

## Task Residual for Tuning Vision-Language Models

Tao Yu<sup>1,2\*</sup> Zhihe Lu<sup>1\*</sup> Xin Jin<sup>1,2</sup> Zhibo Chen<sup>2</sup> Xinchao Wang<sup>1†</sup>

<sup>1</sup>National University of Singapore <sup>2</sup>University of Science and Technology of China

yutao666@mail.ustc.edu.cn, zhihelu@nus.edu.sg, jinxustc@mail.ustc.edu.cn,  
chenzhibo@ustc.edu.cn, xinchao@nus.edu.sg

### Abstract

Large-scale vision-language models (VLMs) pre-trained on billion-level data have learned general visual representations and broad visual concepts. In principle, the well-learned knowledge structure of the VLMs should be inherited appropriately when being transferred to downstream tasks with limited data. However, most existing efficient transfer learning (ETL) approaches for VLMs either damage or are excessively biased towards the prior knowledge, e.g., prompt tuning (PT) discards the pre-trained text-based classifier and builds a new one while adapter-style tuning (AT) fully relies on the pre-trained features. To address this, we propose a new efficient tuning approach for VLMs named Task Residual Tuning (TaskRes), which performs directly on the text-based classifier and explicitly decouples the prior knowledge of the pre-trained models and new knowledge regarding a target task. Specifically, TaskRes keeps the original classifier weights from the VLMs frozen and obtains a new classifier for the target task by tuning a set of prior-independent parameters as a residual to the original one, which enables reliable prior knowledge preservation and flexible task-specific knowledge exploration. The proposed TaskRes is simple yet effective, which significantly outperforms previous ETL methods (e.g., PT and AT) on 11 benchmark datasets while requiring minimal effort for the implementation. Our code is available at <https://github.com/geekyutao/TaskRes>.

### 1. Introduction

Over the past decade, deep learning-based visual recognition models [10, 18, 28, 55, 58] have achieved great success. These state-of-the-art models are often trained on a large amount of image and discrete label pairs. The discrete label is generated by converting a detailed textual description, e.g., “American curl cat”, into a simple scalar, which strikingly eases the computation of loss function. However, this also results in two evident limitations: (i) the rich semantics

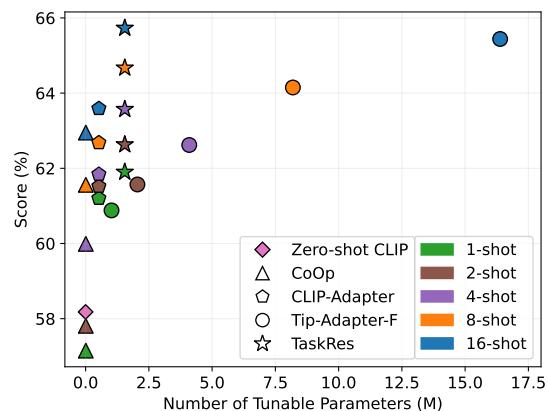


Figure 1. Performance comparison between Zero-shot CLIP [49], CoOp [73], CLIP-Adapter [15], Tip-Adapter-F [71] and our TaskRes on ImageNet with few shot settings.

in the textual description are underused, and (ii) the trained models are limited to recognizing the close-set classes only.

Recent large-scale vision-language model (VLM) pre-training [1, 23, 33, 49, 69] eliminates those limitations by learning visual representations via textual supervision. For instance, texts and images are encoded and mapped into a unified space via a contrastive loss during pre-training [49]. The pre-trained text encoder can then be used to synthesize a text-based classifier for image recognition given the corresponding natural language descriptions as shown in Figure 2 (a). Those pre-trained VLMs have demonstrated a powerful transferability on a variety of downstream tasks in a zero-shot manner. However, the effectiveness of the aforementioned models heavily relies on their large-scale architectures and training datasets. For instance, CLIP [49] has up to 428 million parameters and is trained on 0.4 billion text-image pairs, while Flamingo [1] boasts up to 80 billion parameters and is trained on a staggering 2.1 billion pairs. This makes it impractical to fully fine-tune the model on downstream tasks in a low-data regime.

For that reason, efficient transfer learning (ETL) [15, 71–73] on pre-trained VLMs has gained popularity. ETL represents transfer learning to downstream tasks in both

\*Equal contribution.

†Corresponding author.

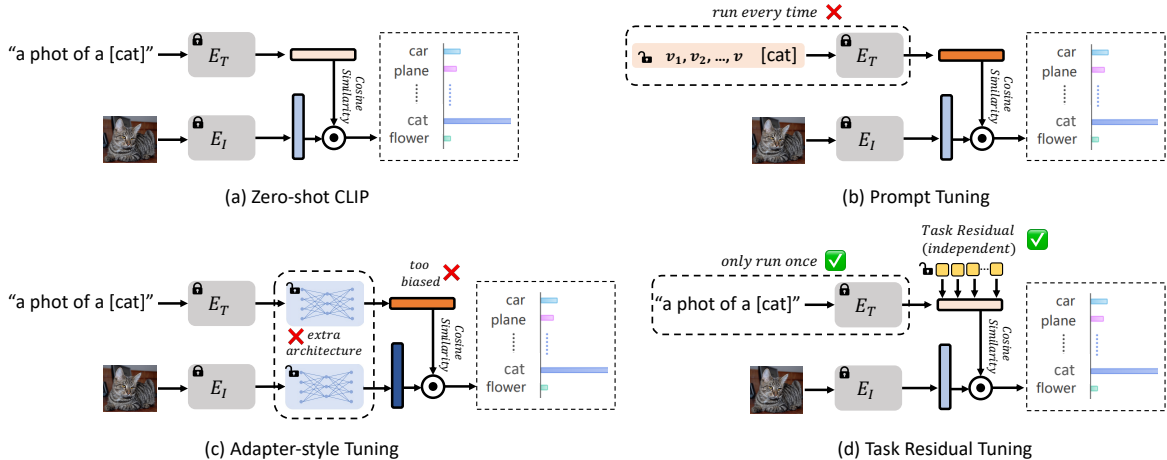


Figure 2. Illustration of (a) Zero-shot CLIP, (b) prompt tuning, (c) adapter-style tuning and (d) our proposed Task Residual Tuning (TaskRes). Our method introduces a *prior-independent* task residual to the fixed pre-trained classifier (*i.e.*, text embeddings of CLIP), being free of running the text encoder every time or extra architecture design.

parameter- and data-efficient manner. The core of ETL is twofold: (i) properly inheriting the well-learned knowledge structure of VLMs, which is already transferable; (ii) effectively exploring the task-specific knowledge given limited data. However, most existing ETL approaches, *i.e.*, prompt tuning (PT) [72, 73] and adapter-style tuning (AT) [15, 71], either damage the prior knowledge of VLMs or learn the new knowledge of a task in an inappropriate/insufficient way. For example, instead of using the pre-trained text-based classifier, CoOp [73] (in Figure 2 (b)) is proposed to learn a continuous prompt for synthesizing a completely new one, which inevitably causes the loss of previous knowledge. Consequently, CoOp underperforms Zero-shot CLIP by 1.03%/0.37% in 1-/2-shot learning on ImageNet (see Figure 1). In contrast, CLIP-Adapter [15] preserves the pre-trained classifier, but is excessively biased towards the prior knowledge when learning a new task, *i.e.*, it transforms the pre-trained classifier weights to be task-specific as illustrated in Figure 2 (c). This results in an inferior new knowledge exploration, thereby a lower accuracy as shown in Figure 1.

For better ETL on pre-trained VLMs, we propose a new efficient tuning approach named Task Residual Tuning (TaskRes), which performs directly on the text-based classifier and explicitly decouples the old knowledge of the pre-trained models and the new knowledge for a target task. The rationale is that the decoupling enables a *better old knowledge inheritance from VLMs* and a *more flexible task-specific knowledge exploration*, *i.e.*, the learned knowledge *w.r.t.* the task is *independent* on the old knowledge. Specifically, TaskRes keeps the original classifier weights frozen and introduces a set of *prior-independent* parameters that are added to the weights. These additive parameters, tuned for adaptation to the target task, are thus named “task residual”.

To gain insight into how TaskRes works, we perform ex-

tensive experiments across 11 benchmark datasets [73] and conduct a systematic investigation of learned task residuals. The experimental results demonstrate that introducing task residual can significantly enhance the transfer performance. We visualize the correlation between the magnitude of learned task residual and the difficulty of transferring a pre-trained model to a downstream task, and observe that the magnitude increases with the transfer difficulty. This suggests the residual is automatically adapted to the task to fully explore the new knowledge, thereby achieving a new state-of-the-art performance on 11 diverse datasets. Furthermore, it is worth noting that our method requires minimal effort for the implementation, *i.e.*, technically adding one line of code only. Our contributions are summarized below:

- We for the first time emphasize the necessity of a proper knowledge inheritance from pre-trained VLMs to downstream tasks via ETL, reveal the pitfalls of existing tuning paradigms, and conduct an in-depth analysis to manifest that decoupling the old pre-trained knowledge and the new task-specific knowledge is the key.
- We propose a new efficient tuning approach named Task Residual Tuning (TaskRes), which achieves a better old knowledge inheritance from VLMs and a more flexible task-specific knowledge exploration.
- TaskRes is convenient for use, which needs a few tuning parameters and effortless implementation.

## 2. Related Work

### 2.1. Vision-Language Models

We mainly review the literature of vision-language models (VLMs) on language-driven visual representation learning (LDVRL) [2, 9, 11, 14, 17, 25, 29, 31, 49, 52, 56]. The core of

LDVRL is to map texts (language) and images into a common space so that the text representations can be used for visual classification. To achieve this, two encoders for texts and images respectively, and specific loss functions as regularization are needed.

Early works explore unsupervised pre-trained models [56] or the skip-gram text modeling [14, 44, 45] for text embedding while sparse coding and vector quantization [7, 56] or Classeme features [11, 60] for visual encoding. For training objectives, MSE [56], self-supervised topic probabilities matching [17] or multi-class logistic loss are adopted. However, those methods are limited to small datasets [9, 25, 31] and weakly representative encoding backbones [7, 56], which tremendously hinders their transferability.

In contrast, recent works [23, 35, 49] have advanced previous ones by leveraging billion-level data of image-text pairs from internet and super powerful neural networks, *e.g.*, Transformers [10, 61, 62, 68, 70], for representation learning. By end-to-end pre-training with a contrastive loss, large-scale VLMs demonstrate remarkable transferability on various downstream tasks in a zero-shot evaluation manner. Standing on the strong transfer capability of large-scale vision-language pre-training, we further explore its potential in efficient transfer learning.

## 2.2. Efficient Transfer Learning

Pre-training a neural network on large-scale datasets, *e.g.*, ImageNet [8] and WebImageText [49], then fine-tuning it on downstream tasks has been a common step of transfer learning [36–38, 65–67]. Here, we mainly focus on efficient transfer learning (ETL) on pre-trained VLMs. ETL represents both parameter- and data-efficient transfer learning, in which a small number of parameters are tuned and a small set of data is utilized. Existing works on ETL can be grouped into two categories: prompt tuning [40, 72, 73] and adapter-style tuning (AT) [15, 71]. Specifically, prompt engineering [24, 54] is first explored to generate appropriate discrete prompts for downstream tasks. Later, learning continuous prompts [39, 73] that can better adapt to a new task shows a more advanced performance. However, these methods face two issues: (i) they need the pre-trained text encoder to engage in the whole training, limiting its scalability and increasing the computational overhead; (ii) they abandon the well pre-trained text-based classifier and generate a new one, which causes the loss of prior knowledge of VLMs.

AT-based methods [15, 71] solve the above issues by using the text encoder once for text-based classifier generation and then focusing on adapting text/image features only. This straightforward design can achieve even better performance, but it heavily relies on the prior knowledge/pre-trained features, which causes an inferior new knowledge exploration. To address this, we propose a new efficient tuning approach, Task Residual Tuning (TaskRes), for a better and more flexi-

ble task-specific knowledge learning by a *prior-independent* “task residual”.

## 2.3. Few-Shot Learning

Few-shot learning (FSL) aims to adapt a model to novel tasks/classes with a handful of labeled examples. Conventional FSL methods [5, 13, 16, 41, 42, 53] often meta-learn on abundant data from base classes for adaptive capability. However, the need of learning on base datasets restricts their scalability. Recent VLM pre-training works [23, 49] offer an effective alternative, which does not require a base dataset. They show that the pre-trained models can already achieve astonishing performance on many downstream tasks in a zero-shot manner. ETL [15, 73] can further boost the performance. In this work, we propose a novel ETL method for adapting VLMs to downstream tasks and evaluate its effectiveness on few-shot tasks.

## 3. Preliminaries

We briefly introduce the adopted VLM, *i.e.*, contrastive language-image pre-training (CLIP) [49], and recap two mainstream approaches for ETL on VLMs, *i.e.*, prompt tuning and adapter-style tuning.

### 3.1. Contrastive Language-Image Pre-training

The CLIP model [49] is designed to obtain visual representations through natural language supervision. It is trained on 0.4 billion image-text pairs, where image features from an image encoder and text features from a text encoder are aligned within a unified embedding space using a contrastive learning loss, allowing CLIP to effectively capture broad visual concepts and learn general visual representations. When testing, CLIP can classify a query image into  $K$  possible categories. This is achieved by calculating the cosine similarity between a query image embedding  $\mathbf{z}$ , obtained from the image encoder with projection, and the text embeddings  $\{\mathbf{t}_i\}_{i=1}^K$ , which are derived by inputting texts (*e.g.*, “a photo of a {class}”) into the text branch. The predicted probability for class  $i$  is formulated as

$$p(y = i | \mathbf{z}) = \frac{\exp(\text{sim}(\mathbf{z}, \mathbf{t}_i) / \tau)}{\sum_{j=1}^K \exp(\text{sim}(\mathbf{z}, \mathbf{t}_j) / \tau)}, \quad (1)$$

where  $\text{sim}(\cdot, \cdot)$  indicates cosine similarity and  $\tau$  is the learned temperature of CLIP.

### 3.2. Revisiting Previous Tuning Paradigms

Inspired by the success of ETL approaches in natural language processing, *e.g.*, prompt tuning [30, 34] and Adapter [22], recent advances (*e.g.*, CoOp [73] and CLIP-Adapter [15]) borrow their ideas to ETL on VLMs.

CoOp, for the first time, introduces prompt tuning to VLMs. Instead of using fixed text prompt contexts such as

“a photo of a”, CoOp proposes to use  $M$  learnable context vectors  $\{\mathbf{v}_m\}_{m=1}^M$  as the task-specific templates. The prompt given to the text encoder then becomes  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M, \mathbf{c}_i\}$  where  $\mathbf{c}_i$  is the embedding of class  $i$ . During the whole training process, CoOp keeps the parameters of pre-trained VLMs frozen and only tunes the learnable vectors  $\{\mathbf{v}_m\}_{m=1}^M$ .

Adapter-style tuning introduces additional modules  $\phi_\omega(\cdot)$  with tunable parameters  $\omega$  to the pre-trained models to transform the pre-trained features  $\mathbf{f}$  to new ones  $\mathbf{f}'$ . In general, adapter-style tuning can be formulated as

$$\mathbf{f}' = \mathbf{f} + \alpha\phi_\omega(\mathbf{f}), \quad (2)$$

where  $\alpha$  is a scaling factor. In CLIP-Adapter, the adapter module  $\phi_\omega$  is composed of two linear transformation layers and a ReLU activation between them. CLIP-Adapter investigates both visual and text adapters, *i.e.*, applying the adapter module to the image and text branches of CLIP, respectively, and shows that they have comparable performance. During training on downstream tasks, adapter-style methods only tune their adapter modules.

## 4. Approach

In this section, we start by identifying the pitfalls of the existing ETL paradigms on VLMs. Those pitfalls motivate us to propose a new ETL approach for VLMs, named *Task Residual Tuning* (TaskRes). The proposed TaskRes is simple yet can effectively avoid the pitfalls.

### 4.1. Pitfalls of Existing ETL Paradigms on VLMs

We rethink the usage of the prior knowledge from pre-trained VLMs and the acquisition of the new knowledge regarding downstream tasks. On the one hand, large-scale VLMs trained with a vast amount of data have learned broad visual concepts that are general for a wide range of downstream vision tasks, enabling homogenization [3]. The prior knowledge should be well preserved when performing a transfer. On the other hand, despite the huge data used in the pre-training, inevitably, there are domain shifts or uncertain concepts in downstream tasks. New knowledge specific to the downstream tasks should be appropriately supplemented to the prior knowledge. However, existing ETL paradigms do not well consider the above principles and have the following two issues.

**Pitfall 1: Lack of guarantees of prior knowledge preservation in prompt tuning.** While the weights of pre-trained text branch modules (*e.g.*, text encoder and projection) are frozen in the prompt tuning paradigm, the original well-learned classification boundary is more or less damaged. This is because the tuning of input prompts ends up with a new boundary that may forget the old knowledge without explicit regularization. As a result, the performance of prompt tuning is limited. For instance, the performance of

CoOp [73] is not as good as Zero-shot CLIP in 1-/2-shot learning on ImageNet as shown in Figure 1.

**Pitfall 2: Limited flexibility of new knowledge exploration in adapter-style tuning.** The data distributions in downstream tasks often shift from the pre-training distribution, and some task-specific or fine-grained visual concepts/representations may not be well learned by pre-trained VLMs, *e.g.*, from CLIP [49] to satellite image dataset EuroSAT [19]. Thus, the new knowledge regarding downstream tasks needs to be appropriately explored. We observe that adapter-style tuning may not sufficiently explore the task-specific knowledge as the input of the adapter is strictly limited to old/pre-trained features, as shown in Figure 2 (c). *Regardless of whether the pre-trained features are suitable for the task, the results of the adapter depend only on them, making adapter-style tuning have limited flexibility for learning new knowledge.*

### 4.2. Task Residual Tuning

Given that the existing ETL paradigms face the aforementioned issues, we propose Task Residual Tuning (TaskRes) to address the issues in a simple way. TaskRes explicitly decouples the maintenance of the old knowledge from the pre-trained VLMs and the learning of task-specific knowledge that is not overly biased on the pre-trained features. We elaborate our TaskRes below.

**Fixed base classifier.** As illustrated in Figure 2 (d), our TaskRes performs the tuning directly on the text-based classifier (*i.e.*, text embeddings)<sup>1</sup>. *Base classifier* is the text embeddings of a pre-trained vision-language model, *e.g.*, CLIP. We denote the base classifier as  $\mathbf{t} \in \mathbb{R}^{K \times D}$  where  $K$  is the number of categories and  $D$  is the feature dimension. We keep the base classifier weights frozen to explicitly prevent the base classifier from being damaged.

**Prior-independent task residual.** To learn task-specific knowledge without being restricted by prior knowledge, we propose *task residual*, which is a set of tunable parameters  $\mathbf{x} \in \mathbb{R}^{K \times D}$  that are not dependent on the base classifier. Our task residual is scaled by a factor  $\alpha$  and element-wisely added to the base classifier to formulate a new classifier  $\mathbf{t}'$  for the target task, written as

$$\mathbf{t}' = \mathbf{t} + \alpha\mathbf{x}. \quad (3)$$

**Tuning for downstream tasks.** During tuning, we only tune the prior-independent task residual while keeping the base classifier (together with the image branch) fixed, enabling reliable old knowledge preservation and flexible new

<sup>1</sup>While we can tune on image embeddings, it may encounter overfitting since the testing image embeddings are different from the training ones. Tuning on text embeddings is free of the issue. Besides, large diversity between image embeddings is not conducive to learning stable task-level information.



knowledge exploration. Given an image, the fixed image branch of CLIP extracts its embeddings  $\mathbf{z} \in \mathbb{R}^D$ . The predicted probability for class  $i$  is then calculated as

$$p(y = i|\mathbf{z}) = \frac{\exp(\text{sim}(\mathbf{z}, \mathbf{t}'_i)/\tau)}{\sum_{j=1}^K \exp(\text{sim}(\mathbf{z}, \mathbf{t}'_j)/\tau)}. \quad (4)$$

Based on the predicted probabilities, the downstream task loss (*e.g.*, cross-entropy loss) only updates the task residual through standard backpropagation.

## 5. Experiment

### 5.1. Setup

We follow prior efficient transfer learning (ETL) work [15, 71, 73] to conduct a few-shot evaluation for ETL models on 11 benchmark datasets *i.e.*, ImageNet [8], Caltech101 [12], OxfordPets [48], StanfordCars [27], Flowers102 [47], Food101 [4], FGVC Aircraft [43], SUN397 [64], DTD [6], EuroSAT [19] and UCF101 [57]. Those datasets include a wide range of visual recognition tasks, *e.g.*, classifying general objects, fine-grained objects, actions, scenes, *etc.* Specifically, ETL models are trained using 1/2/4/8/16 shots per class from the training sets, respectively, and are tested on the full test sets. Additionally, following CoOp [73], we test the generalization performance of our models from ImageNet to its variants (ImageNetV2 [51], ImageNet-Sketch [63], ImageNet-A [21] and ImageNet-R [20]).

### 5.2. Implementation

Our approach has two main components, *i.e.*, base classifier and task residual. For the base classifier, we develop two versions: *regular base classifier* that directly adopts the weights of the text embeddings of the pre-trained CLIP (using pre-defined text templates following prior work [49, 71]), and *enhanced base classifier* obtained by tuning the text projection layer of CLIP on the target task before starting our task residual tuning<sup>2</sup>. Task residual is a matrix filled with learnable parameters that are initialized by zeros. Task residual has the same shape as the base classifier. We scale the task residual and element-wisely add it to the base classifier, as in Eq. 3. By default, the scaling factor  $\alpha$  is set to 0.5 for all datasets except for Flowers102 using 1. Additionally, we explore the use of a learnable  $\alpha$  later in our ablation study. All models in this work are built upon the pre-trained CLIP models. Unless otherwise specified, we use the ResNet-50 [18] version of CLIP which has a ResNet-50 backbone in the image branch. We train our models for 100 epochs for 1-/2-/4-shot experiments and 200 epochs for 8-/16-shot experiments, with batch size 256. Besides, the aforementioned enhanced base classifier is tuned for 50 epochs. Our models

<sup>2</sup>The enhanced base classifier is introduced to help us investigate whether there are consistent performance improvements when the base classifier is enhanced.

are optimized by Adam [26] with an initial learning rate of  $2e-3$  except for ImageNet using  $2e-4$ . Following CoOp [73], our optimization process adopts a cosine learning rate decay schedule and a warmup scheme (*i.e.*, fixing the learning rate in the first epoch to  $1e-5$ ). We test the performance of models after training, and all experimental results are averaged over three random seeds.

## 5.3. Performance Comparison

### 5.3.1 Few-Shot Learning

We develop two versions of our TaskRes, *i.e.*, *Ours/Ours\** using the regular/enhanced base classifier, respectively. We compare our models with Zero-shot CLIP [49], together with the state-of-the-art (SOTA) ETL methods including CoOp [73], CLIP-Adapter [15] and Tip-Adapter-F [71]<sup>3</sup>. The comparison results of the 11 benchmark datasets are shown in Figure 3. Overall, both Ours and Ours\* achieve the SOTA/SOTA-comparable performance on the 11 datasets across all few-shot settings and significantly surpass Zero-shot CLIP [49]. Besides, a few observations can be made. First, with more shots, Tip-Adapter-F reaches a closer performance to ours, but as shown in Figure 1 the number of its tunable parameters increases linearly with that of shots, substantially limiting its scalability. Second, our method shows inferior performance on OxfordPets and Food101. Training with a few shots on these two fine-grained datasets is prone to be overfitting, which is also found in CoOp [73]. Finally, we found the proposed method is superior when needed to recognize a large amount of classes, *e.g.*, 1000 classes of ImageNet and 397 classes of SUN397 in 1-shot case. This is because our TaskRes, not excessively biased to the pre-trained features, can better explore the task-specific knowledge even in such extreme scenarios.

### 5.3.2 Domain Generalization

As pointed out by CoOp [73], ETL models that are trained on a specific domain are at risk of learning spurious correlations when generalizing to unseen domains. We evaluate our models concerning domain generalization. We train our models on 16-shot ImageNet and test the generalization performance of the trained models on four unseen ImageNet variant datasets (ImageNet-V2, -Sketch, -A and -R). As shown in Table 1, our models consistently outperform the compared models (Zero-shot CLIP [49], Linear Probe CLIP [49] and CoOp [73]) across various CLIP visual backbones (ResNet-50 [18], ResNet-101 [18], ViT-B/32 [10] and ViT-B/16 [10]). We also observe that our TaskRes using the enhanced base classifier (Ours\*) achieves higher accuracy than that using the regular base classifier (Ours) on the source dataset while

<sup>3</sup>Original Tip-Adapter [71] has two versions: training-free Tip-Adapter and a fine-tuning version called Tip-Adapter-F. We report the one with higher performance, *i.e.*, Tip-Adapter-F.

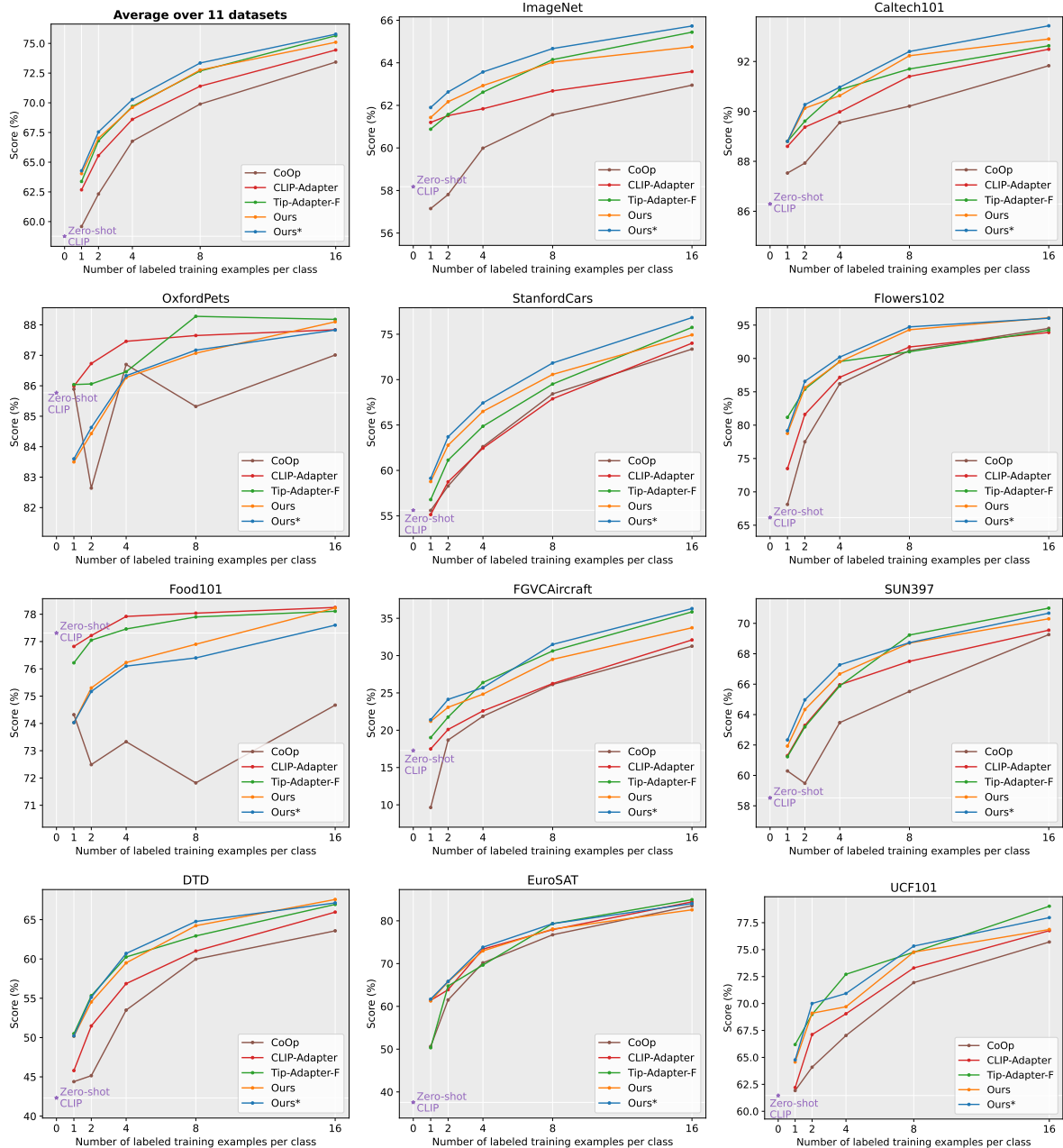


Figure 3. Performance comparison on few-shot learning, *i.e.*, 1-/2-/4-/8-/16-shot, on 11 benchmark datasets. The top-left is the averaged accuracy over the 11 datasets. The full numerical results can be found in the supplementary.

reducing a certain degree of generalization, which is due to a slight overfitting on the source dataset.

#### 5.4. Ablation Study

**Importance of prior knowledge preservation and prior-independent tuning.** We build four target classifiers  $\mathbf{t}'$  to investigate the importance of prior knowledge preservation and prior-independent tuning (where the tunable parameters are not dependent on the pre-trained features), *i.e.*, original

pre-trained classifier  $\mathbf{t}' = \mathbf{t}$  (*i.e.*, our regular base classifier), directly-adapted classifier  $\mathbf{t}' = \phi_{\omega}(\mathbf{t})$ , adapter-style classifier  $\mathbf{t}' = \mathbf{t} + \alpha\phi_{\omega}(\mathbf{t})$  and our TaskRes classifier  $\mathbf{t}' = \mathbf{t} + \alpha\mathbf{x}$ . We conduct experiments on ImageNet across all few-shot settings, using the same scaling factor  $\alpha = 0.5$ . The results in Table 2 support our motivations that (i) explicit preservation of prior knowledge matters as the directly-adapted classifier performs much worse than the original pre-trained classifier and adapter-style classifier, and (ii) tuning prior-

Method	Visual Backbone	Source	Target				
		ImageNet	-V2	-Sketch	-A	-R	Average
Zero-Shot CLIP [49]	ResNet-50	58.18	51.34	33.32	21.65	56.00	40.58
Linear Probe CLIP [49]		55.87	45.97	19.07	12.74	34.86	28.16
CoOp [73]		62.95	55.11	32.74	22.12	54.96	41.23
Ours		64.75	56.47	<b>35.83</b>	<b>22.80</b>	<b>60.70</b>	<b>43.95</b>
Ours*		<b>65.73</b>	<b>57.00</b>	34.43	21.50	58.13	42.77
Zero-Shot CLIP [49]		ResNet-101	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP [49]	59.75		50.05	26.80	19.44	47.19	35.87
CoOp [73]	66.60		58.66	39.08	28.89	63.00	47.41
Ours	67.70		59.50	<b>41.70</b>	<b>29.87</b>	<b>68.07</b>	<b>49.79</b>
Ours*	<b>68.73</b>		<b>60.00</b>	40.3	28.00	64.80	48.28
Zero-Shot CLIP [49]	ViT-B/32		62.05	54.79	40.82	29.57	65.99
Linear Probe CLIP [49]		59.58	49.73	28.06	19.67	47.20	36.17
CoOp [73]		66.85	58.08	40.44	30.62	64.45	48.40
Ours		68.20	59.20	<b>42.50</b>	<b>31.43</b>	<b>69.33</b>	<b>50.62</b>
Ours*		<b>69.17</b>	<b>59.47</b>	40.87	29.70	66.27	49.08
Zero-Shot CLIP [49]		ViT-B/16	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP [49]	65.85		56.26	34.77	35.68	58.43	46.29
CoOp [73]	71.92		64.18	46.71	48.41	74.32	58.41
Ours	73.07		65.30	<b>49.13</b>	<b>50.37</b>	<b>77.70</b>	<b>60.63</b>
Ours*	<b>73.90</b>		<b>65.85</b>	47.70	49.17	75.23	59.49

Table 1. Performance comparison on generalization (from ImageNet to ImageNet-V2/-Sketch/-A/-R) with multiple CLIP visual backbones.

independent parameters is more effective than the dependent one as the TaskRes classifier obviously surpasses the adapter-style classifier.

Target Classifier $t'$	1-shot	2-shot	4-shot	8-shot	16-shot
$t$	60.33	60.33	60.33	60.33	60.33
$\phi_\omega(t)$	28.03	34.90	42.57	48.17	54.13
$t + \alpha\phi_\omega(t)$	61.03	61.23	61.27	62.00	63.17
$t + \alpha x$	<b>61.43</b>	<b>62.17</b>	<b>62.93</b>	<b>64.03</b>	<b>64.75</b>

Table 2. Ablation study of various combinations of three main components: base classifier  $t$ , adapter module  $\phi_\omega(\cdot)$  and our “task residual”  $x$ , for building a target classifier  $t'$ . The experiment is conducted on ImageNet.

**Effectiveness of task residual learning.** We conduct an ablation study of the effect of introducing our task residual learning (TaskRes). Table 3 shows the few-shot results averaged over 11 datasets of the regular base classifier and that equipped with our TaskRes (Ours), together with the enhanced counterparts. Our TaskRes notably improves the regular base, e.g., with 16.14% accuracy gain in 16-shot case. Similarly, the performance gain is also impressive even on the enhanced base. This indicates that TaskRes can always benefit a pre-trained base when adapted to downstream tasks.

Classifier	1-shot	2-shot	4-shot	8-shot	16-shot
Regular Base	58.96	58.96	58.96	58.96	58.96
Regular Base + TaskRes	<b>64.04</b>	<b>67.02</b>	<b>69.61</b>	<b>72.76</b>	<b>75.10</b>
Enhanced Base	61.11	62.82	64.44	69.91	73.09
Enhanced Base + TaskRes	<b>64.28</b>	<b>67.55</b>	<b>70.28</b>	<b>73.35</b>	<b>75.78</b>

Table 3. Exploring the effectiveness of the proposed task residual tuning (TaskRes). The accuracy results averaged on 11 datasets are reported under all shots. Refer to Appendix 2.3 for the full comparison results of the 11 datasets.

**Variants of TaskRes.** The proposed “task residual” can also be applied to the image branch of CLIP, denoted as *TaskRes-I*. We also denote the default TaskRes applied on the text branch as *TaskRes-T* (i.e., Ours) and on the both branches as *TaskRes-I&T*. We compare their performance on ImageNet in Table 4. TaskRes-T consistently outperforms TaskRes-I, as shown in Table 4. The performance of TaskRes-I increases much slower than TaskRes-T as the number of shots increases. This is because TaskRes-I tends to encounter overfitting where training and testing image embeddings are different. Tuning on text embeddings can avoid the issue. Additionally, the diversity of image embeddings is larger than that of text embeddings, making it difficult to learn task-level information. TaskRes-I&T performs slightly worse than TaskRes-T. This might be caused by a misalignment between the updated image embeddings and our base classifier.

Model	1-shot	2-shot	4-shot	8-shot	16-shot
Regular Base	60.33	60.33	60.33	60.33	60.33
TaskRes-I	61.20	61.27	61.47	61.53	61.60
TaskRes-T	61.43	<b>62.17</b>	<b>62.93</b>	<b>64.03</b>	<b>64.75</b>
TaskRes-I&T	<b>61.53</b>	62.10	62.73	63.83	64.67

Table 4. Performance comparison of three TaskRes variants on ImageNet: TaskRes applied on CLIP’s image branch (TaskRes-I), text branch (TaskRes-T) and both branches (TaskRes-I&T).

**Scaling factor.** We investigate the effect of scaling factor  $\alpha$  and show the results of both manually assigned values and a learned value in Table 5. For the manually set  $\alpha$ , we observe a notable improvement with TaskRes scaled even by 0.1, while the highest accuracy is obtained with  $\alpha = 0.5$ . In contrast, the learnable  $\alpha$  can further boost the performance. We show more analyses and results in Appendix 2.3.

$\alpha$	0	0.1	0.3	0.5	0.7	1.0	Learned
Acc.	58.96	61.86	63.76	<b>63.97</b>	63.67	63.02	<b>64.01</b>

Table 5. Ablation study of scaling factor  $\alpha$ . The results are averaged over 11 datasets in 1-shot case.

**Visual backbone.** We further evaluate models on various CLIP visual backbones, namely, ResNet-50 (RN50), ResNet-101 (RN101), ViT-B/32 and ViT-B/16, as shown in Table 6. Our method shows consistent superiority against other alternatives regardless of which backbone is used.

Method	RN50	RN101	ViT-B/32	ViT-B/16
Zero-shot CLIP [49]	58.18	61.62	62.05	66.73
CoOp [73]	62.95	66.60	66.85	71.92
CLIP-Adapter [15]	63.59	65.39	66.19	71.13
Tip-Adapter-F [71]	65.44	68.56	68.65	73.69
Ours	64.75	67.70	68.20	73.07
Ours*	<b>65.73</b>	<b>68.73</b>	<b>69.17</b>	<b>73.90</b>

Table 6. Results of CLIP visual backbones on 16-shot ImageNet.

## 5.5. Investigation of Learned Task Residual

We investigate whether the learned task residuals are in fact related to the difficulty of transferring CLIP to downstream tasks. We first define *relative transfer difficulty* to assess the difficulty of transferring a pre-trained model to a target task. Concretely, the relative transfer difficulty is defined as the ratio of baseline classifier precision to the zero-shot precision of the pre-trained model. We use a random classifier to serve as the baseline classifier where the precision of it equals to  $1/K$  ( $K$  is the number of classes of the downstream task). We then can calculate the relative transfer difficulty of CLIP regarding the 11 benchmark datasets based on the random classifier and Zero-shot CLIP. We count the average magnitudes of the learned task residual of each dataset in 16-shot setting and find that they are highly related to relative transfer difficulty, as shown in Figure 4. This verifies the reliability of the proposed task residual. There are some interesting findings that may be counter-intuitive, *e.g.*, transferring CLIP to ImageNet (with 1000 classes) is the simplest and to EuroSAT (with 10 classes) is the most difficult. This is because the inherent difficulty of ImageNet is large but the data distribution and containing objects are “familiar” to CLIP, whereas EuroSAT is the opposite. Besides, there is more space for TaskRes to play its role in the tasks with large relative transfer difficulty, *e.g.*, EuroSAT (1-shot accuracy gain: 23.71%) and DTD (1-shot accuracy gain: 7.85%). More results and analyses can be found in Appendix 2.4.

## 6. Conclusion, Limitation and Future Work

In this work, we propose a new approach for tuning VLMs, *i.e.*, TaskRes. The proposed TaskRes performs efficient transfer learning (ETL) on VLMs by explicitly de-

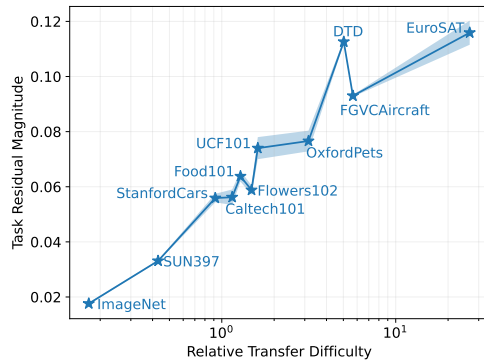


Figure 4. The magnitude of learned task residual increases with relative transfer difficulty (in logarithmic scale). The shadow indicates the standard deviation regarding random seeds.

coupling the classifier into two key parts: undamaged base classifier with rich prior knowledge and task residual independent to the base classifier for better exploring task-specific knowledge. Interestingly, the magnitude of learned task residual is highly related to the difficulty of transferring pre-trained VLMs to a target downstream task. This may inspire the community to consider ETL from a new perspective, *e.g.*, modeling “task-to-task transfer difficulty” for ETL. Extensive experiments have demonstrated the effectiveness of our TaskRes, despite its simplicity.

Nevertheless, our method bears some limitations. For instance, our method encounters negative transfer on two datasets, *i.e.*, OxfordPets (1-shot) and Food101 (1-/2-/4-/8-shot). We conjecture that this happens with two conditions: (i) the downstream tasks are of high relative transfer difficulty as shown in Figure 4, and (ii) Zero-shot CLIP has already achieved a fair precision on them. Besides, the assessment of transfer difficulty in this work is heuristic.

With the rapid advancement of foundation models [3], it is becoming increasingly crucial to establish precise and reliable metrics for assessing the transfer difficulty of pre-trained foundation models to downstream tasks. A comprehensive study of transfer difficulty, which includes distribution analyses [32], is in high demand. Moreover, we can extend the transfer difficulty to a concept-wise level, and investigate the correlation between performance and the occurrence frequency of visual concepts by exploring the CLIP models trained on specific datasets, *e.g.*, SLIP [46] trained on YFCC15M [50, 59]. We will explore these in future work.

## Acknowledgment

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006), the National Research Foundation, Singapore under its Medium Sized Center for Advanced Robotics Technology Innovation, NSFC under Grant U1908209 and 62021001, and ZJNSFC under Grant LQ23F010008.



## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. **1**
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. **2**
- [3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. **4, 8**
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. **5**
- [5] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, pages 4080–4088, 2018. **3**
- [6] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. **5**
- [7] Adam Coates and Andrew Y Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011. **3**
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. **3, 5**
- [9] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, pages 11162–11173, 2021. **2, 3**
- [10] 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, 3, 5**
- [11] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, pages 2584–2591, 2013. **2, 3**
- [12] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pages 178–178. IEEE, 2004. **5**
- [13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135. PMLR, 2017. **3**
- [14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 26, 2013. **2, 3**
- [15] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. **1, 2, 3, 5, 8**
- [16] Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In *ICML*, pages 3607–3616. PMLR, 2020. **3**
- [17] Lluís Gomez, Yash Patel, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, pages 4230–4239, 2017. **2, 3**
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. **1, 5**
- [19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. **4, 5**
- [20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. **5**
- [21] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. **5**
- [22] 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. **3**
- [23] 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 *ICML*, pages 4904–4916. PMLR, 2021. **1, 3**
- [24] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *TACL*, 8:423–438, 2020. **3**
- [25] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, pages 67–84. Springer, 2016. **2, 3**
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2015. **5**
- [27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. **5**
- [28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. **1**
- [29] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, pages 4247–4255, 2015. **2**
- [30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, Online and Punta Cana, Dominican Republic, Nov. 2021. ACL. **3**

- [31] Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *ICCV*, pages 4183–4192, 2017. [2](#), [3](#)
- [32] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*, 2022. [8](#)
- [33] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021. [1](#)
- [34] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, pages 4582–4597, Online, Aug. 2021. *ACL*. [3](#)
- [35] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. [3](#)
- [36] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *NeurIPS*, 2022. [3](#)
- [37] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *NeurIPS*, 2022. [3](#)
- [38] Songhua Liu, Jingwen Ye, Runpeng Yu, and Xinchao Wang. Slimmable dataset condensation. In *CVPR*, 2023. [3](#)
- [39] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 375–383, 2017. [3](#)
- [40] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, pages 5206–5215, 2022. [3](#)
- [41] Zhihe Lu, Sen He, Da Li, Yi-Zhe Song, and Tao Xiang. Prediction calibration for generalized few-shot semantic segmentation. *arXiv preprint arXiv:2210.08290*, 2022. [3](#)
- [42] Zhihe Lu, Sen He, Xi Tian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*, pages 8741–8750, 2021. [3](#)
- [43] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. [5](#)
- [44] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. [3](#)
- [45] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *NeurIPS*, 26, 2013. [3](#)
- [46] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pages 529–544. Springer, 2022. [8](#)
- [47] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICCVGIP*, pages 722–729. IEEE, 2008. [5](#)
- [48] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. [5](#)
- [49] 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 *ICML*, pages 8748–8763. PMLR, 2021. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [50] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. [8](#)
- [51] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. [5](#)
- [52] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, pages 153–170. Springer, 2020. [2](#)
- [53] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987. [3](#)
- [54] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020. [3](#)
- [55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [1](#)
- [56] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 26, 2013. [2](#), [3](#)
- [57] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. [5](#)
- [58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. [1](#)
- [59] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. [8](#)
- [60] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classes. In *ECCV*, pages 776–789. Springer, 2010. [3](#)
- [61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021. [3](#)
- [62] 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. [3](#)
- [63] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. [5](#)
- [64] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society*

- conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. [5](#)
- [65] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *ECCV*, 2022. [3](#)
- [66] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. In *NeurIPS*, 2022. [3](#)
- [67] Jingwen Ye, Songhua Liu, and Xinchao Wang. Partial network cloning. In *CVPR*, 2023. [3](#)
- [68] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, pages 10819–10829, 2022. [3](#)
- [69] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. [1](#)
- [70] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, pages 558–567, 2021. [3](#)
- [71] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022. [1](#), [2](#), [3](#), [5](#), [8](#)
- [72] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. [1](#), [2](#), [3](#)
- [73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)