five images are randomly sampled to form the training set where the mask of each image is annotated in the same way as the target. Our method as well as all baselines that perform meta-learning are optimized given a context set to yield a binary classifier for the target image. For dense representations, we use the last layer *keys* of the MSA-block as we found them to be more accurate than the typical MLP-block output. We observe that our method consistently outperforms the baselines (by 25%) as well as other prompt techniques such as VPT-deep (by 5%). Surprisingly, our method even outperforms [40] that leverages densely annotated datasets (training data per fold) with large language models for a second stage of pretraining. Specifically, on 2/4 folds, EXPRES outperforms [40] by 10% despite training on a significantly smaller densely-annotated dataset (five images). These results are particularly significant as they suggest that highly competitive results can be achieved with models pretrained on image classification. Consequently, results may be further improved by scaling up the image-level annotated datasets which are cheaper to scale than the densely annotated datasets. In the supplementary, we provide visualizations of the segmentation mask predictions by our method as well as Linear, FT-all and VPT-deep methods.

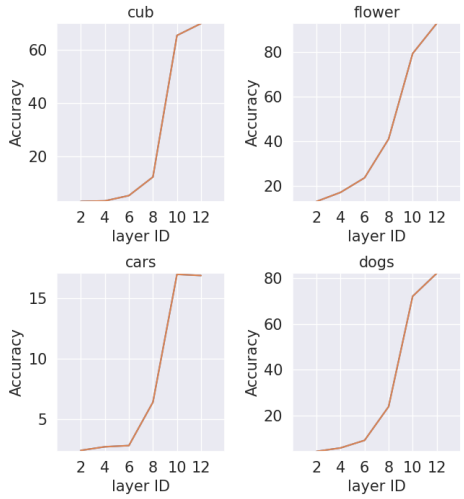

Figure 3. **Prompt Propagation:** Effect of propagating prompts with modulation upto a layer, $l = \{2, \ldots, 12\}$ of the ViT-B/16 encoder with total 12 layers. The datasets are sampled from the FGVC benchmark.

4.3. Ablation Studies

We conduct extensive ablations to evaluate key design decisions used to develop EXPRES such as feature construction, residual prompting, number of prompts *etc.* For all ablations, we use the same ViT-B/16 backbone. When

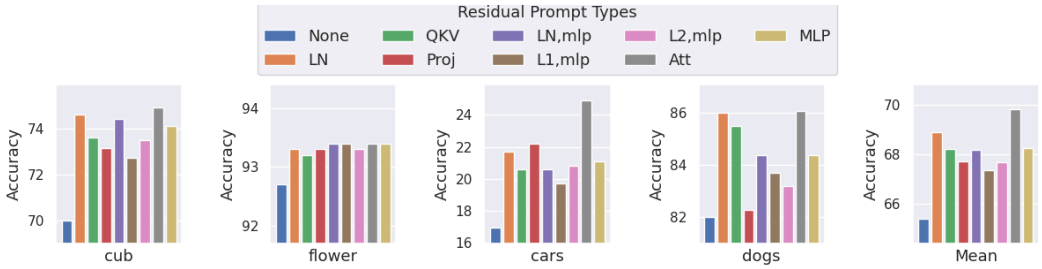

Figure 4. **Residual Prompt Types:** Evaluating the importance of different type of residual prompts on FGVC datasets.

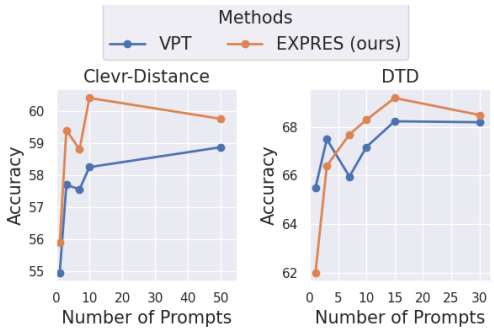

Figure 5. **Computational Efficency Plots:** Comparing Accuracy vs Number of Prompts for VPT and EXPRES

using FGVC datasets, we use only 10% of each dataset with official splits provided in [32].

Prompt Propagation: We evaluate the importance of propagating prompts through the ViT-B/16 encoder with layer-wise modulation. Specifically, upto a layer l, we allow all tokens (patch, class and prompt) to attend to each other at every layer. Beyond l, prompts are not allowed to interact with other tokens at all; they are simply projected by the value heads of MSA-block followed by MLP processing at every layer. In our experiments, we fix the number of prompts to 10 irrespective of the dataset. From Figure 3, we observe that downstream performance depends directly on the extent of prompt propagation with modulation: accuracy improves as more layers allow prompts to interact with other tokens. These results motivate our method, EXPRES, that facilitates finegrained layerwise modulation via residual tokens.

Residual Prompt Type: In Figure 4 we evaluate the importance of different residual prompt types (one-at-a-time) including attention, LayerNorm (*LN*), query-key-value projections (*QKV*) and linear multi-head projection (*Proj*). We also evaluate residual prompting in the MLP block: after LayerNorm (*LN,mlp*), first linear projection (*L1,mlp*) and second linear projection (*L2,mlp*). We evaluate the composite effect of multiple prompt types within a computational block *i.e.*, *Att* (MSA block) and *MLP* (MLP block) As a baseline, we provide the per-dataset results for shallow prompting, referred to as *None* in the figure. The number of prompts at each layer are fixed at 10. We observe that, within the MSA block, adding residual prompts to LayerNorm and query-key-value projections yields the most improvements (3.5% and 2.8% on an average respectively) over *None*. Moreover, LayerNorm prompts in the MLP block are also more effective than prompting the two linear layers. Comparing blockwise performance, prompting the MSA block (*Att*) yields significantly better performance (by 1.6%) than prompting the MLP block (*MLP*). The performance gap between *Att* and *MLP* highlights the importance of directly modulating layerwise interaction between tokens for better adaptation. Consequently, we use *Att* as our default setting for all experiments.

Computational Efficiency of Prompting: In prompting techniques, the primary computational overhead arises from the quadratic complexity (in number of tokens) of the transformer encoder. So to evaluate computation efficiency of prompting, we investigate the rate at which performance improves with number of prompts and provide comparisons between our method and VPT in Figure 5. We evaluate on two different datasets sampled from different categories of VTAB-1k. For a given accuracy, *e.g.* 58.5% on Clevr-Distance, EXPRES requires an order less prompts than VPT, resulting in 2 orders less computations. Overall EXPRES achieves higher optimal performance with far fewer prompts than VPT.

5. Conclusion

In this work we propose a novel prompting technique for adapting large vision models. Our method demonstrates strong performance across variety of downstream tasks with varying dataset sizes. Further, our method outperforms commonly used finetuning approach as well as the recently proposed VPT method on standard benchmarks. We also demonstrate diverse adaptation ability of our method from classification to semantic segmentation tasks in the few-shot setting. Lastly, our method is more parameter efficient that existing weight-space and prompt based adaptation techniques. In the future, we plan to extend our method to additional settings including vision-language learning.

References

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagn'e. Associative alignment for few-shot image classification. In *ECCV*, 2020. 3

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. 2022. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022. 2

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *NeurIPS*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 2, 3

[5] Maxime Bucher, Tuan-Hung VU, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6

[6] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. *NeurIPS*, 33:11285–11297, 2020. 2, 6, 7

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[8] 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. *CoRR*, abs/1606.00915, 2016. 1

[9] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 3

[10] Jonathan Conder, Josephine Jefferson, Khurram Jawed, Alireza Nejati, Mark Sagar, et al. Efficient transfer learning for visual tasks via continuous optimization of prompts. In *International Conference on Image Analysis and Processing*, pages 297–309. Springer, 2022. 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019. 2

[12] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *ICLR*, 2020. 3

[13] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In *NeurIPS*, 2020. 2

[14] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018. 3

[15] 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. In *ICLR*, 2020. 1, 2

[16] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2014. 5

[17] Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang. {MELR}: Meta-learning via modeling episode-level relationships for few-shot learning. In *ICLR*, 2021. 3

[18] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 2

[19] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 3

[20] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, and Balaji Krishnamurthy. Simpropnet: Improved similarity propagation for few-shot image segmentation. In *IJCAI*, 2020. 3

[21] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017. 3

[22] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 3

[23] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17. AAAI Press, 2017. 5

[24] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. *CVPR*, 2018. 3

[25] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, pages 244–253, 2019. 2

[26] Demi Guo, Alexander Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4884–4896, Online, Aug. 2021. Association for Computational Linguistics. 2

[27] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. ECCV, 2020. 3

[28] Haoyu He, Jing Zhang, Bhavani Thuraisingham, and Dacheng Tao. Progressive one-shot human parsing. In *AAAI*, 2021. 3

[29] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *ICLR*, 2022. 2

[30] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799. PMLR, 2019. 2, 7

[31] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[32] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7, 8

[33] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. 3

[34] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE Access*, 7:128837–128868, 2019. 1

[35] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 3

[36] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011. 5

[37] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 1

[38] Namyeong Kwon, Hwidong Na, Gabriel Huang, and Simon Lacoste-Julien. Repurposing pretrained models for robust out-of-domain few-shot learning. In *ICLR*, 2021. 3

[39] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 3

[40] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2, 6, 7

[41] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *CVPR*, 2021. 3

[42] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, 2020. 3

[43] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, Aug. 2021. Association for Computational Linguistics. 3

[44] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 2

[45] Binghao Liu, Yao Ding, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Anti-aliasing semantic reconstruction for few-shot semantic segmentation. In *CVPR*, 2021. 3

[46] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *AAAI*, 2021. 3

[47] Lizhao Liu, Junyi Cao, Minqian Liu, Yong Guo, Qi Chen, and Mingkui Tan. Dynamic extension nets for few-shot semantic segmentation. In *ACM MM*, 2020. 3

[48] 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. 2

[49] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. In *CVPR*, 2020. 4

[50] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 3

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[52] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*, 2021. 3

[53] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*, 2021. 3

[54] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019. 4

[55] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics 'I&' Image Processing*, pages 722–729, 2008. 5

[56] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv*, 2018. 6

[57] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *ECCV*, 2020. 3

[58] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 2, 7

[59] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online, 2020. Association for Computational Linguistics. 2, 7

[60] 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*, pages 8748–8763. PMLR, 2021. 3

[61] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 2

[62] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 3

[63] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *NeurIPS*, 30, 2017. 2

[64] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *CVPR*, pages 8119–8127, 2018. 1

[65] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 2, 6

[66] Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Andrew Jackson. Fine-tuning image transformers using learnable memory. In *CVPR*, pages 12155–12164, 2022. 3

[67] Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Andrew Jackson. Fine-tuning image transformers using learnable memory. *CVPR*, 2022. 5

[68] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017. 3

[69] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *Procedings of the British Machine Vision Conference 2017*, 2017. 5

[70] Guangyuan SHI, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In *NeurIPS*, 2021. 3

[71] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 3

[72] Mennatullah Siam, Naren Doraiswamy, Boris N Oreshkin, Hengshuai Yao, and Martin Jagersand. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings. In *IJCAI*, 2020. 3

[73] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *ICCV*, 2019. 3

[74] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 3

[75] Yisheng Song, Ting Wang, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities, 2022. 3

[76] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *CVPR*, pages 7262–7272, 2021. 2

[77] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 3

[78] Pinzhuo Tian, Zhangkai Wu, Lei Qi, Lei Wang, Yinghuan Shi, and Yang Gao. Differentiable meta-learning model for few-shot semantic segmentation. In *AAAI*, 2020. 3

[79] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? *arXiv*, 2020. 3

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[81] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*. 2016. 3

[82] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *NeurIPS*, 32, 2019. 2

[83] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. 2

[84] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic seg-

mentation with democratic attention networks. In *ECCV*, 2020. 3

[85] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, pages 5463–5474, 2021. 2

[86] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019. 4

[87] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, pages 14733–14743, 2022. 2

[88] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022. 3

[89] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1, 5

[90] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *ICCV*, 2021. 3

[91] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8257, 2019. 6

[92] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *CVPR*, 2021. 3

[93] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *ICCV*, 2021. 3

[94] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *ECCV*, 2020. 4

[95] Xianghui Yang, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. In *BMVC*, 2020. 3

[96] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 3

[97] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2019. 5

[98] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *CVPR*, 2021. 3

[99] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *ICCV*, 2019. 3

[100] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *CVPR*, 2019. 3

[101] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *ECCV*, pages 698–714. Springer, 2020. 1, 2, 6, 7

[102] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 2

[103] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 3

[104] Kai Zhu, Wei Zhai, Zheng-Jun Zha, and Yang Cao. Self-supervised tuning for few-shot segmentation. In *IJCAI*, 2020. 3

[105] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2

[106] Yunzhi Zhuge and Chunhua Shen. Deep reasoning network for few-shot semantic segmentation. In *ACM MM*, 2021. 3