Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement

Xiangyang Zhu*[1], Renrui Zhang*†‡[2,3], Bowei He[1], Aojun Zhou[2], Dong Wang[3], Bin Zhao[3], Peng Gao‡[3]

* Equal contribution † Project leader ‡ Corresponding author

[1]City University of Hong Kong [2]The Chinese University of Hong Kong
[3]Shanghai Artificial Intelligence Laboratory

{xiangyzhu6-c, boweihe2-c}@my.cityu.edu.hk,
{zhangrenrui, gaopeng}@pjlab.org.cn

Abstract

*The popularity of Contrastive Language-Image Pre-training (CLIP) has propelled its application to diverse downstream vision tasks. To improve its capacity on downstream tasks, few-shot learning has become a widely-adopted technique. However, existing methods either exhibit limited performance or suffer from excessive learnable parameters. In this paper, we propose **APE**, an **A**daptive **P**rior r**E**finement method for CLIP's pre-trained knowledge, which achieves superior accuracy with high computational efficiency. Via a prior refinement module, we analyze the inter-class disparity in the downstream data and decouple the domain-specific knowledge from the CLIP-extracted cache model. On top of that, we introduce two model variants, a training-free APE and a training-required **APE-T**. We explore the trilateral affinities between the test image, prior cache model, and textual representations, and only enable a lightweight category-residual module to be trained. For the average accuracy over 11 benchmarks, both APE and APE-T attain state-of-the-art and respectively outperform the second-best by +1.59% and +1.99% under 16 shots with ×**30 less** learnable parameters. Code is available at https://github.com/yangyangyang127/APE.*

1. Introduction

The advent of contrastive visual-language pre-training has provided a new paradigm for multi-modal learning [16, 17, 22, 42]. Its popularity has been observed across diverse downstream vision tasks, including 2D or 3D classification [14, 39, 41, 9], segmentation [27, 48, 36, 44], and detection [38, 45, 29]. CLIP [26] is one of the most acknowledged contrastive visual-language models and has attained widespread attention for its simplicity and superiority. Pre-trained by massive image-text pairs sourced from the Internet, CLIP exhibits remarkable aptitude in aligning vision-language representations with favorable zero-shot performance on downstream tasks. To further enhance CLIP in low-data regimes, many efforts propose few-shot learning techniques with additional learnable modules upon the frozen CLIP for new semantic domains.

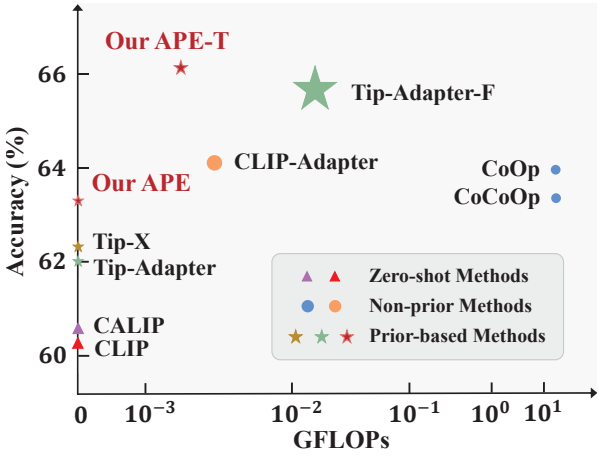
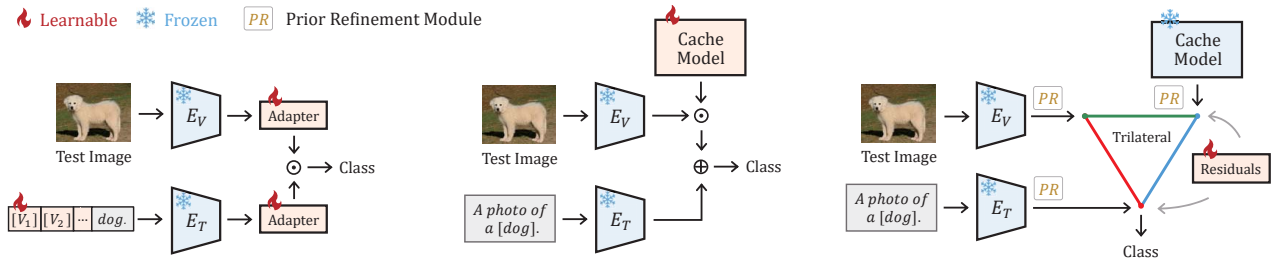


Figure 1: **Comparison of Accuracy, Training GFLOPs, and Learnable Parameters** on 16-shot ImageNet [3] classification. We compare the training GFLOPs including gradient back-propagation, and the icon sizes denote the number of learnable parameters. Our APE and APE-T achieve superior performance with high implementation efficiency.

As shown in Figure 2 (a) and (b), existing CLIP-based few-shot methods can be categorized as two groups concerning whether to explicitly construct learnable modules by CLIP's prior knowledge. **1) Non-prior Methods** ran-



(a) Non-prior methods (b) Prior-based methods (c) Our Proposed APE-T

Figure 2: **Comparison of Existing CLIP-based Few-shot Methods.** We only show the training-required model variants of prior-based methods and our APE-T. E_V, E_T denote CLIP's pre-trained visual and textual encoders, respectively.

domly initialize the learnable modules without CLIP's prior, and optimize them during few-shot training. For instance, CoOp series [47, 46] adopt learnable prompts before CLIP's textual encoder, and CLIP-Adapter [7] instead learns two residual-style adapters after CLIP. Such networks only introduce lightweight learnable parameters but suffer from limited few-shot accuracy, since no pre-trained prior knowledge is explicitly considered for the additional modules. **2) Prior-based Methods** construct a key-value cache model via CLIP-extracted features from the few-shot data and are able to conduct recognition in a training-free manner, including Tip-Adapter [40], Tip-X [33], and CaFo [43]. Then, they can further regard the cache model as a well-performed initialization and fine-tune the cache keys for better classification accuracy. These prior-based methods explicitly inject prior knowledge into the training process but are cumbersome due to the large cache size with enormous learnable parameters. We then ask the question, *can we integrate their merits to make the best of both worlds, namely, not only equipping efficient learnable modules, but also benefiting from CLIP's prior knowledge?*

To this end, we propose **A**daptive **P**rior r**E**finement, termed as **APE**, which efficiently adapts CLIP for few-shot classification by refining its pre-trained knowledge in visual representations. APE can not only achieve superior performance via CLIP's prior, but also consumes less computation resource than non-prior methods, as shown in Figure 1. We observe that not all CLIP's prior, *i.e.*, the extracted visual features of the cache model or test image, are significant for downstream tasks along the channel dimension. In Figure 3, we divide the feature channels of CLIP-extracted visual representations into two groups, and respectively visualize their similarity maps with the textual representation in ImageNet [3]. Features in the first group (a) can observe much better vision-language alignment than the second one (b). Motivated by this, we propose a prior refinement module to adaptively select the most significant feature channels by two criteria, inter-class similarity and variance. By maximizing the inter-class disparity in few-shot training data, the refined feature channels can discard redundant information and reduce the cache size with less memory cost.

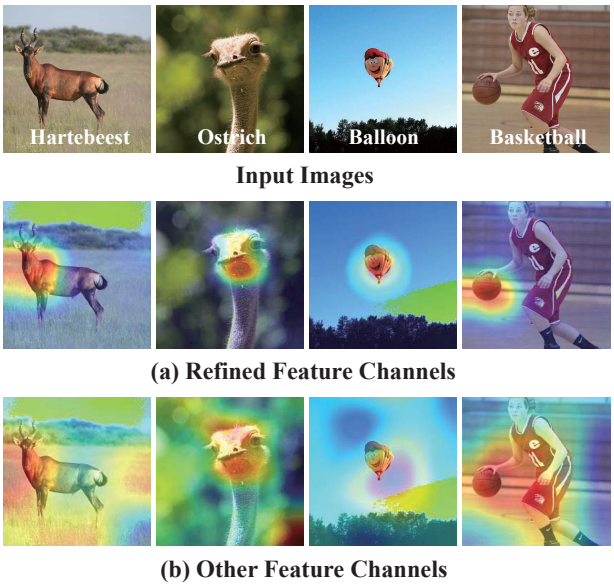


Figure 3: **Similarity Maps for Vision-language Alignment.** We utilize CLIP with ResNet-50 [11] visual encoder and refine 512 feature channels from 1024 ones, where the refined features are more attentive towards object targets.

On top of this, we present two variants of our approach, denoted as APE and APE-T. The first one is a training-free model that directly utilizes the refined cache model for inference. APE novelly explores the trilateral affinities between the test image, the refined cache model, and the textual representations for robust training-free recognition. The second one, APE-T (Figure 2(c)), simply trains lightweight category residuals on top, other than costly fine-tuning the entire cache model. Such category residuals further update the refined cache model and are shared between modalities to ensure the vision-language correspondence. Our APE and APE-T respectively achieve *state-of-the-art* performance compared with existing training-free and training-required methods on 11 few-shot benchmarks, surpassing the second-best by +1.59% and +1.99% for the average 16-shot accuracy.

The contributions of our work are summarized below:

- We propose **A**daptive **P**rior r**E**finement (**APE**), an adaption method of CLIP to explicitly utilize its prior knowledge while remain computational efficiency.
- After prior refinement, we explore the trilateral affinities among CLIP-extracted vision-language representations for effective few-shot learning.
- Our training-free APE and APE-T exhibit state-of-the-art performance on 11 few-shot benchmarks, demonstrating the superiority of our approach.

2. Related Work

Zero-shot CLIP. For a test image within the C-category dataset, CLIP [26] utilizes its encoders to extract the D-dimensional visual and textual representations, denoted as $\mathbf{f} \in \mathbb{R}^D$ and $\mathbf{W} \in \mathbb{R}^{C\times D}$, respectively. Then, the zero-shot classification logits are calculated by their similarity as

$$\mathbf{R}_{fW} = \mathbf{f}\mathbf{W}^\top \in \mathbf{R}^{1\times C}. \tag{1}$$

Based on such a zero-shot paradigm, existing few-shot adaption methods are categorized into two groups.

Non-prior Methods append additional learnable modules on top of CLIP and randomly initialize them without explicit CLIP's prior. Such methods include CoOp [47], CoCoOp [46], TPT [30], and CLIP-Adapter [7]. These approaches only introduce a few learnable parameters, *e.g.*, prompts or adapters, but attain limited accuracy for downstream tasks for lack of CLIP's prior knowledge.

Prior-based Methods can achieve higher classification accuracy by explicitly utilizing CLIP priors with a cache model, including Tip-Adapter [40], Causal-FS [18], Tip-X [33], and CaFo [43]. For a C-category dataset with K samples per class, a key-value cache model is built on top. The cache keys and values are initialized with the CLIP-extracted training-set features, $\mathbf{F} \in \mathbb{R}^{CK\times D}$, and their one-hot labels, $\mathbf{L} \in \mathbb{R}^{CK\times C}$, respectively. Then the similarity $\mathbf{R}_{fF}$ between the test image and training images is calculated as

$$\mathbf{R}_{fF} = \exp\left(-\beta(1-\mathbf{f}\mathbf{F}^\top)\right) \in \mathbb{R}^{1\times CK}, \tag{2}$$

where β is a smoothing scalar. Then, the relation $\mathbf{R}_{fF}$ is regarded as weights to integrate the cache values, *i.e.*, the one-hot labels $\mathbf{L}$, and blended with the zero-shot prediction as few-shot logits,

$$\text{logits} = \mathbf{R}_{fW} + \alpha\mathbf{R}_{fF}\mathbf{L}, \tag{3}$$

where α denotes a balance factor. In this way, prior-based methods can leverage the bilateral relations of $\mathbf{R}_{fW}$ and $\mathbf{R}_{fF}$ to achieve training-free recognition. On top of this, they can further enable the cache model to be learnable, and optimize the training-set features $\mathbf{F}$ during training. Although the initialization of learnable modules has explicitly incorporated CLIP's prior knowledge, these methods suffer from excessive parameters derived from the cache model.

Different from all above methods, our APE and APE-T can not only perform competitively via CLIP's prior knowledge, but also introduce lightweight parameters and computation resources by an adaptive prior refinement module.

3. Method

In Section 3.1, we first illustrate the prior refinement module in our APE by two inter-class metrics. Then in Section 3.2 and Section 3.3, we respectively present the details of our training-free and training-required variants, APE and APE-T, based on the refined representations.

3.1. Prior Refinement of CLIP

For a downstream dataset, the CLIP-extracted visual representations could comprise both domain-specific and redundant information along the channel dimension. The former is more discriminative at classifying downstream images, and the latter represents more general visual semantics. Therefore, we propose two criteria, inter-class similarity and variance, to adaptively select the most significant feature channels for different downstream scenarios.

3.1.1 Inter-class Similarity

This criterion aims to extract the feature channels that minimize the inter-class similarity, namely, the most discriminative channels for classification. For a downstream image, we represent its CLIP-extracted feature as $\mathbf{x} \in \mathbb{R}^D$, where D denotes the entire channel number and we seek to refine Q feature channels from D. We then set a binary flag $\mathbf{B} \in \{0,1\}^D$, where $B_k = 1$ $(k = 1, ..., D)$ denotes the k^{th} element x_k is selected, and $\mathbf{B}\mathbf{B}^\top = Q$. Now, our goal turns to find the optimal $\mathbf{B}$ to produce the highest inter-class divergence for downstream data.

For a C-category downstream dataset, we calculate the average similarity S between categories of all training samples. We adopt cosine similarity, $\delta(\cdot,\cdot)$, as the metric as

$$S = \sum_{i=1}^{C} P^i \sum_{\substack{j=1\\ j\neq i}}^{C} P^j \frac{1}{M^i}\frac{1}{M^j}\sum_{m=1}^{M^i}\sum_{n=1}^{M^j}\delta(\mathbf{x}^{i,m}, \mathbf{x}^{j,n}), \tag{4}$$

where $i, j \in \{1, ..., C\}$ represent two different categories. P^i, P^j denote the prior probability of the two categories, and M^i, M^j denote their total number of training samples.

However, calculating S for the whole dataset, even few shots, is computational expensive. Considering CLIP's contrastive pre-training, where the vision-language representations have been well aligned, the textual features of downstream categories can be regarded as a set of visual proto-

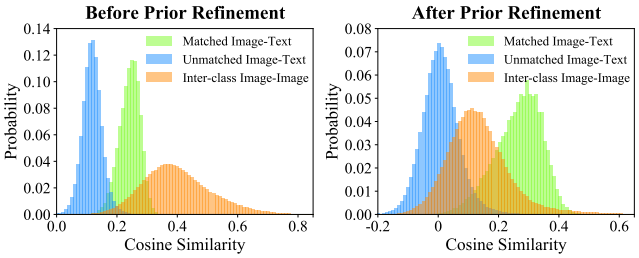


Figure 4: **The Effectiveness of Prior Refinement Module,** which minimizes the inter-class visual similarity and improves the text-image alignment.

types [31, 4, 13]. Such prototypes can approximate the clustering centers in the embedding space for the visual features of different categories [8, 35]. To obtain the textual features, we simply utilize the template 'a photo of a [CLASS]' and place all category names into [CLASS] as the input for CLIP. We then denote the textual features of downstream categories as $\mathbf{x}^i \in \mathbb{R}^D$, where $i \in \{1, ..., C\}$.

Therefore, we adopt these textual features to substitute the image ones for each category, which determines $M^1 = ... = M^C = 1$. Under open-world settings, we can also assume $P^1 = ... = P^C = \frac{1}{C}$. Then, we define the optimization problem to minimize the inter-class similarity,

$$\begin{aligned} \min_{\mathbf{B}} \quad & S = \frac{1}{C^2} \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} \delta(\mathbf{x}^i \odot \mathbf{B}, \mathbf{x}^j \odot \mathbf{B}), \\ s.t. \quad & \mathbf{B}\mathbf{B}^\top = Q, \end{aligned} \tag{5}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{x} \odot \mathbf{B}$ only selects the domain-specific feature channels. We further suppose the textual features have been L2-normalized, so we can simplify the cosine similarity as

$$S = \sum_{k=d_1}^{d_Q} S_k = \sum_{k=d_1}^{d_Q} \left(\frac{1}{C^2} \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} x_k^i \cdot x_k^j \right), \tag{6}$$

where $k = \{d_1, d_2, ..., d_Q\}$ denotes the indices of selected feature channels with $B_k = 1$, and $S_k = \frac{1}{C^2} \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} x_k^i \cdot x_k^j$ represents the average inter-class similarity of the k-th channel. From Equation 6, we observe that solving the optimization problem in Equation 5 equals selecting Q elements with the smallest average similarity. That is, we sort all D elements by their average similarities and select the top-Q smallest ones. In this way, we can derive the binary flag $\mathbf{B}$ and obtain the most discriminative feature channels for downstream classification.

3.1.2 Inter-class Variance

Besides the inter-class similarity, we introduce another criterion to eliminate the feature channels that remain almost constant between categories, which exhibit no inter-class difference with little impact for classification. For efficiency, we also adopt the category textual features, $\mathbf{x}^i \in \mathbb{R}^D$, where $i \in \{1, ..., C\}$, as visual prototypes for the downstream datasets. For the k^{th} feature channel, we formulate its inter-class variance as

$$V_k = \frac{1}{C} \sum_{i=1}^{C} (x_k^i - \bar{x_k})^2, \tag{7}$$

where $\bar{x_k} = \sum_{i=1}^{C} x_k^i$ denotes the average variance of the k^{th} channel across categories. Likewise to Equation 6, the variance criterion can also be regarded as a ranking problem, but instead selecting the top-Q channels with the highest variances. By this, we can effectively filter out the redundant and less informative channels within CLIP's prior knowledge for the downstream dataset.

Finally, we blend the similarity and variance criteria with a balance factor λ as the final measurement. For the k^{th} feature channel, we formulate it as

$$J_k = \lambda S_k - (1 - \lambda) V_k, \tag{8}$$

where $k = 1, ..., D$. The top-Q smallest J_k are selected as the final refined feature channels, which indicate the most inter-class divergence and discrimination.

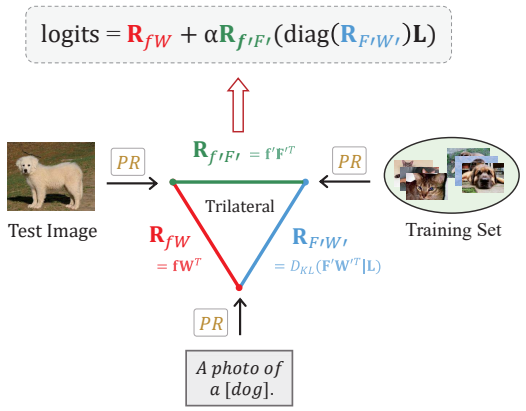
3.1.3 Effectiveness

Figure 4 shows the benefit brought by our adaptive refinement module. We conduct the refinement by textual features on ImageNet [3] validation set and visualize the statistic, where the category number C equals 1000. We experiment with ResNet-50 [11] as CLIP's visual encoder, where we refine $Q = 512$ feature channels from the entire $D = 1024$ ones. We compare three types of metrics referring to [33]. As shown, for the refined 512 feature channels, the inter-class similarity between images ('Inter-class Image-Image') has been largely reduced, indicating strong category discrimination. Meanwhile, our refinement better aligns the paired image-text features ('Matched Image-Text'), and pushes away the unpaired ones ('Unmatched Image-Text'), which enhances the multi-modal correspondence of CLIP for downstream recognition.

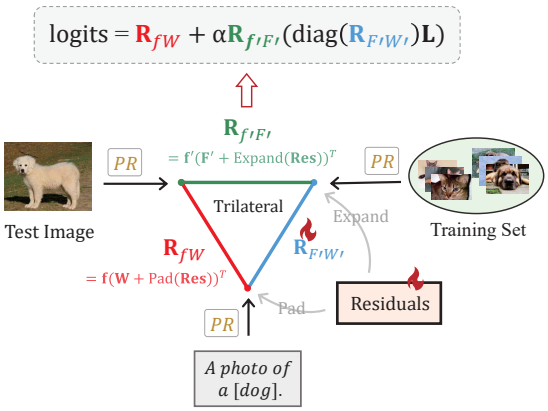
On top of the refined CLIP-extracted features, we present two few-shot adaption methods for CLIP, the training-free APE, and training-required APE-T.

3.2. Training-free APE

In essence, CLIP is a zero-shot similarity-based classifier, which relies on the distance between the test image and category textual representations in the embedding space. Considering this, our APE is based on the refined CLIP's prior and explores the trilateral embedding distances among the test image, downstream category texts, and the training images in the cache model, as shown in Figure 5.



Training-free APE



Training-required APE-T

Figure 5: **Framework of APE.** Based on the prior refinement (PR), APE explores trilateral relations of vision-language representations in a training-free manner.

Figure 6: **Framework of APE-T.** Our training-required variant appends learnable category residuals along with $\mathbf{R}_{F'W'}$ on top of APE for few-shot training.

For a C-way-K-shot downstream dataset with K training samples per category, we adopt CLIP to extract the L2-normalized features of the test image, category texts, and the training images, respectively denoted as $\mathbf{f} \in \mathbb{R}^{D}$, $\mathbf{W} \in \mathbb{R}^{C\times D}$, and $\mathbf{F} \in \mathbb{R}^{CK\times D}$. We then conduct our adaptive prior refinement module to obtain the most Q informative channels for the three features, formulated as $\mathbf{f}' \in \mathbb{R}^{Q}$, $\mathbf{W}' \in \mathbb{R}^{C\times Q}$, and $\mathbf{F}' \in \mathbb{R}^{CK\times Q}$. This not only discards the redundant signals in pre-trained CLIP, but also reduces the cache model with less computation cost during inference.

As for the trilateral relations, we first denote the relation between $\mathbf{f}$ and $\mathbf{W}$ as

$$\mathbf{R}_{fW} = \mathbf{f}\mathbf{W}^{\top} \in \mathbb{R}^{1\times C}, \tag{9}$$

which represents the cosine similarity between the test image and category texts, *i.e.*, the original classification logits of CLIP's zero-shot prediction as described in Section 2. Then, we formulate the affinities between $\mathbf{f}'$ and $\mathbf{F}'$ as

$$\mathbf{R}_{f'F'} = \exp\left(-\beta(1 - \mathbf{f}'\mathbf{F}'^{\top})\right) \in \mathbb{R}^{1\times CK}, \tag{10}$$

which indicates the image-image similarities from the cache model with a modulating scalar β, referring to the prior-based methods [40, 33]. Further, we take the relationship between $\mathbf{F}'$ and $\mathbf{W}'$ into consideration, and formulate their cosine similarity as $\mathbf{F}'\mathbf{W}'^{\top}$, which denotes CLIP's zero-shot prediction to the few-shot training data. To evaluate such downstream recognition capacity of CLIP, we calculate the KL-divergence, $D_{KL}(\cdot|\cdot)$, between CLIP's prediction and their one-hot labels, $\mathbf{L}$. We formulate it as

$$\mathbf{R}_{F'W'} = \exp\left(\gamma D_{KL}(\mathbf{L}|\mathbf{F}'\mathbf{W}'^{\top})\right) \in \mathbb{R}^{1\times CK}, \tag{11}$$

where γ serves as a smooth factor. $\mathbf{R}_{F'W'}$ can be regarded as a score for each training feature in the cache model, indicating its representation accuracy extracted by CLIP and how much it contributes to the final prediction.

Finally, integrating all trilateral relations, we obtain the overall classification logits of APE as

$$\text{logits} = \mathbf{R}_{fW} + \alpha\mathbf{R}_{f'F'}\Big(\text{diag}(\mathbf{R}_{F'W'})\mathbf{L}\Big), \tag{12}$$

where α serves as a balance factor and $\text{diag}(\cdot)$ denotes diagonalization. The first term represents the zero-shot prediction of CLIP and contains its pre-trained prior knowledge. The second term denotes the few-shot prediction from the cache model, which is based on the refined feature channels and $\mathbf{R}_{F'W'}$'s reweighing. Therefore, by the adaptive prior refinement and trilateral relation analysis, our APE can enhance few-shot CLIP both efficiently and effectively.

3.3. Training-required APE-T

To further improve the few-shot performance of APE, we introduce a training-required framework, APE-T, in Figure 6. Existing prior-based methods [40, 18] directly fine-tune all the training features in the cache model, which leads to large-scale learnable parameters and computational cost. In contrast, APE-T freezes the cache model, and only trains a group of additional lightweight category residuals, $\mathbf{Res} \in \mathbb{R}^{C\times Q}$, along with the cache scores $\mathbf{R}_{F'W'} \in \mathbb{R}^{1\times CK}$.

Specifically, the category residuals $\mathbf{Res}$ are implemented by a set of C learnable embeddings. Each embedding corresponds to a downstream category, which aims to optimize the refined Q feature channels for different categories during few-shot training. To preserve the vision-language correspondence in the embedding space, we apply $\mathbf{Res}$ to both textual features $\mathbf{W}$ and training-set features $\mathbf{F}'$.

For Equation 9, we first pad the Q-channel $\mathbf{Res}$ into D channels as $\mathbf{W}$ by filling the redundant channel indices with zero. Then, we element-wisely add the padded $\mathbf{Res}$ with

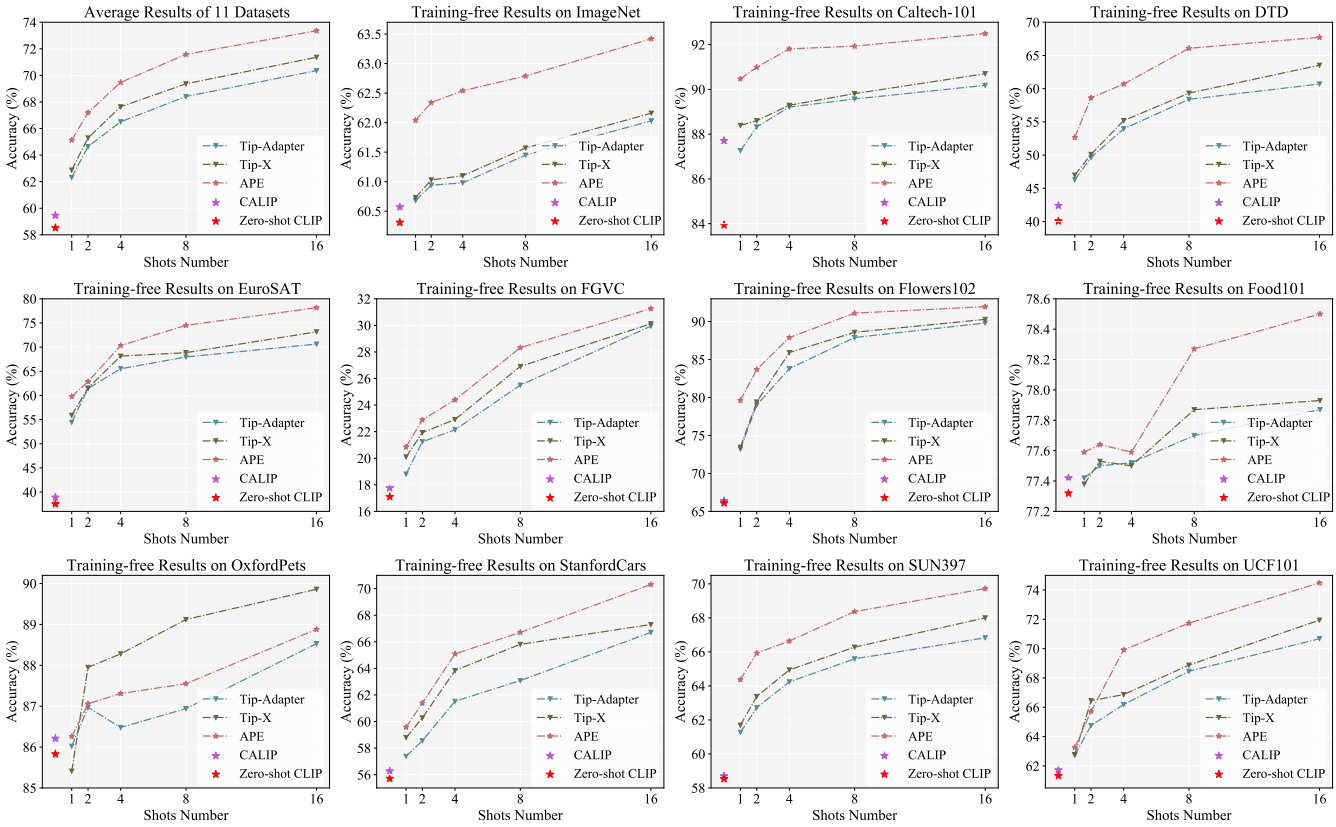


Figure 7: **Few-shot Performance of APE and other Training-free Methods** on 11 image classification datasets.

$\mathbf{W}$, which updates CLIP's zero-shot prediction by the optimized textual features, formulated as

$$\mathbf{R}_{fW} = \mathbf{f}\Big(\mathbf{W} + \text{Pad}(\mathbf{Res})\Big)^{\top}. \tag{13}$$

For Equation 10, we first broadcast the C-embedding $\mathbf{Res}$ into CK as $\mathbf{F}'$ by repeating the residual within each category. Then, we element-wisely add the expanded $\mathbf{Res}$ with $\mathbf{F}'$, which improves the cache model's few-shot prediction by optimizing training-set features, formulated as

$$\mathbf{R}_{f'F'} = \exp\Big(-\beta\big(1 - \mathbf{f}'(\mathbf{F}' + \text{Expand}(\mathbf{Res}))^{\top}\big)\Big).$$

For Equation 11, we directly enable the $\mathbf{R}_{F'W'}$ to be learnable during training without manual calculation. By this, APE-T can adaptively learn the optimal cache scores for different training-set features and determine which one to contribute more to the prediction.

Finally, we also leverage Equation 12 to obtain the final classification logits for APE-T. By only training such small-scale parameters, APE-T avoids the expensive fine-tuning of the cache model and achieves superior performance by updating the refined features for both modalities.

4. Experiments

In Section 4.1, we first present the detailed settings of APE and APE-T. Then in Section 4.2, we evaluate our approach on 11 widely-adopted benchmarks.

4.1. Experimental Settings

Datasets. We adopt 11 image classification benchmarks for comprehensive evaluation: ImageNet [3], Caltech-101 [6], DTD [2], EuroSAT [12], FGVCAircraft [21], Flowers102 [23], Food101 [1], OxfordPets [24], StanfordCars [15], SUN397 [37], and UCF101 [32]. In addition, ImageNet-Sketch [34] and ImageNet-V2 [28] are adopted to test the generalization ability. Given the few-shot training data from each dataset, we tune our models on the official validation set and evaluate the result on the full test set.

Experiment Settings. For APE and APE-T, we adopt ResNet-50 [11] as the visual encoder of CLIP by default, which outputs vision-language features with $D = 1024$ channels. We follow existing works [47, 40, 7] to conduct 1/2/4/8/16-shot learning and utilize the textual prompt in Tip-X [33] and CuPL [25]. For the prior refinement module, we set λ in Equation 8 to 0.7 for APE, and 0.2 for APE-T. To train APE-T, we adopt a batch size 256 and AdamW [20]

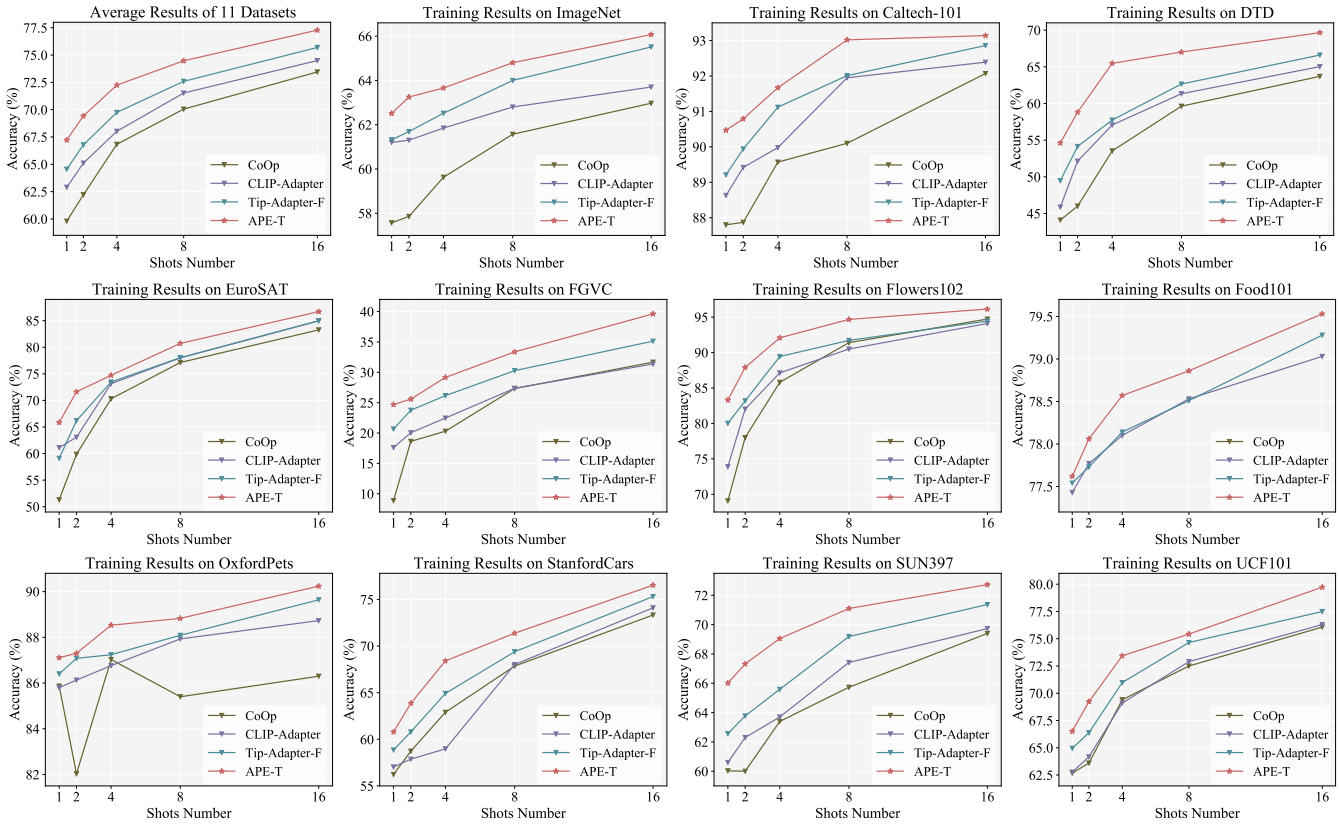


Figure 8: **Few-shot Performance of APE-T and other Training-required Methods** on 11 image classification datasets.

optimizer with a cosine annealing scheduler [19]. We utilize a learning rate of 0.0001 for ImageNet and Food101, and 0.001 for the rest datasets.

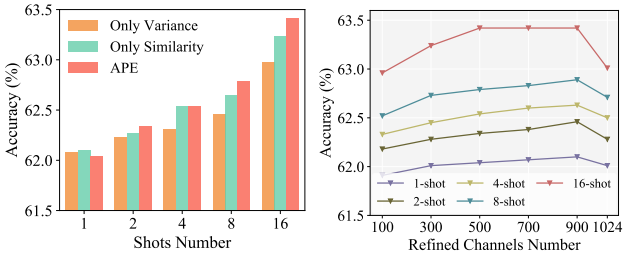
4.2. Performance Analysis

APE Results. Under the training-free settings, we compare our APE with Tip-Adapter [40] and Tip-X [33] in Figure 7. They are both prior-based methods and also training-free with a cache model. As shown by the average results across 11 datasets, APE exhibits consistent advantages over other methods for 1 to 16 shots, indicating our strong few-shot adaption capacity. Although We lag behind Tip-X on OxfordPets, remarkable gains are observed on DTD and EuroSAT datasets, *i.e.*, +7.03% and +7.53% over Tip-Adapter under the 16-shot setting. This demonstrates the effectiveness of refining domain-specific knowledge and exploiting the trilateral relations for different downstream scenarios.

APE-T Results. In Figure 8, we compare APE-T with three other training-required methods, CoOp [47], CLIP-Adapter [7], and Tip-Adapter-F [40]. Our APE-T outperforms existing ones on every benchmark and achieves *state-of-the-art* results for all few-shot settings. On average, APE-T's 16-shot accuracy of 77.28% surpasses Tip-Adapter-F by +1.59%. Particularly, we observe APE-T contributes to substantial improvements of +3.05% and +4.50% classification accuracies respectively on DTD and FGVCAircraft than Tip-Adapter-F. These superior results fully verify the significance of updating the refined feature channels by our learnable category residuals.

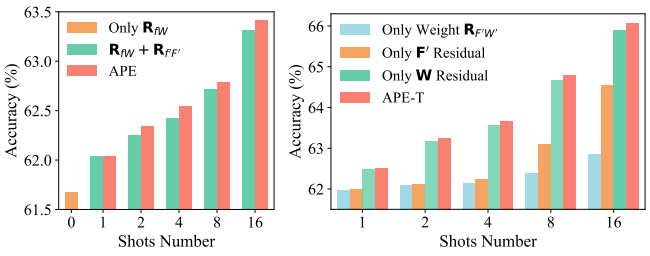
Computation Efficiency. We also compare the computing overhead between our approach and existing methods in Table 1. We test by an NVIDIA RTX A6000 GPU and report the performance on 16-shot ImageNet. As presented, CoOp involves the least learnable parameters but requires numerous training time and GFLOPs to back-propagate the gradients across the whole textual encoder. Tip-Adapter-F reduces the training time but brings large-scale learnable parameters by fine-tuning the full cache model along with no small GFLOPs for the gradients. In contrast, our APE-T not only attains the highest accuracy, but also achieves advantageous computation efficiency: **$\times$5000 fewer GFLOPs than CoOp, and $\times$30 fewer parameters than Tip-Adapter-F.**

Generalization Ability. In Table 2, we train the models by in-domain ImageNet and test their generalization ability on out-of-distribution datasets. With the best in-domain performance, our APE and APE-T both achieve



(a) Similarity and Variance Criteria (b) Channel Numbers to Refine

Figure 9: **Ablation Study on Prior Refinement.**



(a) Training-free APE (b) Training-required APE-T

Figure 10: **Ablation Study on APE and APE-T.**

Methods	Training	Epochs	GFLOPs	Param.	Acc.
Zero-shot					
CLIP [26]	-	-	-	-	60.33
CALIP [10]	-	-	-	-	60.57
Training-free					
Tip-Adapter [40]	0	0	-	0	62.03
Tip-X [33]	0	0	-	0	62.11
APE	0	0	-	0	**63.41**
Training-required					
CoOp [47]	14 h	200	>10	0.01 M	62.95
CLIP-Adapter [7]	50 min	200	0.004	0.52 M	63.59
Tip-Adapter-F [40]	5 min	20	0.030	16.3 M	65.51
APE-T	5 min	20	0.002	0.51 M	**66.07**

Table 1: **Comparison of Accuracy (%) and Efficiency** on 16-shot ImageNet [3]. "GFLOPs" are calculated during training with gradient back-propagation.

Datasets	**Source**	**Target**	
	ImageNet [3]	-V2 [3]	-Sketch [28]
Zero-Shot			
CLIP [26]	60.33	53.27	35.44
CALIP [10]	60.57	53.70	35.61
Training-free			
Tip-Adapter [40]	62.03	54.60	35.90
APE	**63.42**	**55.94**	**36.61**
Training			
CoOp [47]	62.95	54.58	31.04
CLIP-Adapter [7]	63.59	55.69	35.68
Tip-Adapter-F [40]	65.51	57.11	36.00
APE-T	**66.07**	**57.59**	**36.36**

Table 2: **Domain Generalization Performance (%) of APE and APE-T.** We utilize 16-shot ImageNet [3] as the training data before out-of-distribution test.

significant out-of-distribution performance on ImageNet-V2. For ImageNet-Sketch with more distribution shifts, our training-free APE outperforms all existing methods including the training-required ones. However, as we train the category residuals on the in-domain ImageNet, APE-T performs worse than APE by testing on ImageNet-Sketch.

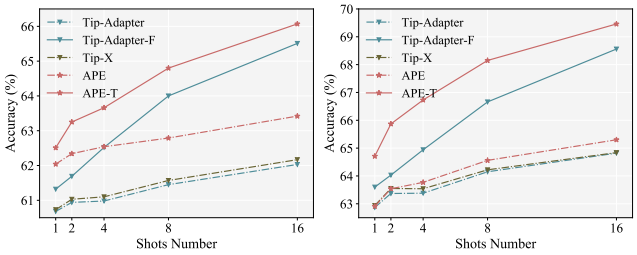
5. Ablation Study

In this section, we perform extensive ablation experiments to investigate the contribution of our method, respectively for the prior refinement module, the training-free APE, and training-required APE-T. All experiments are conducted on ImageNet.

Prior Refinement Module. In Figure 9 (a), we evaluate the impact of our two refinement criteria, inter-class similarity and variance, and adopt our training-free APE with ResNet-50 [11] as the baseline. As shown, the absence of either similarity or variance would harm the performance. In addition, we observe that the similarity criterion plays a more important role than variance, which better selects the most discriminative channels from CLIP-extracted representations. Then in Figure 9 (b), we investigate the influence of refined channel number Q. For all shots, the channel number within the range $[500, 900]$ yields better performance. This indicates the more significance of our refined feature channels than other redundant ones.

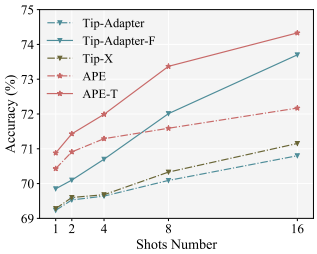
Training-free APE. In Figure 10 (a), we decompose the proposed trilateral relations and reveal their roles respectively. For the 0-shot result, 'Only $\mathbf{R}_{fW}$' denotes the performance of zero-shot CLIP with 61.64% accuracy. By equipping '$\mathbf{R}_{fW}$ + $\mathbf{R}_{f'F'}$', the cache model with prior refinement can help to attain higher performance under the few-shot settings. Finally, considering all three relations ('APE') builds the best-performing framework, which demonstrates the effective boost from our trilateral analysis.

Training-required APE-T. In Figure 10 (b), we compare the impact of different learnable modules in APE-T, including the category residuals $\mathbf{Res}$ for the visual $\mathbf{F}'$ and the textual $\mathbf{W}$, and the cache scores, $\mathbf{R}_{F'W'}$. From the presented results, each learnable component is necessary to best unleash the potential of APE-T. We observe that tuning the refined feature channels in $\mathbf{W}$ is more significant than $\mathbf{F}'$. This suggests the role of textual zero-shot prediction is more critical than the cache model, since CLIP's original pre-training target lies in the vision-language contrast.

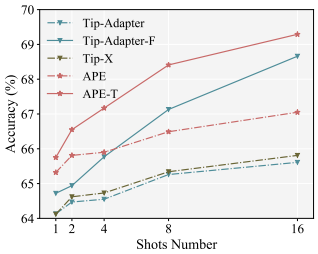


(a) Results with ResNet-50

(b) Results with ResNet-101



(c) Results with ViT-B/16



(d) Results with ViT-B/32

Figure 11: **Ablation Study with Different Backbones.** The dashed and solid lines represent training-free and training-required methods, respectively. Totally four network structures are involved: ResNet-50 [11], ResNet-101 [11], ViT-B/16 [5], and ViT-B/32 [5].

Different Backbones. We implement our approach and existing models under different CLIP encoders in Figure 11. We utilize the best settings and only substitute the encoder network. The ResNet [11] and Vision Transformer (ViT) [5] backbones are investigated, with which we still achieve the best accuracy, under training or training-free settings.

Different Prompt. We consider the influence of prompt in Figure 12. Three types of prompts are involved. The template prompt is the widely utilized version, *e.g.*, ensembling 7 different templates for ImageNet, following [40]. The CuPL prompt proposed in [25] is generated by GPT-3. We ensemble template and CuPL prompt in our work, denoted as "CuPL+t". From Figure 12, CuPL+t prompt can advance all few-shot approaches. Besides, our APE and APE-T perform the best under all sorts of prompts.

Balance Factor λ and Smooth Factor γ. We explore the effect of the values of λ in Equation 8 and γ in Equation 11. Balance factor λ controls the weights of the similarity and variance criteria to the final blending criterion J_k. We observe for APE, $\lambda = 0.7$ yields the best accuracy. This suggests that the similarity criterion is more important for APE. The smooth factor γ controls the contribution of each training sample to the final prediction. We observe a significant accuracy improvement when γ increases from 0 to 0.1, which indicates the efficacy of relation $\mathbf{R}_{F'W'}$. When γ increases from 0.2 to 0.3, *i.e.*, relation $\mathbf{R}_{F'W'}$ becomes sharper, the performance reduces rapidly, which suggests the few-shot performance is sensitive to hyper-parameter γ. The positive value of γ indicates that Equation 11 is closer to finding difficult samples.

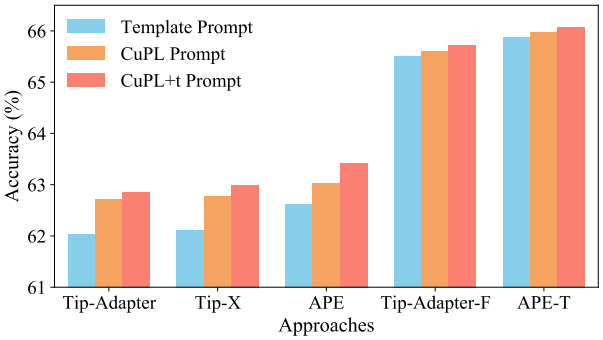


Figure 12: **Different Prompt for APE and APE-T.**

Balance Factor λ	0.1	0.3	0.5	0.6	**0.7**	0.8
	63.02	63.15	63.27	63.33	**63.42**	63.37
Smoothing Factor γ	0.0	0.05	0.1	0.15	**0.2**	0.3
	62.64	63.04	63.31	63.34	**63.42**	63.06

Table 3: **Ablation Studies (%) for Hyper-parameters** of APE on ImageNet [3]. We investigate blending balance factor λ in Equation 8, and smooth factor γ in Equation 11. The experiments are conducted under 16-shot settings with ResNet-50 [11] backbone.

6. Conclusion

In this paper, we propose an **A**daptive **P**rior r**E**finement method (APE) to adapt CLIP for downstream datasets. Our APE extracts the informative domain-specific feature channels with two criteria and digs into trilateral relations between three CLIP-extracted representations. On top of this, we present two model variants of APE, respectively for training-free and training-required few-shot learning. Extensive experiments have demonstrated our approach can not only achieve leading few-shot results but also obtain superior efficiency. Our future direction will focus on extending APE for wider CLIP-based downstream tasks besides classification, *e.g.*, open-world object detection, segmentation, and 3D point cloud recognition.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (Grant No.62206272), and by the National Key R&D Program of China (NO.2022ZD0160100).

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461, 2014. 6

[2] 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*, pages 3606–3613, 2014. 6

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2, 4, 6, 8, 9

[4] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018. 4

[5] 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. 9

[6] 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 *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004. 6

[7] 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. 2, 3, 6, 7, 8

[8] Samantha Guerriero, Barbara Caputo, and Thomas Mensink. Deepncm: Deep nearest class mean classifiers. 2018. 4

[9] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *IJCAI 2023*, 2023. 1

[10] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022. 8

[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*, pages 770–778, 2016. 2, 4, 6, 8, 9

[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. 6

[13] Saumya Jetley, Bernardino Romera-Paredes, Sadeep Jayasumana, and Philip Torr. Prototypical priors: From improving classification to zero-shot learning. *arXiv preprint arXiv:1512.01192*, 2015. 4

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[15] 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*, pages 554–561, 2013. 6

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[17] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022. 1

[18] Guoliang Lin and Hanjiang Lai. Revisiting few-shot learning from a causal perspective. *arXiv preprint arXiv:2209.13816*, 2022. 3, 5

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[22] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 1

[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*, pages 722–729, 2008. 6

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. 6

[25] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022. 6, 9

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 3, 8

[27] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 1

[28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to im-

agenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 6, 8

[29] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022. 1

[30] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022. 3

[31] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[32] 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. 6

[33] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. *arXiv preprint arXiv:2211.16198*, 2022. 2, 3, 4, 5, 6, 7, 8

[34] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 6

[35] Wenguan Wang, Cheng Han, Tianfei Zhou, and Dongfang Liu. Visual recognition with deep nearest centroids. *arXiv preprint arXiv:2209.07383*, 2022. 4

[36] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022. 1

[37] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 6

[38] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 1

[39] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021. 1

[40] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2, 3, 5, 6, 7, 8, 9

[41] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 1

[42] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 1

[43] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *CVPR 2023*, 2023. 2, 3

[44] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 1

[45] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 2, 3, 6, 7, 8

[48] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 1