ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230595

Instance-Wise Adaptive Tuning and Caching for Vision-Language Models

Chunjin Yang, Fanman Meng*, Shuai Chen, Mingyu Liu and Runtong Zhang

University of Electronic Science and Technology of China

Abstract. Large-scale vision-language models (LVLMs) pre-trained on massive image-text pairs have achieved remarkable success in visual representations. However, existing paradigms to transfer LVLMs to downstream tasks encounter two primary challenges. Firstly, the text features remain fixed after being calculated and cannot be adjusted according to image features, which decreases the model's adaptability. Secondly, the model's output solely depends on the similarity between the text and image features, leading to excessive reliance on LVLMs. To address these two challenges, we introduce a novel two-branch model named the Instance-Wise **A**daptive **T**uning and **C**aching (*ATC*). Specifically, one branch implements our proposed $ConditionNet$, which guides image features to form an adaptive textual cache that adjusts based on image features, achieving instance-wise inference and improving the model's adaptability. The other branch introduces the similarities between images and incorporates a learnable visual cache, designed to decouple new and previous knowledge, allowing the model to acquire new knowledge while preserving prior knowledge. The model's output is jointly determined by the two branches, thus overcoming the limitations of existing methods that rely solely on LVLMs. Additionally, our method requires limited computing resources to tune parameters, yet outperforms existing methods on 11 benchmark datasets.

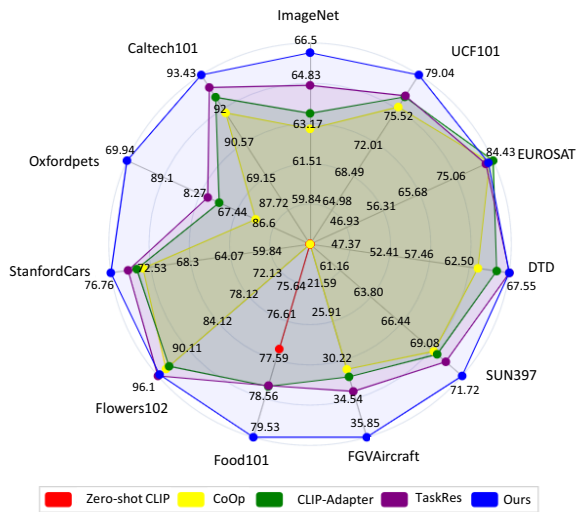


Figure 1: Performance comparison between Zero-shot CLIP [25], CoOp [38], CLIP-Adapter [9], TaskRes [34] and our *ATC* on ImageNet [6] with 16-shot settings.

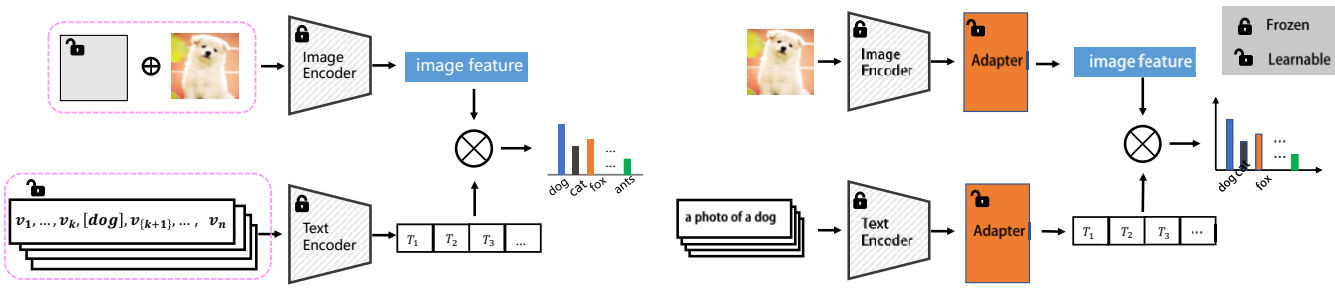
1 INTRODUCTION

Large-scale vision-language models (LVLMs) are trained through contrastive learning on a vast amount of image-text pairs. These models map images and texts to the same space through textual encoders and visual encoders. LVLMs, for instance CLIP [25], ALIGN [5], and ALBEF [18], have shown excellent performance in downstream tasks such as semantic segmentation [33, 39], object detection [21, 35], VQA [28], and so on. However, LVLMs have a large number of parameters, for example, the CLIP [25] model has 428 millions parameters, directly fine-tuning a model could potentially compromise the valuable knowledge obtained during the large-scale pre-training phase, and can pose a risk of over-fitting to the downstream task.

There are currently two main paradigms, as shown in Figure 2, to address above issues: input-level prompt, such as CoOp [38], CoCoOp [37], and feature-level adapter, such as CLIP-Adapter [9], TaskRes [34]. However, for input-level prompt, during the training process, damage would be susceptible to prior knowledge, as demonstrated by CoOp's [38] 1-shot classification accuracy being lower than the Zero-shot CLIP [25]. For feature-level adapter, due to the excessive coupling of prior and new knowledge, the models' ability to learn new knowledge is limited. Additionally, both paradigms suffer from the same problem, that is, the final classification result is only dependent on the similarity between textual features and visual features, leading to excessive reliance on LVLMs, and models' performance upper limit is determined by the LVLMs. Moreover, even within the same category of images, there are differences in visual features. Class-wise text features are not sufficient to cover the large changes in appearance context and geometry of the current category, making it impossible to adapt to the unique features of each test image. Therefore, if only fixed class-wise text features are used as the final classification criteria, the stability of the model will be compromised. Neither paradigm provides a solution for this problem.

In response to the above-mentioned challenges, we propose a novel two-branch model named *ATC*: one branch introduces the similarities between train and val images, and employs training data to create a learnable visual cache, which decouples old and new knowledge and allows our method to retain previous knowledge of LVLMs while maximizing the acquisition of new knowledge. The other branch uses our proposed $ConditionNet$ to direct the visual feature to generate textual biases, which are overlaid on the text feature to produce an adaptive textual cache. This cache automatically

* Corresponding Author. Email: fmmeng@uestc.edu.cn.



(a) Input-level prompt style (b) Feature-level adapter style

Figure 2: **Exsiting paradigms**. (a)Input-level prompt style. Replace the originally discrete text or image with continuous, learnable vectors. (b)Feature-level adapter style. Use an adapter after image or text features to fine-tune the features.

adjusts the text feature based on the image characteristics, achieving instance-wise inference and enhancing the adaptability and generalization of the model. The final output of the model is a combination of the two branches, breaking the existing methods' excessive reliance on LVLMs. Additionally, our *ATC* only requires minimal computing resources to be trained.

We benchmarked our *ATC* on 11 datasets covering various visual recognition tasks such as classification for generic objects, scenes, actions, and fine-grained categories, and specialized tasks such as texture and satellite image recognition. Our results show that *ATC* can effectively transfer pre-trained vision-language models to downstream tasks with limited data, and with better efficiency than existing methods, as demonstrated in Figure 1. In summary, our contributions can be summarized as follows:

- We utilized our proposed $ConditionNet$ to guide image features to fine-tune textual features and developed an adaptive textual cache that can be adjusted based on image characteristics, thus achieving instance-wise inference and enhancing the model's adaptive capabilities.
- We developed a learnable visual cache that decouples new and prior knowledge, enabling our *ATC* to attain maximum acquisition of new knowledge while preserving previously acquired knowledge.
- Our proposed two-branch structure reduces the over-dependence on LVLMs adopted by existing methods. Code is available at https://github.com/Susato9/ATC-main.

2 RELATED WORK

2.1 Large-Scale Vision-Language Models

Large-scale vision-language models (LVLMs) combine visual and textual inputs, enabling them to process, understand, and generate associations between images and natural language text, such as CLIP [25], ALBEF [18], ALIGN [5] and BEiT-v3 [31]. These models are pre-trained on large datasets containing text and images to develop an understanding of both types of inputs. We will use CLIP [25] as an example to elaborate on.

CLIP [25] is one of the most popular LVLMs and is pre-trained on 400 million image-text pairs using contrastive learning techniques [3, 10]. Additionally, it exhibits strong zero-shot classification ability. The CLIP [25] model's training methodology is grounded in two assumptions: (i) text and images can complement each other's information; (ii) different images and texts can be compared for similarity. To achieve this goal, the CLIP [25] uses self-supervised and contrastive learning methods to obtain model parameters by learning the similarity relationship between images and text features. During pretraining, CLIP [25] maps images and text to the same space to update the model parameters by comparing the similarity of positive and negative samples. During the inference stage, given an image and a series of image captions, they are respectively processed by the visual encoder and textual encoder to obtain the image feature vector z and the text feature vector t. Then, their cosine similarity is calculated using the following formula:

$$p(y=i|z) = \frac{exp(sim(z,t_i)/\tau)}{\sum_{j=1}^{K} exp(sim(z,t_j)/\tau)} \quad (1)$$

where $sim(\cdot,\cdot)$ indicates cosine similarity, and τ is the learned temperature of CLIP [25]. Recently, researchers have found that utilizing text supervision can greatly improve the visual representation ability of models. Our work aims to transfer large-scale pre-trained vision-language models to specific tasks via text supervision and a small amount of training data.

2.2 Data-efficient Transfer Learning

Data-efficient transfer learning is a subfield of machine learning that facilitates prior knowledge from pre-trained models to be applied to new target tasks while employing limited training data. Given the financial and temporal costs of acquiring and annotating sizable data, data-efficient transfer learning has emerged as a prevalent research domain. In this regard, pretrained models obtained from large-scale datasets, such as ImageNet [6], by utilizing advanced hardware and neural network structures might be used to expedite the target task's learning process. However, because of the discrepancy in data and distributions between source and target tasks, achieving optimal performance becomes a challenge, and several strategies are employed in data-efficient transfer learning, such as meta-learning and data augmentation.

Recently, new paradigms have emerged with the increasing number of large-scale vision-language models. CoOp [38] and CoCoOp [37] proposed to replace the fixed prompt templates in large-scale vision-language models with continuous and learnable vectors, and fine-tune them based on downstream task data. CLIP-Adapter [9] uses simple linear projection to adapt features to downstream tasks. TaskRes [34] improves model performance by decoupling old and new knowledge of large-scale vision-language models through a simple approach. In this work, we propose a new two-branch model for data-efficient transfer learning that achieves better performance than the aforementioned methods.

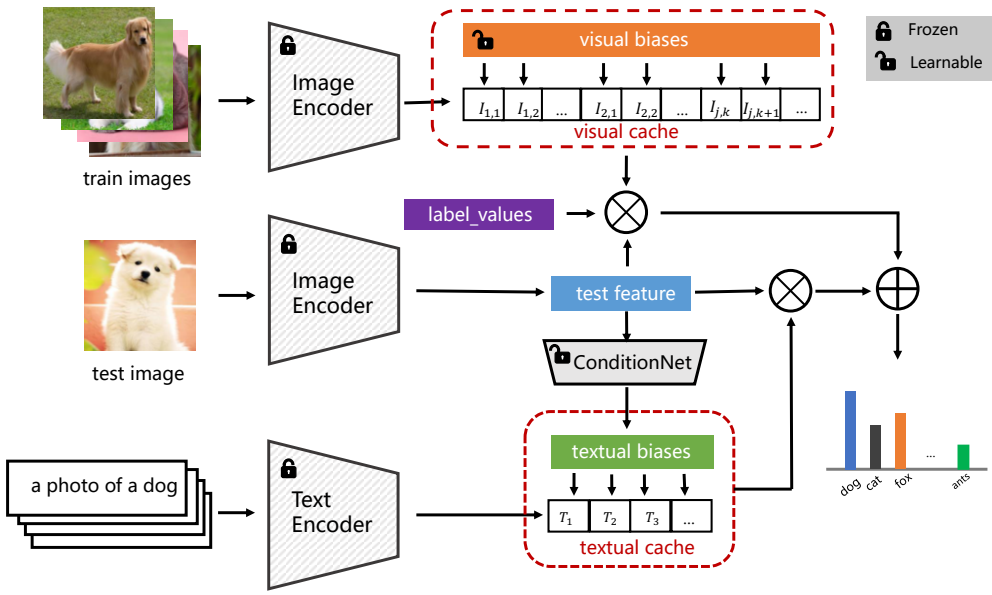


Figure 3: **The Pipeline of *ATC***. Our proposed model adopts a two-branch structure. The upper branch constructs a learnable visual cache by applying an adjustable training matrix to all support image features. The lower branch features our proposed $ConditionNet$ that generates biases for the test image features, which adjusts the textual cache. Combining the two branches results in the production of the final output, thereby addressing the limitations of existing methods that rely only on LVLMs.

3 METHOD

The overall pipeline of our proposed method is shown in Figure 3. In this section, we will first review the challenges of current data-efficient transfer learning based on LVLMs, and then specifically introduce our proposed method, *ATC*.

3.1 *Defects of Existing Data-efficient Transfer Learning based on LVLMs*

LVLMs possess massive parameters, and fine-tuning the model with insufficient data presents a risk of over-fitting and causes damage to model's original knowledge. However, due to domain shift between the training data of LVLMs and downstream tasks, transfer learning is necessary. Thus, effective data-efficient transfer learning should enable the model to acquire new knowledge substantively while upholding its previous knowledge. However, there are still some issues that need to be addressed when applying LVLM-based data-efficient transfer learning currently.

Methods that employ prompt engineering at input level, such as CoOp [38] and CoCoOp [37], allow models to alter static text prompt templates into continuous learnable prompt templates for adaptation to new tasks. Nevertheless, using these methods could lead to loss of some of CLIP's original knowledge. For example, CoOp's [38] 1-shot and 2-shot accuracy is inferior to CLIP's [25] zero-shot accuracy, while the authors of CoCoOp [37] reported that the computing resources and time consumed by the model during training are considerable. Even though the visual and textual encoders are set with fixed parameters, and the model only modifies a few parameters, saving the complete model's gradient is resource-intensive during training for backpropagation. Similarly, Hyojin [1] and MaPLe [16] introduced prompt engineering [15, 19, 20, 27] from natural language processing into image processing, which sacrifices prior knowledge and consumes resources for LVLMs to adapt to downstream tasks.

Approaches that fine-tune features on the feature side, such as CLIP-Adapter [9] only utilize adapters after the visual or textual encoder, which maps new knowledge onto old knowledge in an excessively coupled manner. Due to this excessive coupling, the learning capacity for new knowledge is limited. Although TaskRes [34] decouples new and old knowledge by implementing learnable masks on text features, the mask remains static after the completion of training, leading to limited adaptability. Moreover, both paradigms inevitably suffer from a problem where the final classification result only relies on the similarity between visual and textual features, resulting in the inadequate utilization of training data and excessive dependence on LVLMs. Moreover, both paradigms have poor adaptability because text features remain unchanged once calculated and cannot be adjusted based on image characteristics.

3.2 *ATC*

In response to the issues with existing methods, we propose a novel two-branch model. On one branch, we use training data to construct a learnable visual cache. On the other branch, we employ our proposed $ConditionNet$ to adjust the adaptive textual cache. The final result is jointly determined by these two branches. In the following sections, we will provide a detailed introduction to our *ATC*.

3.2.1 $ConditionNet$

Even the visual features of images in the same category vary, class-wise text features are insufficient to cover the significant changes in the appearance and context geometry of the current category, making it impossible to adapt to the unique features of each test image. Therefore, if only fixed and invariant class-wise text features are used as the final classification criteria, the model's generalization and stability will be compromised. Therefore, to enable text features to automatically adjust based on the features of test images, we propose $ConditionNet$. This network can generate textual biases to adjust the textual cache by perceiving the features of test images. The proposed $ConditionNet$ facilitates communication between image features and text features, achieving adaptive adjustment of text cache

during the testing process. By utilizing instance-wise inference, it improves the model's generalization and adaptability. Here, we use a small number of parameters LSTM as the $ConditionNet$.

3.2.2 Adaptive Textual cache

We propose an adaptive textual cache that can fine-tune the text features based on the characteristics of test images which disrupts the paradigm of traditional methods, where the text features remain constant once computed. Here is the process of constructing the adaptive textual cache. Firstly, the category labels of the entire dataset are placed in a fixed manual prompt template (e.g., "a photo of a {class}"), denoted by V, $V = [v_1, v_2, \ldots, v_c]$, then, V is fed into the textual encoder of the CLIP [25], $Encoder\_txt$, to obtain the initial textual cache $P_{txt} = [t_1, t_2, \ldots, t_c] \subset R^{c \times dim}$. Here, c represents the number of categories in the dataset, and dim represents the output dimension of the $Encoder\_txt$. Calculating the cache only once significantly reduces computational cost. Meanwhile, test image I is processed by the visual encoder, $Encoder\_img$, resulting in image features, $f_{test} \subset R^{1 \times dim}$, the feature f_{test} is fed to our proposed $ConditionNet$, a model with few parameters, to generate textual biases, $textual\_biases \subset R^{c \times dim}$,

$$s = ConditionNet(f_{test}) \tag{2}$$

$$textual\_biases = [s, s, \ldots, s] \tag{3}$$

Overlaying the $textual\_biases$ onto the textual cache generates the latest cache, $\hat{P}_{txt} \subset R^{c \times dim}$,

$$\hat{P}_{txt} = P_{txt} + textual\_biases \tag{4}$$

3.2.3 Learnable Visual cache

To eliminate the excessive reliance on LVLMs, we proposed a novel branch, and constructed a learnable visual cache on this branch. Through this cache, our ATC has achieved the decoupling of new and existing knowledge, enabling the model to fully acquire new knowledge while retaining prior knowledge. Without loss of generality, let us consider an experiment with k shots and n classes, with $n \times k$ training images that have undergone data augmentation, represented as S_{aug}. We obtain the initial visual cache by encoding the images once, $P_{img} \subset R^{n \cdot k \times dim}$,

$$P_{img} = Encoder\_img(S_{aug}) = [I_{1,1}, I_{1,2}, \ldots I_{i,j}, \ldots I_{c,k}] \tag{5}$$

and generate a one-hot matrix of labels for all training data in accordance with the order of the images, denoted as $label\_values \subset R^{n \cdot k \times c}$. $I_{i,j}$ represents the visual features of the $j-th$ image corresponding to the $i-th$ category. To enable the image cache to be automatically adjusted for the task, we initialize a matrix of zeros, denoted as $visual\_biases \subset R^{n \cdot k \times dim}$, which is automatically updated during model training and aggregated with the original visual cache.

$$\hat{P}_{img} = P_{img} + visual\_biases \tag{6}$$

We incorporate the design concept of the Tip-Adapter [36] caching model when creating the visual cache. Our method differs from Tip-Adapter [36] in that Tip-Adapter [36] initializes a linear layer with training data, while we overlay a learnable mask initialized at zero on image features from the training data. Our method offers a larger learning space for the model.

Given an image feature $f_{test} \subset R^{1 \times dim}$, the final predicted probability distribution $f \subset R^{1 \times c}$ is obtained through the above three modules using the following formula:

$$f_1 = f_{test} \otimes \hat{P}_{img}^{\mathsf{T}} \otimes label\_values \tag{7}$$

$$f_2 = f_{test} \otimes \hat{P}_{txt}^{\mathsf{T}} \tag{8}$$

$$f = \alpha f_1 + \beta f_2 \tag{9}$$

Here, $\otimes$ represents the hadamard product of matrices, f_1 represents the cosine similarity between the f_{test} and the visual cache, while f_2 represents the cosine similarity between the f_{test} and the textual feature cache. The variables α and β are weighting coefficients.

3.3 EXPERIMENT

3.4 Experiment Setup

We follow previous work [9, 34, 38] to conduct a few-shot evaluation on 11 benchmark datasets, including ImageNet [6], Caltech101 [8], OxfordPets [24], StanfordCars [17], Flowers102 [23], Food101 [2], FGVCAircraft [22], SUN397 [32], DTD [4], EuroSAT [12], and UCF101 [29]. These datasets cover various computer vision tasks, specifically, ImageNet and Caltech101 are used for classification of generic objects, while OxfordPets, StanfordCars, Flowers102, Food101, and FGVCAircraft are used for fine-grained classification, SUN397 is used for scene recognition, UCF101 is used for action recognition, DTD is used for texture classification, and finally, EuroSAT is used for satellite imagery recognition. We train the model by randomly sampling 1/2/4/8/16 samples from each class in the training data and test the model on the entire test dataset. During the training process, the following cross entropy loss is utilized:

$$L = -[ylog(\hat{y}) + (1-y)log(1-\hat{y})] \tag{10}$$

Additionally, we performed domain generalization experiments following the experimental settings of CoCoOp [37], Tip-Adapter [36], and TaskRes [34]. For the domain generalization experiments, we used ImageNet [6] as the source dataset and four datasets ImageNet-V2 [26], ImageNet-Sketch [30], ImageNet-A [14], and ImageNet-R [13] , which have certain domain differences from ImageNet, as our target datasets.

3.5 Baseline models

For the few-shot classification experiment, we compare our approach with Zero-shot CLIP [25], Linear-Probe CLIP [25], CoOp [38], Tip-Adapter-F [36], TaskRes [34], and TaskRes* [34]. Both Zero-shot CLIP and Linear-Probe CLIP [25] utilized the same handcrafted prompt template, such as "a photo of a {class}", while CoOp [38] replaced the fixed handcrafted prompt template with a continuous learnable template. Tip-Adapter-F [36] represents a variant of Tip-Adapter [36] that utilizes a small amount of data to train a caching model. TaskRes [34] improves the classification performance of models by augmenting the text features with a learnable matrix initialized at zero, while TaskRes* [34] further boosts the performance by enhancing the classifier based on TaskRes [34].

For the domain generalization experiment, we compare our approach with Zero-shot CLIP [25], Linear-Probe CLIP [25], CLIP+CoOp (M=16) [38], CLIP+CoOp (M=4) [38], TaskRes* [34].

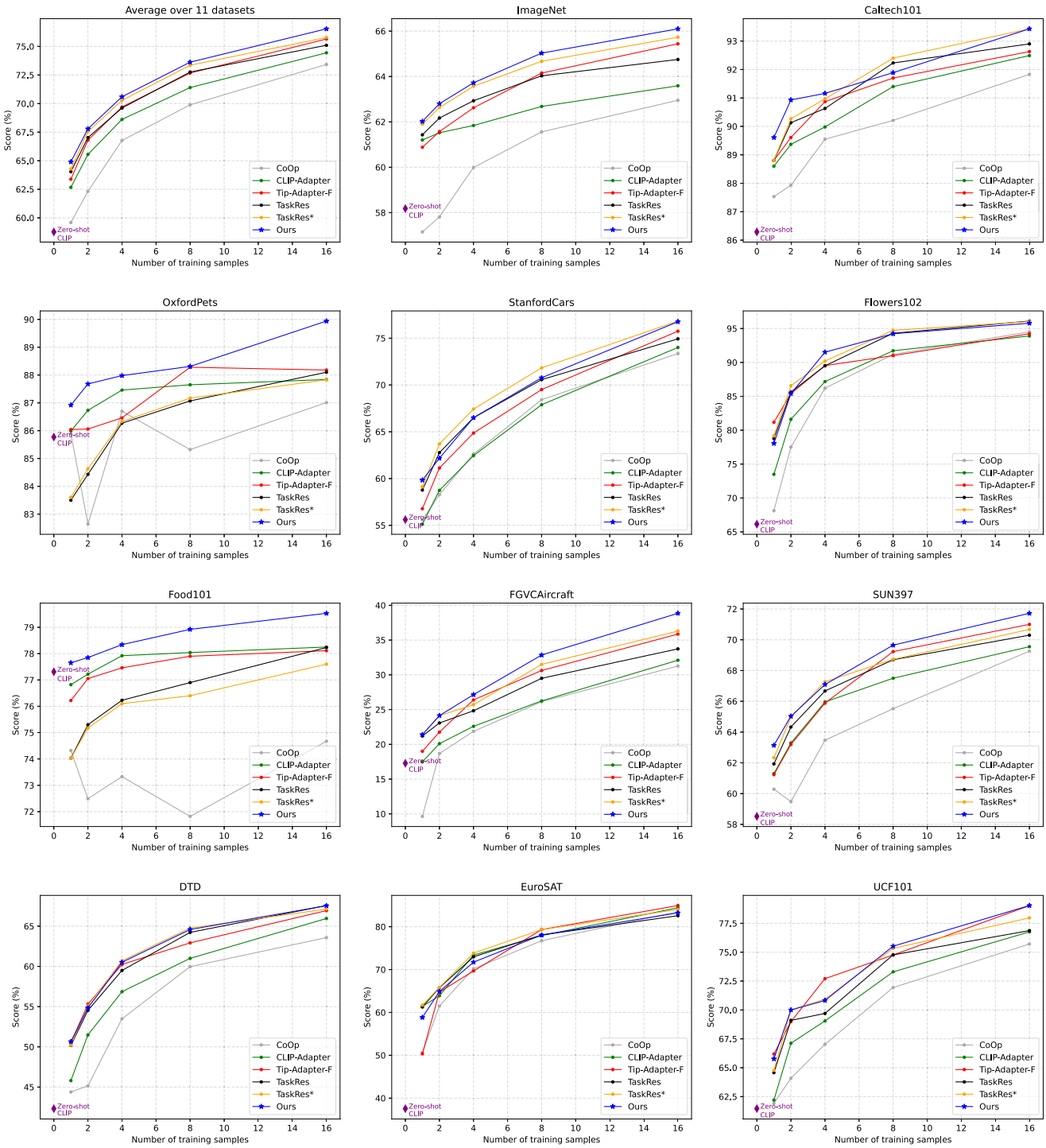


Figure 4: **Main results of few-shot learning on 11 datasets**. Our approach *ATC* consistently shows better performance over previous baselines across different training shots, the top-left is the averaged accuracy over the 11 datasets.

3.6 Performance Comparison

few-shot learning The primary experimental results, as depicted in Figure 4, showcase a comparison between our proposed 1/2/4/8/16-shot experimental outcomes and the state-of-the-art few-shot transfer learning techniques based on CLIP [25], including Linear-Probe CLIP [25], CoOp [38], Tip-Adapter-F [36], TaskRes [34], and TaskRes* [34]. Our method achieves the highest average classification accuracy on 11 datasets, as indicated in the upper left corner of the Figure 4 , and notably outperforms other algorithms on prominent datasets such as ImageNet, OxfordPets, Food101, FGVCAircraft, SUN397, especially on Food101, where other methods fail to demonstrate satisfactory results. Remarkably, our approach achieves the highest accuracy on eight datasets out of eleven, excluding StanfordCars, Flowers102, and EuroSAT, upon conducting 16-

Table 1: **Performance comparison on generalization (from ImageNet to ImageNet-V2/-Sketch/-A/-R) with multiple CLIP visual backbones**. Our proposed method achieved the highest average accuracy across four different visual backbones.

Method	Vision backbone	Source	Target				
		ImageNet	-V2	-sketch	-A	-R	Average
Zero-shot CLIP [25]	ResNet-50 [11]	58.18	51.34	33.32	21.65	56.00	40.58
Linear-Probe CLIP [25]		55.87	45.97	19.07	12.74	34.86	28.16
CLIP+CoOp(M=16) [38]		62.95	55.11	32.74	22.12	54.96	41.23
CLIP+CoOp(M=4) [38]		63.33	55.40	34.67	**23.06**	56.60	42.43
TaskRes* [34]		65.73	**57.00**	34.43	21.50	58.13	42.77
Ours		**66.10**	56.78	**35.39**	22.19	**59.25**	**43.40**
Zero-shot CLIP [25]	ResNet-101 [11]	61.62	54.81	38.71	28.05	64.38	46.49
Linear-Probe CLIP [25]		59.75	50.05	26.80	19.44	47.19	35.87
CLIP+CoOp(M=16) [38]		66.60	58.66	39.08	28.89	63.00	47.41
CLIP+CoOp(M=4) [38]		65.98	56.80	40.40	29.60	64.98	47.95
TaskRes* [34]		68.73	60.00	40.30	28.00	64.80	48.28
Ours		**69.28**	**61.20**	**41.55**	**30.51**	**67.81**	**50.27**
Zero-shot CLIP [25]	ViT- B/32 [7]	62.25	54.79	40.82	29.57	65.99	47.79
Linear-Probe CLIP [25]		59.58	49.73	28.06	19.67	47.20	36.17
CLIP+CoOp(M=16) [38]		66.85	58.08	40.44	30.62	64.45	48.40
CLIP+CoOp(M=4) [38]		66.34	58.24	41.48	31.34	65.78	49.21
TaskRes* [34]		69.17	59.47	40.87	29.70	66.27	49.08
Ours		**69.45**	**61.09**	**42.10**	**32.13**	**68.82**	**51.04**
Zero-shot CLIP [25]	ViT- B/16 [7]	66.73	60.83	46.15	47.77	73.96	57.18
Linear-Probe CLIP [25]		65.85	56.26	34.77	35.68	58.43	46.29
CLIP+CoOp(M=16) [38]		71.92	64.18	46.71	48.41	74.32	58.41
CLIP+CoOp(M=4) [38]		71.73	64.56	47.89	49.93	75.14	59.38
TaskRes* [34]		73.90	65.85	47.70	49.17	75.23	59.49
Ours		**74.34**	**66.02**	**48.89**	**50.38**	**77.38**	**60.67**

shot experiments. Furthermore, our proposed method outperforms Zero-shot CLIP [25] on all 11 datasets, as depicted in Figure 5, as compared to the Zero-shot CLIP [25], our approach has significant improvement in accuracy.

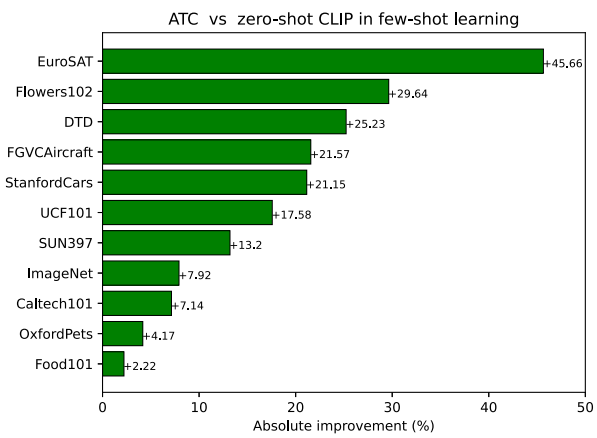


Figure 5: **Comparison with Zero-shot CLIP [25] in few-shot setting**. On 11 benchmark datasets, our method demonstrates a significant improvement in accuracy compared to Zero-shot CLIP.

3.6.1 Domain Generalization

The primary objective of this experiment is to assess the model's generalization ability. We randomly select 16 samples per class from ImageNet as training data and independently test our model on the entire test sets of variants ImageNet datasets, including ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. We compare our experimental results with those of other state-of-the-art few-shot transfer learning techniques, including Zero-shot CLIP [25], Linear-Probe CLIP [25], CLIP+CoOp(M=16) [38], CLIP+CoOp(M=4) [38], and TaskRes* [34]. Our method achieves the highest average accuracy on all four datasets by using different visual backbone networks except for ResNet-50 [11], as shown in the summarized results in Table 1. When used as the visual encoder, ResNet-50's [11] encoding dimension is 1024, while other visual backbone networks have an encoding dimension of only 512, which results in unsatisfactory performance. Therefore, using ResNet-50 [11] as the visual backbone network requires the model to learn more parameters when constructing the adaptive textual cache and $ConditionNet$, leading to some degree of over-fitting on the source dataset. Furthermore, as illustrated in Figure 6, our approach demonstrated excellent generalizability as evidenced by its higher accuracy on all four target datasets compared to Zero-shot CLIP [25].

3.6.2 Ablation Study

In this section, we mainly conducted ablation experiments on the two main modules we proposed adaptive textual cache and learnable visual cache and four different visual backbone networks, namely ResNet-50 [11], ResNet-101 [11], ViT-B/32 [7], and ViT-B/16 [7], using different weighting coefficients α and β. The experiments were conducted primarily on the ImageNet dataset.

Adaptive Textual cache We conducted ablation experiments on adaptive textual cache by comparing it with fixed textual cache. Table 2 summarizes the experimental results, indicating that using adaptive textual cache improves accuracy across 1/2/4/8/16-shot experiments, effectively validating the efficacy of this module.

Learnable Visual cache The effectiveness of the learnable visual cache is validated through experiments. We conducted ablation ex-

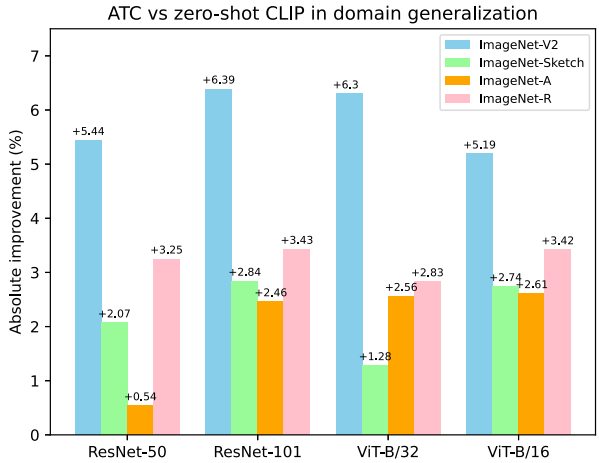


Figure 6: **Comparisions with Zero-shot CLIP [25] in domain generalization**. Our *ATC* has achieved improvement in accuracy compared to zero-shot CLIP on four benchmark datasets by using different visual backbone networks.

Table 2: **Ablation Study of adaptive textual cache**.In different few-shot experimental settings, utilizing an adaptive textual cache is beneficial in enhancing the model's performance.

Textual cache	1-shot	2-shot	4-shot	8-shot	16-shot
Fixed	61.79	62.45	63.23	64.42	65.72
Adaptive	**62.02**	**62.80**	**63.72**	**65.03**	**66.10**
	+0.23	+0.35	+0.49	+0.61	+0.38

periments on the ImageNet dataset to compare our method of constructing visual caches by adding learnable biases with the methods of constructing fixed caches and using learnable linear layers in Tip-Adapter-F [36], the experimental results are presented in the Table 3. The experimental results are as follows: from the experimental results, it can be seen that our proposed method of constructing caches outperforms the linear layer initialization in Tip-Adapter-F, demonstrating the effectiveness of our proposed method.

Table 3: **Ablation study of learnable visual cache**. Under various few-shot experimental settings, constructing a learnable visual cache by adding biases can greatly enhance the model's performance.

Visual cache	1-shot	2-shot	4-shot	8-shot	16-shot
Fixed	61.87	62.39	62.78	63.75	64.68
Learnable linear layer	61.96	62.31	62.78	64.49	65.84
Adding biases	**62.02**	**62.80**	**63.72**	**65.03**	**66.10**
	+0.06	+0.49	+0.94	+0.54	+0.26

Vision Backbone Furthermore, we conducted few-shot experiments on ImageNet dataset for different visual backbone networks, including ResNet-50 [11], ResNet-101 [11], ViT-B/32 [7] and ViT-B/16 [7]. The experimental results are presented in the Table 4, which demonstrates that our method outperforms other methods across all visual backbone networks.

α and β Based on the Figure 3 and Formula 9, the final classification probability is obtained by combining two branches, with α and β as the corresponding weighting coefficients. We conducted separate experiments on these coefficients. When we varied α, we fixed β at **1**, and when we varied β, we fixed α at **1**. According to the

Table 4: Results of CLIP visual backbones on 16-shot ImageNet.

Method	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16
Zero-shot CLIP [25]	58.18	61.62	62.05	66.73
CoOp [38]	62.95	66.60	66.85	71.92
CLIP-Adapter [9]	63.59	65.39	66.19	71.13
Tip-Adapter-F [36]	65.44	68.56	68.65	73.69
TaskRes [34]	64.75	67.70	68.20	73.07
TaskRes* [34]	65.73	68.73	69.17	73.90
Ours	**66.10**	**69.10**	**69.65**	**74.34**
	+0.37	+0.37	+0.48	+0.44

experimental results, the highest accuracy was achieved when both coefficients were set to **1**. To validate our findings, we conducted 16-shot experiments using ResNet-50 [11] on the ImageNet dataset, the result of experiment are present in Table 5.

Table 5: **Ablation study of α and β**. The best performance is achieved when α and β are both equal to 1.

α	0.00	0.50	1.00	1.50	2.00
accuracy	58.18	65.03	**66.10**	65.89	65.98
β	1.00	1.50	2.00	2.50	3.00
accuracy	**66.10**	66.01	65.92	65.75	64.26

4 Conclusion, Limitations and Future Work

Our proposed method aims to address the primary issue of data-efficient transfer learning found in LVLMs. We propose a new two-branch model called ATC. In the first branch, We constructed a learnable visual cache from the training data, which enables the decoupling of new and old knowledge. This allows our model to acquire new knowledge while retaining the prior knowledge of LVLMs. In the second branch, the proposed $ConditionNet$ guides visual features to generate textual biases that are overlaid on the text feature, creating an adaptive textual cache. This cache automatically adjusts the text feature based on the image feature, providing strong adaptability. The final output of the model combines the two branches, reducing the existing methods' over-reliance on LVLMs, and our approach generates higher accuracy on 11 datasets compared to previous methods.

However, our proposed method has one primary limitation, an increase in training data will result in an visual cache expansion, leading to a higher number of parameters and resource consumption.

The integration of vision-language, and multimodal pre-training is a growing field and requires further research and exploration to efficiently transfer various large-scale models to downstream tasks. We anticipate that the empirical findings and insights we present here can lay the groundwork for future research on efficient adaptation methods for emerging fundamental models, which still require more investigation.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0112000, the National Natural Science Foundation of China under Grant 62271119, the Natural Science Foundation of Sichuan Province under Grant 2023NSFSC1972.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola, 'Exploring visual prompts for adapting large-scale models', *arXiv preprint arXiv:2203.17274*, **1**(3), 4, (2022).

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, 'Food-101–mining discriminative components with random forests', in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, (2014).

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, 'A simple framework for contrastive learning of visual representations', in *International conference on machine learning*, pp. 1597–1607. PMLR, (2020).

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi, 'Describing textures in the wild', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, (2014).

[5] Gerson H Cohen, 'Align: a program to superimpose protein coordinates, accounting for insertions and deletions', *Journal of applied crystallography*, **30**(6), 1160–1161, (1997).

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, (2009).

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

[8] 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 *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, (2004).

[9] 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).

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, 'Momentum contrast for unsupervised visual representation learning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, (2020).

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

[12] 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).

[13] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, (2021).

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song, 'Natural adversarial examples', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, (2021).

[15] 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).

[16] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan, 'Maple: Multi-modal prompt learning', *arXiv preprint arXiv:2210.03117*, (2022).

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, '3d object representations for fine-grained categorization', in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, (2013).

[18] 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', *Advances in neural information processing systems*, **34**, 9694–9705, (2021).

[19] Xiang Lisa Li and Percy Liang, 'Prefix-tuning: Optimizing continuous prompts for generation', *arXiv preprint arXiv:2101.00190*, (2021).

[20] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang, 'Gpt understands, too', *arXiv preprint arXiv:2103.10385*, (2021).

[21] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu, 'Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14074–14083, (2022).

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi, 'Fine-grained visual classification of aircraft', *arXiv preprint arXiv:1306.5151*, (2013).

[23] Maria-Elena Nilsback and Andrew Zisserman, 'Automated flower classification over a large number of classes', in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, (2008).

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar, 'Cats and dogs', in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, (2012).

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

[26] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, 'Do imagenet classifiers generalize to imagenet?', in *International conference on machine learning*, pp. 5389–5400. PMLR, (2019).

[27] 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).

[28] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei, 'Clip models are few-shot learners: Empirical studies on vqa and visual entailment', *arXiv preprint arXiv:2203.07190*, (2022).

[29] 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).

[30] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing, 'Learning robust global representations by penalizing local predictive power', *Advances in Neural Information Processing Systems*, **32**, (2019).

[31] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al., 'Image as a foreign language: Beit pretraining for all vision and vision-language tasks', *arXiv preprint arXiv:2208.10442*, (2022).

[32] 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*, pp. 3485–3492. IEEE, (2010).

[33] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai, 'Side adapter network for open-vocabulary semantic segmentation', *arXiv preprint arXiv:2302.12242*, (2023).

[34] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang, 'Task residual for tuning vision-language models', *arXiv preprint arXiv:2211.10277*, (2022).

[35] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang, 'Open-vocabulary object detection using captions', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, (2021).

[36] Renrui Zhang, Wei Zhang, 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pp. 493–510. Springer, (2022).

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, 'Conditional prompt learning for vision-language models', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, (2022).

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

[39] Ziqin Zhou, Bowen Zhang, Yinjie Lei, Lingqiao Liu, and Yifan Liu, 'Zegclip: Towards adapting clip for zero-shot semantic segmentation', *arXiv preprint arXiv:2212.03588*, (2022).