# Fine-Grained Visual Prompt Learning of Vision-Language Models for Image Recognition

Hongbo Sun
Wangxuan Institute of Computer Technology
& National Key Laboratory for Multimedia Information
Processing, Peking University
Beijing, China
sunhongbo@pku.edu.cn

Xiangteng He
Wangxuan Institute of Computer Technology
& National Key Laboratory for Multimedia Information
Processing, Peking University
Beijing, China
hexiangteng@pku.edu.cn

Jiahuan Zhou
Wangxuan Institute of Computer Technology
& National Key Laboratory for Multimedia Information
Processing, Peking University
Beijing, China
jiahuanzhou@pku.edu.cn

Yuxin Peng[*]
Wangxuan Institute of Computer Technology
& National Key Laboratory for Multimedia Information
Processing, Peking University
Beijing, China
pengyuxin@pku.edu.cn

## ABSTRACT

Large-scale pre-trained vision-language (VL) models have shown powerful generic representation capabilities for adapting to downstream tasks with limited training data, which are data-efficient solutions to various applications such as image recognition. In order to enhance the adaption performance, most existing methods attempt to introduce learnable vectors into the text prompt to generate adaptive classification weights for the class in the downstream task. However, they generally focus on the text side while neglecting adaptive visual feature generation on the image side, which is insufficient to fit the downstream task data. In this paper, we propose **fine-grained visual prompt learning (FG-VPL)** of vision-language models for image recognition with few training samples, and the main contributions are: (1) Fine-grained visual prompt is introduced into the image encoder of the vision-language model for focusing on the target object and conducting information interaction within the object, which facilitates generating discriminative visual features for image recognition. (2) A two-pathway adaptive recognition module is proposed to narrow the domain gap and utilize both the cross-modal knowledge of the vision-language model and the visual information of the few-sample training set for classifying images with the help of feature adapters. We conduct extensive experiments on **11** image recognition benchmark datasets under the few training samples setting, which demonstrate that our proposed approach can achieve state-of-the-art performance. The code is available at https://github.com/PKU-ICST-MIPL/FG-VPL_ACMMM2023.

[*] Corresponding author.

*MM '23, October 29-November 3, 2023, Ottawa, ON, Canada*
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612403

## CCS CONCEPTS

• **Computing methodologies → Object recognition**.

## KEYWORDS

Fine-Grained visual prompt learning; Vision-language models; Image recognition with few training samples

**ACM Reference Format:**
Hongbo Sun, Xiangteng He, Jiahuan Zhou, and Yuxin Peng. 2023. Fine-Grained Visual Prompt Learning of Vision-Language Models for Image Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612403

## 1 INTRODUCTION

Exiting state-of-the-art image recognition models, such as the ResNet series [10] and ViT series [7], are generally trained on separate datasets to predict discrete image class labels of a fixed set. In this paradigm, they are limited to closed-set visual concepts, which significantly affects their generalization ability. The recently emerged large-scale pre-trained ***vision-language (VL)*** models bring a new paradigm for recognizing open-set visual concepts with the powerful generic representation capability. Specifically, VL models, such as the well-known CLIP [25], project the images and corresponding raw texts into the common feature space with separate encoders for alignment, which pulls the matched image-text pairs together while pushing the unmatched image-text pairs away with contrastive learning. After pre-training on the large-scale image-text pairs, various visual concepts from the natural language texts are captured by the model and the learned representation can be readily transferred to wide-ranging downstream tasks. For example, the downstream image recognition can be conducted with few training samples through the similarity calculation between the projected image feature and text feature extracted from the natural language description of new classes. However, there exists a gap between the pre-trained data and downstream task data, which affects the generalization performance of VL models. For adapting VL models to downstream tasks, recent works

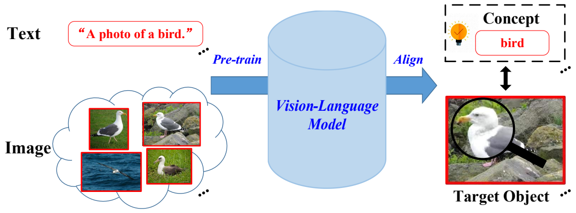

**Figure 1: Vision-language models align the concept to the target object by pre-training with massive matched image-text pairs.**

resort to prompt learning and feature adapters for fine-tuning VL models. CoOp [34] proposes to add learnable vectors into the text prompt, which tunes VL models to generate adaptive classification weights for downstream image data. Tip-Adapter [32] proposes to construct and fine-tune the key-value cache from the training samples set based on VL models for image recognition. Though achieving promising performance, the above methods generally ignore the discriminative visual features generation of the downstream data, which limits their adaptation ability. Inspired by the pre-training procedure of VL models, as shown in Figure 1, we observe that visual concepts are captured through the matching training of image-text pairs. In the process, the image encoder of the VL model is forced to focus on the target object in the image for aligning the concept in the text. Thus, inducing the VL model to ***concentrate on the new target object (concept)*** in the downstream task to extract discriminative visual feature becomes the core idea of our approach for ***improving the adaptation ability***.

Object information is generally mined in the fine-grained image recognition task. For example, Peng et al. [24] propose the typical object-part attention model for locating the object and selecting significant parts for classification. Sun et al. [28] propose to introduce the object structure information into the vision transformer to highlight significant regions and boost discriminative feature learning. Hu et al. [12] propose to utilize the recurrent attention multi-scale transformer to amplify object region for recognition.

Inspired by the above observations, we propose the fine-grained visual prompt learning of VL models for image recognition with few training samples, termed as ***FG-VPL***. Concretely, we introduce a learnable prompt token in the image encoder of the VL model and construct the fine-grained branch for focusing on the target object and extracting discriminative visual features. To further narrow the gap between the pre-trained data and downstream task data, we propose a two-pathway adaptive recognition module. On the one hand, the text features extracted by the text encoder of the VL model are projected as classification weights by the feature adapter. On the other hand, we utilize the visual information contained in the few-sample training set for querying test images. Therefore, the cross-modal matching knowledge from the VL model and the visual information of the few-sample training set can be fully utilized for improving image recognition performance.

The key contributions of this paper can be summarized as follows:

- We propose fine-grained visual prompt learning of vision-language models, which guides the vision-language model to focus on the target object of the downstream task data. Adaptive discriminative visual features are thus obtained for improving image recognition performance with few training samples.
- A two-pathway adaptive recognition module is proposed to utilize both the cross-modal matching knowledge in the vision-language model and the visual information of the few-sample training set for classifying images, where the feature projection is adopted to narrow the gap between pre-trained data and downstream task data.
- Extensive experiments are conducted on 11 widely-used image recognition benchmark datasets, including generic object classification, fine-grained visual categorization, etc., which demonstrates the effectiveness of our proposed FG-VPL approach.

## 2 RELATED WORK

This section briefly reviews the related work of vision-language models and prompt learning.

### 2.1 Vision-Language Models

Recently, vision-language (VL) models have attracted more and more attention for their strong generic representation and generalization ability. The VL models can be roughly classified into two types according to the encoder types. The first kind of methods[3, 16, 17, 31], such as UNITER[3] and ALIGN[16], utilize the fusion encoder to model the cross-modal interaction between image and text, which can learn the matching relation between image and text sufficiently. However, they need to input all possible image-text pairs for inference, which limits their efficiency to a large extent. The second kind of methods [13, 19, 25], such as the well-known CLIP[25], utilize single-modal encoder for the image and text respectively. They align the image and text through contrastive learning in the embedding space. Image and text features can be computed separately, improving the computation efficiency. With the help of large-scale image-text pairs, VL models show promising generalization ability to a wide range of downstream tasks, which has significant application value.

However, the existing gap between the pre-trained data and downstream task data significantly affects the vision-language model's transfer performance, which led to recent research works about adapting VL models to downstream tasks, such as typical prompt learning methods.

### 2.2 Prompt Learning

Prompt learning is first proposed in the natural language processing area. It aims to extract the useful information of large-scale pre-trained language models, such as GPT [26] and BERT [6], to adapt to downstream tasks in the prompt design way. Recently, Continuous prompt learning is studied in Prefix-tuning [18] to utilize the learnable vectors as prompt, which is optimized in an end-to-end training way. In the computer vision research area, CoOp [34] proposes to utilize the learnable vectors as the context of a text prompt, which helps generate adaptive classification weights for downstream task data. Based on CoOp, CoCoOp [33] proposes to utilize a neural network to generate an input-conditional vector for each image, which is added to the learnable context vectors of the text prompt for boosting the model's generalization ability. Besides, Zhang et al. [32] propose Tip-Adapter to construct the key-value cache model from the few training samples set, which generates the complementary prediction of the original CLIP's prediction through comparing the cache model's information with the test image. Wherein the cache keys can be further fine-tuned to improve the recognition performance.

Though having achieved promising performance, the above methods generally focus on the text side, which generally neglects the generation of adaptive visual features to downstream tasks. We propose fine-grained visual prompt learning to induce the vision-language model to focus on the target object, which helps the model capture discriminative visual information for the downstream image recognition task.

## 3 APPROACH

The overview of our proposed FG-VPL approach is shown in Figure 2. We select the well-known vision-language model ***CLIP as the backbone***. In this section, we first review the framework of CLIP and elaborate on the proposed fine-grained visual prompt (FVP) module in Sect. 3.1, which help extract discriminative visual features. The two-pathway adaptive recognition (TAR) module is introduced in Sect. 3.2, which is utilized for the final classification. It is worth noting that ***our proposed FG-VPL approach can be employed on any two-stream vision-language model*** for broad application.

### 3.1 Fine-grained Visual Prompt

The CLIP model is a typical two-stream pre-trained vision-language (VL) model, which aims to learn a common embedding space for aligning the paired image and text. Specifically, the CLIP model comprises an image encoder and a text encoder. The text encoder projects the natural language text into text feature representation with the transformer network [29]. The image encoder projects the image into image feature representation with ResNet 50 [10] or ViT [7]. Contrastive learning is adopted in the pre-training process, which pulls together the matched image-text pairs and pushes away the unmatched ones in the common embedding space. With the help of 400 million image-text pairs, CLIP is empowered to align massive concepts and target objects and learn powerful generic representation. Therefore, CLIP can be naturally generalized to the downstream image recognition task. Given the input image $\boldsymbol{x}$, the image encoder in CLIP projects it into the image feature $\boldsymbol{f}(x)$. The class name, such as "dog", is added to a prompt such as " a photo of a {CLASS}.", which is further projected into $\boldsymbol{w_i}$ by CLIP's text encoder. $k$ is the number of classes and $i \in \{1, 2, ..., k\}$. The prediction probability of the input image is calculated as follows:

$$p(y = i|x) = \frac{\exp(sim(\boldsymbol{w_i}, \boldsymbol{f(x)})/\tau)}{\sum\limits_{j=1}^{k} \exp(sim(\boldsymbol{w_j}, \boldsymbol{f(x)})/\tau)}, \quad (1)$$

where $y$ is the prediction label, $sim(,)$ denotes the cosine similarity and $\tau$ is the temperature parameter. The set of $\boldsymbol{w_i}$ is denoted as $\boldsymbol{W}$, which can be viewed as classification weights for the class. However, there exists the domain gap between the pre-trained data and downstream task data, which affects the generalization performance of the vision-language model.

To address the problem, we propose the fine-grained visual prompt (FVP) module, which guides the vision-language model to focus on the target object in the downstream data. The adaptive discriminative visual feature is thus extracted for improving the recognition performance. As shown in the right part of Figure 2, the vision transformer is selected as the vision-language model's image encoder, and a learnable fine-grained visual prompt is input into the first transformer layer. Specifically, the image is first split into patches as the input of the image encoder, which is denoted as $\boldsymbol{x} \in R^{H \times W \times 3}$, where the $H$ and $W$ are the height and width of the image, respectively. The patch size is denoted as $P$ and the number of patches is thus obtained as $N = \lfloor \frac{H}{P} \rfloor \times \lfloor \frac{W}{P} \rfloor$. The $i_{th}$ image patch $\boldsymbol{x}^i_{patch}$ is then linearly projected into the $D$-dimensional embedding vector. After combing the class token $\boldsymbol{x_{cls}}$ and adding the position embeddings, the input token sequence is $[\boldsymbol{x_{cls}}, \boldsymbol{F}(x^1_{patch}), \boldsymbol{F}(x^2_{patch}), ..., \boldsymbol{F}(x^N_{patch})]$. $\boldsymbol{x_{cls}}$ is utilized to represent the whole image. We introduce a learnable $D$-dimensional fine-grained visual prompt $\boldsymbol{x_{FP}}$ in the last of the token sequence. And the input of the first transformer layer is denoted as:

$$\boldsymbol{z_0} = [\boldsymbol{x}^0_{cls}, \boldsymbol{F^0}(x^1_{patch}), \boldsymbol{F^0}(x^2_{patch}), ..., \boldsymbol{F^0}(x^N_{patch}), \boldsymbol{x}^0_{FP}]. \quad (2)$$

The transformer layer comprises a multi-head self-attention (MSA) module and a feed-forward neural network module. When $\boldsymbol{z_{k-1}}$ is input into the $k_{th}$ transformer layer, the output is calculated as follows:

$$z^{'}_k = LN(MSA(\boldsymbol{z_{k-1}}) + \boldsymbol{z_{k-1}}), \quad (3)$$

$$\boldsymbol{z_k} = LN(FFN(z^{'}_k) + z^{'}_k), \quad (4)$$

where $LN(\cdot)$ is the layer normalization operation [1]. The global interaction among the class token, image patch tokens, and the fine-grained visual prompt are conducted through the above self-attention mechanism in each transformer layer.

Self-attention weights in the transformer layer indicate the influence of each image patch token on the class token, which is used for the final classification. Intuitively, the self-attention weights are highly correlated with whether the image patch contains the target object information, which is further utilized for detecting discriminative patches within the object. Specifically, there are $H$ attention heads in the $l_{th}$ transformer layer. $\boldsymbol{Q}$ and $\boldsymbol{K}$ are projected query vectors and key vectors of tokens, then the self-attention weights among tokens are obtained as follows:

$$\boldsymbol{Att^l_h} = softmax(\frac{\boldsymbol{QK^T}}{\sqrt{H}}), \quad (5)$$

where the $\boldsymbol{Att^l_h} \in R^{(N+2)\times(N+2)}$ $(h = 1, 2, ..., H)$ denotes the attention weight of $h_{th}$ attention head and $N$ is the number of image patch tokens. Considering the $L - 1$ transformer layers, the total attention weight of $h_{th}$ attention head is obtained by recursive matrix multiplication:

$$\boldsymbol{Att_h} = \prod_{l=1}^{L-1} \boldsymbol{Att^l_h}. \quad (6)$$

The $\boldsymbol{Att_h}$ is obtained based on multi-layer self-attention weights analyses, which can better help locate the target object. In concrete,

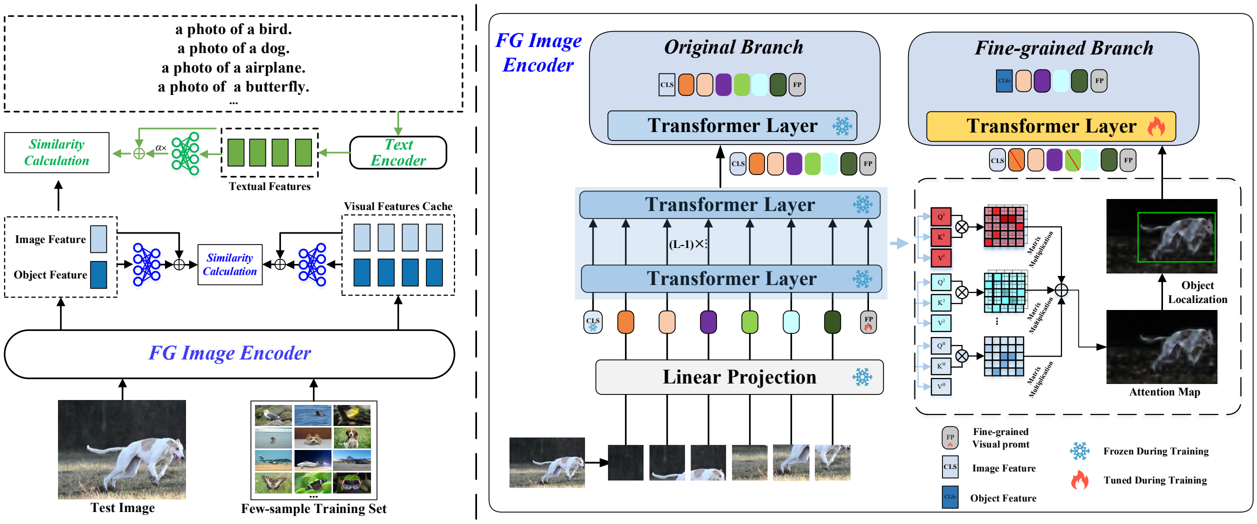

**Figure 2: The framework of our proposed *FG-VPL* approach. The left part describes the two-pathway adaptive recognition *(TAR)*. The right part depicts introducing the fine-grained visual prompt *(FVP)* into the image encoder and constructing the original and fine-grained branch, dubbed as FG Image Encoder in the figure.**

the attention weight between the image patch token and the class token indicates the significance of the image patch token, which is extracted from $\boldsymbol{Att_h}$ and denoted as $\boldsymbol{AM_h^{cls}} \in R^{N\times 1}$. Considering all the $H$ attention heads, we obtain the final attention map as follows:

$$AM = \sum_{h=1}^{H} \boldsymbol{AM_h^{cls}} \,. \tag{7}$$

As shown in the right part of Figure 2, the attention map $AM$ presents the significance of each image patch where the target object patches are highlighted and the background patches are ignored. Thus, the target object, i.e., the dog in the figure, is accurately located with Otsu binarization operation [22] and the largest connected area analysis.

To obtain comprehensive discriminative visual information, we construct two branches, i.e., the original branch and the fine-grained branch, to extract features from the input image and the target object, respectively. Wherein the fine-grained branch is constructed with a new transformer layer to conduct the information interaction among the class token, image patch tokens within the object, and the fine-grained visual prompt. Specifically, the input token sequence of the original branch is as follows:

$$\boldsymbol{z_{L-1}} = [\boldsymbol{x_{cls}^{L-1}}, \boldsymbol{F^{L-1}}(x_{patch}^{1}), \boldsymbol{F^{L-1}}(x_{patch}^{2}), ..., \boldsymbol{F^{L-1}}(x_{patch}^{N}), \boldsymbol{x_{FP}^{L-1}}] \tag{8}$$

The information interaction among the class token, all the image patch tokens, and the fine-grained visual prompt is conducted in the original branch with Eq. 3 and Eq. 4. The class token of the original branch is then viewed as the image feature $\boldsymbol{f}(x)$, which captures the global image information and can be improved by tuning the fine-grained visual prompt. The input token sequence of the fine-grained branch is as follows:

$$\boldsymbol{z_{L-1}^{Obj}} = [\boldsymbol{x_{cls}^{L-1}}, \boldsymbol{F^{L-1}}(x_{patch}^{O_1}), \boldsymbol{F^{L-1}}(x_{patch}^{O_2}), ..., \boldsymbol{F^{L-1}}(x_{patch}^{O_M}), \boldsymbol{x_{FP}^{L-1}}] \tag{9}$$

where $\boldsymbol{x_{patch}^{O_1}}$, $\boldsymbol{x_{patch}^{O_2}}$, and $\boldsymbol{x_{patch}^{O_M}}$ denote the image patch tokens within the target object. The information interaction within the object is conducted in the fine-grained branch with the newly added transformer layer using Eq. 3 and Eq. 4. By this means, the class token of the fine-grained branch, denoted as $\boldsymbol{f_O}(x)$, can capture discriminative information of the target object relevant to the downstream data through tuning the fine-grained visual prompt and the newly added transformer layer during training.

Thus, the fine-grained image encoder (FG Image Encoder for short) in Figure 2 is constructed by introducing the fine-grained visual prompt and designing the fine-grained branch. FG Image Enecoder focuses on the target object and extracts comprehensive discriminative visual features for image classification, which can be jointly optimized with the subsequent two-pathway adaptive recognition module.

## 3.2 Two-pathway Adaptive Recognition

The vision-language model contains cross-modal matching knowledge for generalizing to downstream tasks. Meanwhile, in the downstream task of image recognition with few training samples, the visual information in the training set also plays an important role to compare with the test image for classification. Thus, leveraging the above cross-modal knowledge and visual information of the training set is an effective way to improve classification performance. However, there exists the domain gap between the pre-trained data and downstream task data to affect the classification accuracy. Thus, we propose the two-pathway adaptive recognition (TAR) module

| Dataset | Classes | Train | Val | Test | Handcrafted text prompt |
|---|---|---|---|---|---|
| ImageNet[5] | 1000 | 1.28M | N/A | 50,000 | itap of a {CLASS}. a bad photo of the {CLASS}. a origami {CLASS}. a photo of the large {CLASS}. a {CLASS} in a video game. art of the {CLASS}. a photo of the small {CLASS}. |
| Caltech101 [8] | 100 | 4,128 | 1,649 | 2,465 | a photo of a {CLASS}. |
| OxfordPets [23] | 37 | 2,944 | 736 | 3,669 | a photo of a {CLASS}, a type of pet. |
| StanfordCars [15] | 196 | 6,509 | 1,635 | 8,041 | a photo of a {CLASS}. |
| Flowers102 [21] | 102 | 4,093 | 1,633 | 2,463 | a photo of a {CLASS}, a type of flower. |
| Food101 [2] | 101 | 50,500 | 20,200 | 30,300 | a photo of {CLASS}, a type of food. |
| FGVCAircraft [20] | 100 | 3,334 | 3,333 | 3,333 | a photo of a {CLASS}, a type of aircraft. |
| SUN397 [30] | 397 | 15,880 | 3,970 | 19,850 | a photo of a {CLASS}. |
| DTD [4] | 47 | 2,820 | 1,128 | 1,692 | {CLASS} texture. |
| EuroSAT [11] | 10 | 13,500 | 5,400 | 8,100 | a centered satellite photo of {CLASS}. |
| UCF101 [27] | 101 | 7,639 | 1,898 | 3,783 | a photo of a person doing {CLASS}. |

**Table 1: The detailed statistics of 11 image recognition benchmark datasets. Handcrafted text prompts in Tip-Adapter [32] are adopted in our FG-VPL approach.**

with feature adapters to narrow the domain gap, as shown in the left part of Figure 2.

In the first path, the cross-modal knowledge stored in the vision-language model is utilized for the similarity calculation between the test image and the text that contains the class name, which is marked with green in the figure. Concretely, the image feature $f(x_{test})$ and object feature $f_O(x_{test})$ of the test image are extracted by the FG Image Encoder, which is described in Sect. 3.1. The class name is added to the text prompt template for generating the text feature $\boldsymbol{w_i}$ in Eq. 1 with the text encoder of the vision-language model. One linear layer with an activation function denoted as $p(\cdot)$ is adopted as a feature adapter to adapt the textual features generated by the vision-language model to the downstream task image in the embedding space. The new textual feature $\boldsymbol{w_i^{new}}$ is calculated as follows:

$$\boldsymbol{w_i^{new}} = \alpha \times p(\boldsymbol{w_i}) + \boldsymbol{w_i}, \tag{10}$$

where $\alpha$ is a modulating parameter. The set of $\boldsymbol{w_i^{new}}$ is denoted as $\boldsymbol{W^{new}}$, and two prediction logits are obtained for the test image feature $f(x_{test})$ and object feature $f_O(x_{test})$, respectively.

$$\boldsymbol{logits_{p1}^{I}} = f(x_{test})(\boldsymbol{W^{new}})^T, \tag{11}$$

$$\boldsymbol{logits_{p1}^{O}} = f_O(x_{test})(\boldsymbol{W^{new}})^T. \tag{12}$$

In the second path, the visual information of the training set is utilized for similarity calculation with the test image, which is marked with blue in the left part of Figure 2. We first extract the image features $\boldsymbol{f(x_{set})}$ and object features $\boldsymbol{f_O(x_{set})}$ from the training set, where the object features are the average value of image patch tokens within the object in the original branch. The visual features are then projected using one linear layer with an activation function denoted as $q(\cdot)$, which aims to adapt the visual features from the vision-language models to downstream task data for similarity calculation. The new visual features for the training set and test image are:

$$\boldsymbol{f^{new}}(x_{set}) = q(f(x_{set})) + f(x_{set}), \tag{13}$$

$$f_O^{new}(x_{set}) = q(f_O(x_{set})) + f_O(x_{set}), \tag{14}$$

$$\boldsymbol{f^{new}}(x_{test}) = q(f(x_{test})) + f(x_{test}), \tag{15}$$

$$f_O^{new}(x_{test}) = q(f_O(x_{test})) + f_O(x_{test}). \tag{16}$$

The one-hot vectors of the training set class labels are denoted as $L_{set}$ and two prediction logits are obtained as follows:

$$\boldsymbol{logits_{p2}^{I}} = \varphi(\boldsymbol{f^{new}}(x_{test})(\boldsymbol{f^{new}}(x_{set}))^T)\boldsymbol{L_{set}}, \tag{17}$$

$$\boldsymbol{logits_{p2}^{O}} = \varphi(f_O^{new}(x_{test})(f_O^{new}(x_{set}))^T)\boldsymbol{L_{set}}, \tag{18}$$

where $\varphi(s) = \exp(-(1-s))$ is a normalization function. Considering the above two recognition paths, we can get the prediction vectors for the image $\boldsymbol{logits^I} = \boldsymbol{logits_{p1}^I} + \boldsymbol{\lambda logits_{p2}^I}$ and object $\boldsymbol{logits^O} = \boldsymbol{logits_{p1}^O} + \boldsymbol{\lambda logits_{p2}^O}$, and the final prediction vector for the test image is calculated as follows:

$$\boldsymbol{logits} = \boldsymbol{logits^I} + \beta \times \boldsymbol{logits^O}, \tag{19}$$

where $\beta$ is a modulating parameter. In the training stage, it is noted that the so-called test image is from the training set, and the cross entropy loss function is utilized for optimizing the proposed whole FG-VPL model.

Overall, the cross-modal knowledge in the vision-language model and the visual information of the training set are fully utilized with our proposed two-pathway adaptive recognition (TAR) module while narrowing the domain gap. Besides, the matching calculation in TAR also facilitates the FG Image Encoder in the FVP module to focus on the target object more accurately, which benefits each other for improving classification performance.

# 4 EXPERIMENTS

In this section, we first introduce the dataset setup, evaluation metric, and implementation details. Comparison experiments and analyses are conducted on **11** benchmark image recognition datasets for evaluating the performance of our proposed FG-VPL approach. Besides, ablation experiments, experiments on different vision backbones, and parameter experiments are conducted to verify the effectiveness of each component of the proposed approach.

## 4.1 Dataset Setup and Evaluation Metric

To evaluate our proposed method sufficiently, we conduct experiments on 11 public image recognition benchmark datasets used in CLIP [25]. The 11 datasets contain ImageNet [5], Caltech101 [8], OxfordPets [23], StandfordCars [15], Flowers102 [21], Food101 [2],

| Methods | Vision Backbone | Training Samples Setup | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 4 | 8 | 16 |
| Zero-shot CLIP [25] | ResNet 50 | 60.3 | | | | |
| Zero-shot CLIP [25] | ViT-B_16 | 68.7 | | | | |
| Linear-probe CLIP [25] | ResNet 50 | 22.2 | 31.9 | 41.2 | 49.5 | 56.1 |
| Linear-probe CLIP [25] | ViT-B_16 | 32.0 | 44.6 | 54.4 | 62.0 | 67.7 |
| CLIP-Adapter [9] | ResNet 50 | 61.2 | 61.5 | 61.8 | 62.7 | 63.6 |
| CoOp [34] | ResNet 50 | 57.2 | 57.8 | 60.0 | 61.6 | 63.0 |
| Tip-Adapter-F [32] | ResNet 50 | 61.3 | 61.7 | 62.5 | 64.0 | 65.5 |
| Tip-Adapter-F* [32] | ViT-B_16 | 69.8 | 70.0 | 70.8 | 71.9 | 73.7 |
| Our FVP Module | ViT-B_16 | 69.7 | 70.7 | 71.3 | 71.5 | 74.5 |
| Our TAR Module | ViT-B_16 | 70.0 | 70.6 | 71.7 | 72.9 | 74.5 |
| **Our FG-VPL Approach** | **ViT-B_16** | **70.3** | **71.1** | **72.5** | **73.0** | **75.5** |

**Table 2: Image classification accuracy (%) comparison with other state-of-the-art methods on the ImageNet dataset under various training sample settings. Our proposed FG-VPL approach surpasses all the comparison methods consistently in all settings. In the table, * denotes the results obtained by running the officially released code of the comparison method. The bold value indicates the best classification accuracy, and the underlined value indicates the sub-optimal classification accuracy.**

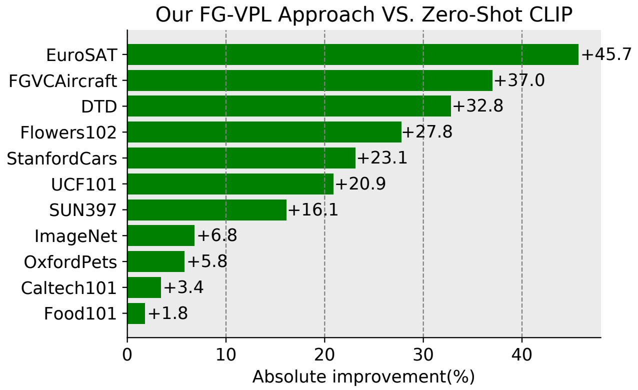

**Figure 3: Image classification accuracy (%) gains of our proposed FG-VPL approach over the zero-shot CLIP method on 11 benchmark datasets under the setting of 16 training samples per class.**

FGVCAircraft [20], SUN397 [30], DTD [4], EuroSAT [11], and UCF 101 [27], which covers generic object classification, fine-grained visual categorization, scene recognition, texture image classification, satellite image classification, and action recognition. The detailed statistics of 11 datasets are shown in Table 1. For fair performance comparison, all the models, except the Zero-shot CLIP, are trained with 1, 2, 4, 8, 16 images per class as the training set, dubbed as 1-sample, 2-sample, 4-sample, 8-sample, 16-sample setting, and *tested on the full test set.*

The widely used image classification accuracy is adopted to evaluate the performance of our proposed FG-VPL approach and other comparison methods.

## 4.2 Implementation Details

Our FG-VPL approach selects the two-stream vision-language model CLIP [25] as the backbone, where ViT-B_16 is selected as the image encoder backbone, and the transformer network is selected as the textual encoder. Pre-trained weights from CLIP are loaded and frozen during the training process. The same data preprocessing method in CLIP and Tip-Adapter [32] is adopted, comprising resizing, random horizontal flip, and so on. Following Tip-Adapter, we utilize the handcrafted text prompt ensemble for ImageNet and the single handcrafted text prompt for the other 10 datasets. In the training stage, we set the batch size to 256, and the number of training epochs is set as 50 for the ImageNet dataset, 400 for the EuroSAT dataset, and 100 for other datasets, respectively. $\alpha$ is set as 0.05 and $\beta$ is set to be 0.5. $\lambda$ is set as 1 as default, which can be tuned with the validation set. AdamW [14] is adopted as the optimizer, and the learning rate is set as 1e-3 with a cosine annealing scheduler. All the experiments are conducted with Pytorch on an NVIDIA A40 GPU.

## 4.3 Comparison Experiments

We conduct extensive comparison experiments with state-of-the-art (SOTA) methods on 11 benchmark datasets under 1-sample, 2-sample, 4-sample, 8-sample, 16-sample settings following CLIP [25] and Tip-Adapter [32]. The comparison with other methods is fair, and the results are shown in Table 2, Figure 3, and Figure 4. We can observe that:

- On the challenging ImageNet dataset, our proposed FG-VPL approach outperforms all the comparison methods in all few-sample settings, as shown in Table 2. The performance of our FG-VPL is improved with the increase of training samples, which presents our model's potential in adapting the vision-language model to the downstream image recognition task with different amounts of training data. Tip-adapter-F [32] constructs the cache model to utilize the information from the few-sample training set for image recognition. By contrast, our FG-VPL approach captures the visual object information related to the downstream image recognition task with the fine-grained prompt design and narrows the domain gap between the pre-trained data and downstream task data with the two-pathway adaptive recognition module. Thus, we achieve better recognition performance than Tip-adapter-F with **1.8%** accuracy gain in the 16-sample setting. Our FG-VPL approach also outperforms CoOp [34] by

| Methods | ViT-B_32 | ViT-B_16 |
|---|---|---|
| Zero-shot CLIP [25] | 63.8 | 68.7 |
| CLIP-Adapter [9] | 66.2 | 71.1 |
| CoOp [34] | 66.9 | 71.9 |
| Tip-Adapter [32] | 65.6 | 70.8 |
| Tip-Adapter-F [32] | 68.7 | 73.7 |
| **Our FG-VPL Approach** | **70.1** | **75.5** |

**Table 3: Image classification accuracy (%) comparison with other state-of-the-art methods on the ImageNet with different vision backbones under the setting of 16 training samples per class. ViT-B_32 and ViT-B_16 denote the ViT-Base network with $32 \times 32$ and $16 \times 16$ as the input image patch size, respectively.**

a margin of **3.6%** in the 16-sample setting with adopting the ViT-B_16 as the vision backbone, as shown in Table 3. Compared with the learnable vectors in the text prompt proposed in the CoOp which affects the classifier of the vision-language model, our proposed fine-grained visual prompt learning can directly guide the model to focus on the target object, which obtains adaptive discriminative visual features for the downstream image recognition task to achieve higher accuracy.

- Figure 4 shows the image recognition performance comparison on all the 11 datasets described in Sect. 4.1. Our proposed FG-VPL approach nearly achieves the best performance in all the few-sample settings. On the average classification performance metric over 11 datasets, our FG-VPL approach achieves the best **85.6%** classification accuracy, bringing **1.8%**, **1.8%**, **2.9%**, **3.1%**, and **4.2%** performance gains over the sub-optimal Tip-adapter-F method in the settings of 1, 2, 4, 8, 16 training samples per class, respectively. It shows the effectiveness and generalization ability of our proposed FG-VPL approach. With more training samples, fine-grained object information can be captured more accurately, and the feature adapters in the two-pathway adaptive recognition module can be trained more sufficiently, which brings higher performance gains.
- Figure 3 shows the absolute improvements brought by our proposed FG-VPL approach over Zero-shot CLIP. Significant gains can be observed on various datasets. For example, on the EuroSAT dataset, our FG-VPL approach brings **45.7%** image recognition accuracy gain. The performance gains on most of the fine-grained datasets, i.e., FGVCAircraft, Flowers102, StanfordCars, are also significant (over **20%**), which corresponds with our model design. We also achieve promising image recognition accuracy improvements on challenging datasets such as ImageNet (**6.8%**) and SUN397 (**16.1%**). Overall, our proposed method can significantly improve the adaptation ability of the vision-language model to various downstream image recognition datasets with only a few training samples.

## 4.4 Experiments on Different Vision Backbones

For evaluating the performance of our proposed FG-VPL approach with different vision backbones of the image encoder in the vision-language model, we conduct comparison experiments based on the ViT-B_32 and ViT-B_16 vision backbones, respectively. The experimental results are shown in Table 3. Our proposed FG-VPL approach performs the best on both the vision backbones compared with other SOTA methods. By utilizing the stronger vision backbone, our FG-VPL shows better recognition accuracy, which indicates its extensibility to various vision backbones. The performance gains are also improved from the **1.4%** to **1.8%** compared with the suboptimal Tip-Adapter-F method. We attribute it to the more accurate image patch significance analyses brought by a stronger vision backbone, which helps the model capture the visual object information well to generate adaptive discriminative visual features for image recognition.

| $\alpha$ | 0.01 | 0.05 | 0.10 | 0.15 | 0.20 |
|---|---|---|---|---|---|
| ACC(%) | 75.2 | 75.5 | 75.3 | 75.2 | 75.1 |

**Table 4: The experiment about parameter $\alpha$ in Eq.10 on the ImageNet dataset in the 16-sample setting.**

| $\beta$ | 0.10 | 0.30 | 0.50 | 0.70 | 0.90 |
|---|---|---|---|---|---|
| ACC(%) | 75.0 | 75.4 | 75.5 | 75.2 | 75.0 |

**Table 5: The experiment about parameter $\beta$ in Eq.19 on the ImageNet dataset in the 16-sample setting.**

## 4.5 Ablation Experiments

Ablation studies on the proposed two components of our FG-VPL approach, i.e., fine-grained visual prompt (FVP) module and two-pathway adaptive recognition (TAR) module, are conducted on the ImageNet dataset in the 16 training samples per class setting. Experimental results are shown in Table 2, we can observe that:

- Compared with the Zero-shot CLIP method, FVP and TAR bring performance improvements in all the few-sample settings. With the increment of training samples, our FVP module can perform better than Tip-Adapter-F, which verifies the effectiveness of introducing a fine-grained visual prompt to utilize the target object feature with our FVP module. Comprehensive and discriminative visual features are thus captured to directly benefits image recognition.
- Our TAR module surpasses the Tip-Adapter-F in all the few-sample settings, which verifies the effectiveness of narrowing the domain gap between the pre-trained data and downstream task data with our two-pathway adaptive recognition module. The combination of the two proposed components further improves the model’s recognition accuracy to verify their complementarity. The matching process in TAR facilitates the target object localization of the FG Image Encoder in FVP, which generates stronger discriminative features for better matching performance in return.

## 4.6 Parameter Experiments

We conduct parameter experiments about $\alpha$ in Eq.10 and $\beta$ in Eq.19 on the ImageNet dataset in the 16-sample setting, and experimental results are shown in Table 4 and Table 5. The proposed FG-VPL approach achieves the best performance when $\alpha$ is set as 0.05. We attribute it to the strong language representation ability of the text encoder of the vision-language model, which only needs a slight adjustment to adapt to downstream tasks. $\beta$ is set as 0.5 to achieve

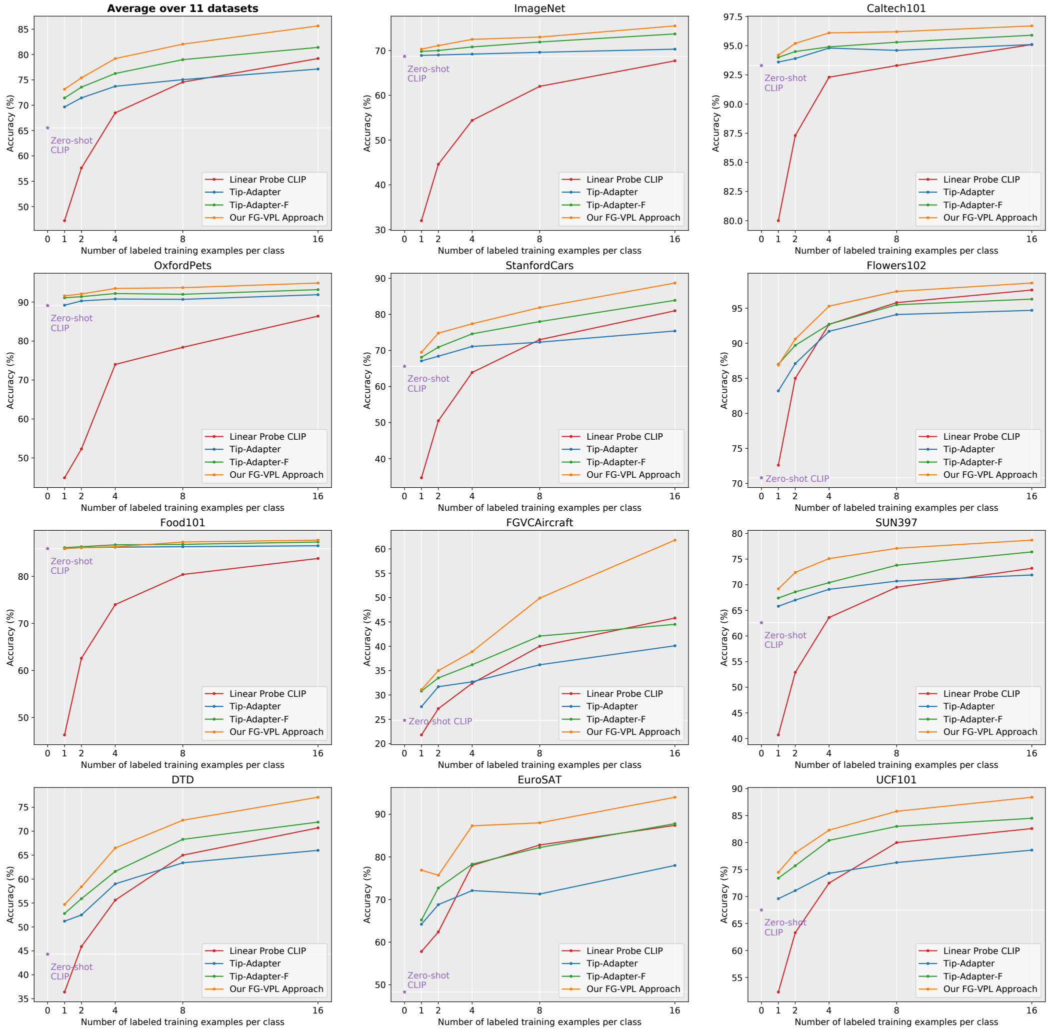

**Figure 4: Image classification accuracy (%) of different methods on the 11 benchmark datasets, which all adopt the ViT-B_16 as the vision backbone of the image encoder. The results of comparison methods are obtained by running their officially released codes. Our proposed FG-VPL approach achieves the best performance, which brings significant gains compared with the state-of-the-art method Tip-Adapter-F.**

the highest image recognition accuracy. It indicates the importance of the object feature extracted by our approach, which verifies the effectiveness of our fine-grained visual prompt design.

## 5 CONCLUSION

In this paper, we propose fine-grained visual prompt learning of vision-language models for image recognition with few training samples. The fine-grained visual prompt is introduced into the image encoder of the vision-language model to induce the model to focus on the target object of the downstream task, which contributes to generating discriminative features for image recognition. A two-pathway adaptive recognition module is proposed to utilize both the cross-modal matching knowledge from the vision-language model and the visual information from the few-sample training set for classifying test images. Wherein the feature adapters are designed to narrow the domain gap between pre-trained data and downstream task data. The proposed two components benefit each other to improve the total image recognition performance. Extensive experiments on **11** public benchmark datasets verify the effectiveness of our proposed FG-VPL approach for adapting vision-language models to image recognition with few training samples.

In the future, we will introduce the knowledge graph that describes the target object attributes in a fine-grained way into the text prompt design, which is hopeful for further improving the generalization ability of vision-language models.

## 6 ACKNOWLEDGEMENT

This work was supported by the grants from the National Key R&D Program of China (2021YFF0901502) and the National Natural Science Foundation of China (61925201, 62132001, 62272013).

# REFERENCES

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European conference on computer vision*. Springer, 446–461.

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

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

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

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

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. 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*. IEEE, 178–178.

[9] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544* (2021).

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

[11] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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 (2019), 2217–2226.

[12] Yunqing Hu, Xuan Jin, Yin Zhang, Haiwen Hong, Jingfeng Zhang, Yuan He, and Hui Xue. 2021. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4239–4248.

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

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

[16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[17] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.

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

[19] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208* (2021).

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

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

[22] Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9, 1 (1979), 62–66.

[23] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3498–3505.

[24] Yuxin Peng, Xiangteng He, and Junjie Zhao. 2017. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing* 27, 3 (2017), 1487–1500.

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

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[27] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[28] Hongbo Sun, Xiangteng He, and Yuxin Peng. 2022. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5853–5861.

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

[30] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3485–3492.

[31] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3208–3216.

[32] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*. Springer, 493–510.

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

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