Dual Memory Networks:
A Versatile Adaptation Approach for Vision-Language Models

Yabin Zhang[1,2] Wenjie Zhu[1] Hui Tang[3] Zhiyuan Ma[1] Kaiyang Zhou[4] Lei Zhang[1,2,*]
[1]HKPolyU [2]OPPO [3]HKUST [4]HKBU
{csybzhang,cslzhang}@comp.polyu.edu.hk

Abstract

With the emergence of pre-trained vision-language models like CLIP, how to adapt them to various downstream classification tasks has garnered significant attention in recent research. The adaptation strategies can be typically categorized into three paradigms: zero-shot adaptation, few-shot adaptation, and the recently-proposed training-free few-shot adaptation. Most existing approaches are tailored for a specific setting and can only cater to one or two of these paradigms. In this paper, we introduce a versatile adaptation approach that can effectively work under all three settings. Specifically, we propose the dual memory networks that comprise dynamic and static memory components. The static memory caches training data knowledge, enabling training-free few-shot adaptation, while the dynamic memory preserves historical test features online during the testing process, allowing for the exploration of additional data insights beyond the training set. This novel capability enhances model performance in the few-shot setting and enables model usability in the absence of training data. The two memory networks employ the same flexible memory interactive strategy, which can operate in a training-free mode and can be further enhanced by incorporating learnable projection layers. Our approach is tested across 11 datasets under the three task settings. Remarkably, in the zero-shot scenario, it outperforms existing methods by over 3% and even shows superior results against methods utilizing external training data. Additionally, our method exhibits robust performance against natural distribution shifts. Codes are available at https://github.com/YBZh/DMN.

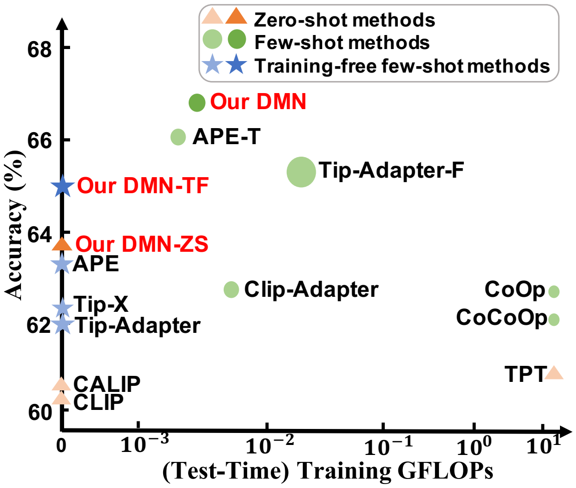


Figure 1. Illustration of the classification accuracy, (test-time) training GFLOPs, and learning parameters on zero-shot and 16-shot ImageNet classification. The icon sizes denote the number of learnable parameters. Our method is unique in its ability to work for all three task settings with superior results.

1. Introduction

Contrastive vision-language pre-training [20, 27, 44, 64] has shown promising results in various downstream vision tasks, including 2D/3D perception [69, 74] and generation [6, 48]. Among these models, CLIP [44] is arguably the most representative one due to its simplicity and effectiveness. Leveraging a vast collection of image-text pairs from the Internet, CLIP aligns features across modalities, leading to notable zero-shot classification capabilities. To further enhance its performance on downstream tasks, numerous adaptation strategies have emerged, primarily employing frozen CLIP encoders in zero-shot and few-shot settings.

Most existing approaches are tailored for one specific task setting. Specifically, enhanced zero-shot performance is achieved by exploring additional insights from the test sample itself [14, 51] or via enhanced text prompts [38, 43]. In the few-shot setting, researchers typically insert adaptive parameters (*e.g.*, Prompt [22, 74], Adapter [13], and Residual [65]) into the pre-trained vision-language models and optimize these parameters using labeled training data. Recently, a training-free variant of few-shot adaptation has been proposed for resource-constrained applications [68].

---

[*] Corresponding author.

Methods	No External Training Data	Task Settings		
		Zero-shot	Few-shot	TF Few-shot
TPT [51] and [38, 40, 43, 46, 75]	✓	✓	✗	✗
DiffTPT [12]	✗	✓	✗	✗
CoOp [74] and [3, 22, 36, 50, 59, 63, 65, 66, 73, 76]	✓	✗	✓	✗
Tip-Adapter [68], and [77]	✓	✗	✓	✓
SuS-X [56]	✗	✗	✓	✓
CALIP [14]	✓	✓	✓	✗
CaFo [70]	✗	✓	✓	✓
DMN (Ours)	✓	✓	✓	✓

Table 1. Summary of adaptation methods for vision-language models. 'Zero-shot', 'Few-shot', and 'TF Few-shot' represent the zero-shot adaptation, few-shot adaptation, and the recently introduced training-free few-shot adaptation, respectively. 'No External Training Data' indicates that the approach does not utilize any synthetic training images from generation models or retrieved images via class names.

In this setting, no parameters are needed to learn, and thus much computational resources are saved. While numerous methods have been introduced, they typically cater to only one or two task settings, as summarized in Tab. 1, thereby limiting their applicability.

In this work, we propose a versatile adaptation approach that works effectively for all the three task settings, as shown in Fig. 1. Specifically, we propose the dual memory networks comprising dynamic and static memory components, producing sample-adaptive classifiers for each test point. The static memory network caches features of training data, generating the adaptive classifier for each test sample by adaptively weighting cached training features and thus enabling training-free few-shot adaptation. In contrast, the dynamic memory network preserves features of historical test samples during the testing process, introducing another adaptive classifier by adaptively weighting cached test features. This allows us to explore additional data insights beyond the training samples, further enhancing the model's performance in the few-shot setting and extending its applications to the zero-shot setting where training data is absent. These two types of memory networks employ the same memory interactive strategy, which is highly flexible. This strategy can be used in a training-free mode for zero-shot and training-free few-shot adaptations. In addition, it can be further enhanced by incorporating learnable projection layers in the traditional few-shot setting.

We evaluate our approach on 11 datasets. In particular, in the setting where external training data are unavailable, our method surpasses existing zero-shot methods by a significant margin of over 3% by leveraging knowledge of historical test samples. Even in comparison to methods that utilize external training data, our model still exhibits substantial advantages, outperforming the recent CaFo [70] by 1.48%. These results highlight the crucial significance of historical test samples in the adaptation process, which is neglected in existing works. It is worth emphasizing the efficiency of incorporating historical test knowledge with the dynamic memory network, as the memory interaction process involves only a single attention module. Through the utilization of historical test knowledge, labeled training data, and vanilla text information, our approach significantly enhances few-shot performance, establishing a new state-of-the-art in both the few-shot and training-free few-shot settings. Moreover, our method demonstrates excellent generalization capabilities to natural distribution shifts. We summarize our contributions as follows:

- We introduce a versatile adaptation strategy for pre-trained vision-language models, termed Dual Memory Networks (DMN), aimming to effectively address the tasks of zero-shot adaptation, few-shot adaptation, and training-free few-shot adaptation. To the best of our knowledge, *this is the first work to enhance vision-language model adaptation across the three settings without the use of external training data*.
- DMN comprises static and dynamic memory networks that gather information from labeled training data and historical test data, respectively. The two memory networks employ a flexible interactive strategy, which can operate in a training-free mode and can be further enhanced with learnable projection layers.
- Our approach has been validated on 11 datasets with three task settings. In the zero-shot setting, it outperforms competitors by over 3% and even surpasses methods using external training data. It also demonstrates robust performance against natural distribution shifts.

2. Related Work

Adaptation of Vision-Language Models. Foundation models [24, 29, 44, 47] have attracted increasing attention in downstream tasks recently [32, 33, 55, 60, 67]. Pre-trained on vast collections of image-text pairs, vision-language models like CLIP exhibit remarkable zero-shot generalization capabilities across a range of downstream datasets [44]. Building upon CLIP, numerous methods have been introduced to adapt it to various downstream classification tasks, especially under the zero-shot and few-shot settings as summarized in Tab. 1. In the zero-shot setting where labeled training data are unavailable, one primary research direction is how to extract richer information from the test samples [12, 14, 51, 75] and class names [12, 38, 40, 43, 56]. For the former group, CALIP [14] enhances feature extraction through attention mechanisms, and instance-adaptive prompts are explored using consistency regularization in [12, 51]. Leveraging class names, some approaches [56, 70] generate synthetic training samples utilizing additional image generation models [7, 47], and others [38, 43, 46] craft advanced text prompts by querying pre-trained large language models.

To further unlock the potential of pre-trained CLIP models for downstream tasks, how to adapt the frozen CLIP model with a limited amount of labeled training data has attracted increasing attention, leading to the few-shot adaptation. Inspired by the parameter-efficient transfer learning [19, 26], many methods propose to tune the pre-trained CLIP models with carefully designed prompts [3, 3, 22, 23, 66, 73, 74] and adapters [13]. Besides, Lin *et al.* [34], Wortsman *et al.* [59], and Yu *et al.* [65] respectively investigate the cross-modal adaptation, weight ensembles, and task residuals for better CLIP adaptation. Recently, a training-free variant of few-shot adaptation has been proposed for resource-constrained applications [68], where computationally intensive model training is prohibited. Specifically, Tip-Adapter [68] is a pioneering training-free few-shot approach, which caches the encoded features and labels of training images as task priors. Predictions are then derived based on the similarity between the test feature and cached features. Tip-Adapter is subsequently augmented with the integration of calibrated intra-modal distance as described in [56], and through adaptive channel prior refinement as elaborated in [77]. These training-free adaptation methods can be enhanced with optional model optimization by either tuning the cached features [68] or adding learnable category residuals [77].

Most aforementioned methods are tailored for a specific task setting and can only cater to one or two of these adaptation paradigms, as summarized in Tab. 1. Although existing few-shot methods can be applied to the zero-shot task by utilizing external training data through generation or searching [56, 70], they may not fully meet the practical requirements of zero-shot applications, such as efficient and rapid adaptation to new tasks. In contrast, we propose a versatile adaptation approach that can effectively handle all the three tasks without relying on any external training data. This is achieved by fully utilizing the training data and historical test samples via the proposed DMN framework, leading to the new state-of-the-art across all three adaptation settings.

Memory Networks. Memory networks were initially introduced in the realm of Natural Language Processing. Inspired by the knowledge accumulation and recalling in human brain [1, 53], they introduce an external memory component, allowing the storage and retrieval of historical knowledge to facilitate decision making [54, 58]. Subsequently, the concept of interactive memory, facilitating the storage and retrieval of historical information, has been adopted in various vision tasks, including classification [21, 49], segmentation [28, 41, 62], and detection [4, 8, 30, 31]. Recently, ideas reminiscent of memory networks have been introduced into CLIP adaptation [56, 68]. However, the memory modules employed in their approaches are typically read-only and do not support real-time writing, akin to the static memory in our method. As expected, these approaches are unable to leverage historical test samples, limiting their performance in few-shot adaptation and impeding their application in zero-shot adaptation. Our method stands out as the first to introduce a dynamic memory that supports both reading and writing operations for test data, while optionally maintaining a static memory for training data. By exploring all available data sources, our method can effectively handle all the three adaptation tasks and achieve superior performance.

3. Method

We first present a flexible memory interactive strategy for both dynamic and static memory networks. Then, we present these memory networks in detail.

3.1. A Flexible Memory Interactive Strategy

Memory networks [54, 58] provide an effective mechanism to explicitly accumulate and recall knowledge, empowering better performance by utilizing the relevant historical information. A memory network typically comprises the following four abstract steps:

1. Convert a new input $\mathbf{x}$ into the feature space.
2. Update the memory $\mathbf{M}$ with $\mathbf{x}$.
3. Read out an output given $\mathbf{x}$ and the current memory.
4. Convert the output into the desired response.

In the following, we demonstrate how to instantiate these steps in CLIP adaptation, where the memory interaction strategy in steps 2 and 3 is our main focus.

We first present how to use CLIP to classify a test sample under the zero-shot setting. For a test image $\mathbf{x}$ within a downstream task of C classes, we extract the visual representation $\mathbf{v} \in \mathbb{R}^D$ and textual representation $\mathbf{C} \in \mathbb{R}^{C\times D}$ with pre-trained CLIP encoders, where D is the feature dimension. Both $\mathbf{v}$ and $\mathbf{C}$ are L_2 normalized along the D dimension. Then, the zero-shot prediction probability can

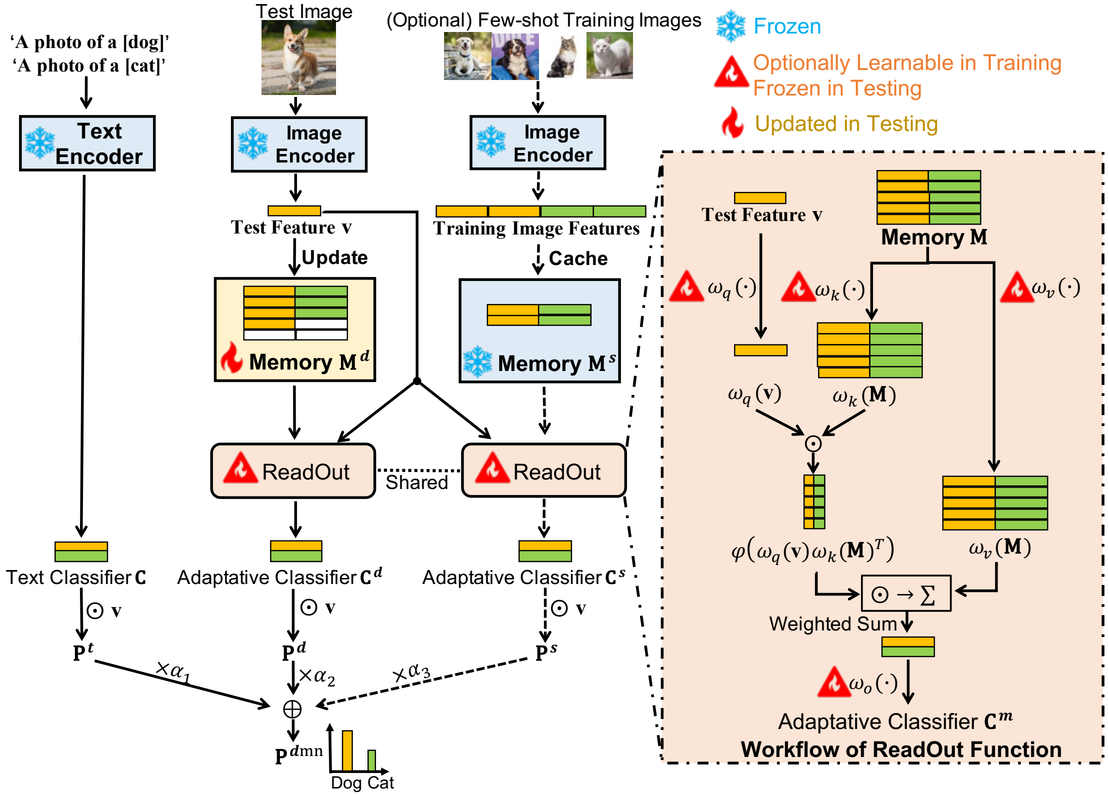


Figure 2. An illustration of the overall framework of our Dual Memory Networks (DMN), which integrates knowledge from three sources (*i.e.*, text input, historical test data, and optional training images) to tackle the three types of adaptation tasks (*i.e.*, zero-shot, few-shot, and the recently-proposed training-free few-shot adaptations).

be achieved by using text features $\mathbf{C}$ as the classifier:

$$\mathbf{P}^t = \text{Softmax}\left(\mathbf{v}\mathbf{C}^\top\right) \in \mathbb{R}^C, \tag{1}$$

where the scaling parameter is omitted for simplicity.

To instantiate the memory networks in CLIP adaption, it is natural to adopt the pretrained image encoders of CLIP to transform the input $\mathbf{x}$ to the image feature $\mathbf{v}$. We construct a category-split memory $\mathbf{M} \in \mathbb{R}^{C\times L\times D}$, where L is the memory length for each category. To update the memory $\mathbf{M}$ with $\mathbf{v}$, we simply store $\mathbf{v}$ in a 'slot' of $\mathbf{M}$. Specifically, given the (pseudo) label $y \in [1, C]$ of the input image, we locate the sub-memory $\mathbf{M}_y \in \mathbb{R}^{L\times D}$ corresponding to the category y, find an empty slot of it, say the i^{th} row, denoted by $\mathbf{M}_{y,i} \in \mathbb{R}^D$, and update the memory as:

$$\mathbf{M}_{y,i} = \mathbf{v}. \tag{2}$$

Besides the image feature, we also cache the corresponding prediction entropy estimated from $\mathbf{P}^t$, which is used to locate the slot to update when $\mathbf{M}_y$ is full. Specifically, if all rows of $\mathbf{M}_y$ are occupied by image features, we replace the row of maximum entropy in $\mathbf{M}_y$ with $\mathbf{v}$ if $\mathbf{v}$ exhibits smaller prediction entropy. In other words, we store samples with lower prediction entropy in the memory.

Given the updated memory $\mathbf{M}$ and the test feature $\mathbf{v}$, we read out a sample adaptive classifier $\mathbf{C}^m \in \mathbb{R}^{C\times D}$ via cross-attention as:

$$\mathbf{C}^m = \text{ReadOut}(\mathbf{v}, \mathbf{M}), \tag{3}$$

where the y^{th} row of $\mathbf{C}^m$ is produced by using $\mathbf{v}$ as query and adopting memory $\mathbf{M}_y$ as key and value:

$$\mathbf{C}^m_y = \omega_o\left(\varphi\left(\omega_q(\mathbf{v})\omega_k(\mathbf{M}_y)^\top\right)\omega_v(\mathbf{M}_y)\right). \tag{4}$$

The ω_q, ω_k, ω_v and ω_o respectively represent the project functions for query, key, value, and the output, $\omega_q(\mathbf{v})\omega_k(\mathbf{M}_y)^\top \in \mathbb{R}^{1\times L}$ measures the cosine similarities between normalized features of $\omega_q(\mathbf{v})$ and $\omega_k(\mathbf{M}_y)$, and $\varphi(x) = \exp(-\beta(1-x))$ modulates the sharpness of x with hyper-parameter β. Intuitively, $\mathbf{C}^m_y$ is the weighted combination of image features in $\mathbf{M}_y$, where the weight is based on the cosine similarity between test and memoried image features. In other words, the sample adaptive classifier $\mathbf{C}^m$ is produced by image features, instead of the text features that produce the text classifier $\mathbf{C}$.

Finally, we follow Eq. (1) to convert the memory output $\mathbf{C}^m$ to the desired classification prediction, leading to the final memory response:

$$\mathbf{P}^m = \mathrm{M2P}(\mathbf{v}, \mathbf{C}^m) = \mathrm{Softmax}\left(\mathbf{v}\mathbf{C}^{m\top}\right) \in \mathbb{R}^C. \quad (5)$$

The $\mathbf{P}^m$ is the classification probability of test feature $\mathbf{v}$ with the sample adaptive classifier $\mathbf{C}^m$.

The versatility of our memory interactive strategy across various task settings stems from the flexibility of the projection layer. Specifically, we define the projection function ω_* (covering $\omega_q, \omega_k, \omega_v$, and ω_o) using a residual architecture:

$$\omega_*(x) = \mathrm{L}_2\left(x + \mathrm{Linear}(x)\right), \quad (6)$$

where $\mathrm{Linear}(\cdot)$ represents a linear layer with all parameters initialized to zero and $\mathrm{L}_2(\cdot)$ indicates the L_2 normalization along feature dimension. In the training-free setting, the projection function $\omega_*(\cdot)$ degenerates to $\omega_*(x) = x$, given the L_2 normalized input x. Therefore, the memory interaction is conducted in the vanilla feature space of CLIP. Given labeled training samples, we can explore a more efficient feature space for memory interaction by optimizing the linear layers with the classification objective. Next, we present the dynamic and static memory networks based on this flexible interactive strategy.

3.2. Dynamic Memory Network

The dynamic memory networks accumulate historical test samples in the test process and is activated for all task settings. Firstly, we introduce a dynamic memory $\mathbf{M}^d \in \mathbb{R}^{C\times L\times D}$ initialized with zero values. Given the test feature $\mathbf{v}$, we update the memory $\mathbf{M}^d$ using Eq. (2) with the estimated pseudo label y from the text classifier:

$$y = \arg\max_j \mathbf{P}^t_j. \quad (7)$$

Given the updated memory $\mathbf{M}^d$ and the test feature $\mathbf{v}$, we can read out a sample adaptive classifier $\mathbf{C}^d$ with the readout function in Eq. (3) as:

$$\mathbf{C}^d = \mathrm{ReadOut}(\mathbf{v}, \widehat{\mathbf{M}}^d), \quad (8)$$

where $\widehat{\mathbf{M}}^d = [\mathbf{M}^d, \mathbf{C}] \in \mathbb{R}^{C\times(L+1)\times D}$ is the extended memory with text feature. Such a memory extension actually initializes the $\mathbf{C}^d$ with the text classifier $\mathbf{C}$, considering that the memory $\mathbf{M}^d$ is initialized with zero values. As more image features are written into the memory, the classifier $\mathbf{C}^d$ is gradually refined with cached image features, utilizing the historical test samples in the testing process. Finally, the sample classification probability with the dynamic memory network is introduced with Eq. (5) as:

$$\mathbf{P}^d = \mathrm{M2P}(\mathbf{v}, \mathbf{C}^d) \in \mathbb{R}^C. \quad (9)$$

The prediction $\mathbf{P}^d$ utilizes knowledge of historical test samples, including the current one, whose effectiveness is analyzed in Sec. 4.3.

Variants	Adaptation Settings	$\mathbf{M}^d$	$\mathbf{M}^s$	ω_*
DMN-ZS	Zero-shot	✓	✗	✗
DMN-TF	Training-free Few-shot	✓	✓	✗
DMN	Few-shot	✓	✓	✓

Table 2. Summary of our DMN variants for different adaptation tasks. The '$\mathbf{M}^d$' and '$\mathbf{M}^s$' respectively represent whether the dynamic and the static memory networks are activated and 'ω_*' indicates whether the projection layers are optimized.

3.3. Dual Memory Networks

In this section, we present the full version of our versatile DMN, which comprises the aforementioned dynamic memory network and the following static memory network. The overall framework is shown in Fig. 2. For a C-way-K-shot task with K training images per category, one may opt to utilize these samples by extending the dynamic memories with image features of these data, *i.e.*, updating $\widehat{\mathbf{M}}^d = [\mathbf{M}^d, \mathbf{M}^s, \mathbf{C}] \in \mathbb{R}^{C\times(L+K+1)\times D}$ in Eq. (8), where $\mathbf{M}^s \in \mathbb{R}^{C\times K\times D}$ is the aggregation of image features of CK training samples. Although this simple strategy brings certain improvement, we argue that the valuable knowledge from labeled data may gradually get diluted as the dynamic memory fills up. This dilution results in a degraded performance (see Fig. 6a for more analyses).

To make full use of labeled data, we additionally maintain one static memory, *i.e.*, $\mathbf{M}^s$, and introduce another sample adaptive classifier using these labeled data only. As described by its name, the static memory $\mathbf{M}^s$ keeps unchanged after creation. Given the static memory $\mathbf{M}^s$ and the test feature $\mathbf{v}$, we can read out a sample adaptive classifier $\mathbf{C}^s$ with the readout function in Eq. (3) as:

$$\mathbf{C}^s = \mathrm{ReadOut}(\mathbf{v}, \mathbf{M}^s). \quad (10)$$

The corresponding prediction probability is:

$$\mathbf{P}^s = \mathrm{M2P}(\mathbf{v}, \mathbf{C}^s) \in \mathbb{R}^C. \quad (11)$$

The prediction $\mathbf{P}^s$ is based on the knowledge of labeled training data, which are complement to the text knowledge in $\mathbf{P}^t$ and historical test knowledge in $\mathbf{P}^d$. The final prediction is obtained by aggregating the three knowledge sources:

$$\mathbf{P}^{dmn} = \alpha_1\mathbf{P}^t + \alpha_2\mathbf{P}^d + \alpha_3\mathbf{P}^s, \quad (12)$$

where $\alpha_{1\sim3}$ denote the weights for text prediction, prediction of dynamic memory network, and prediction of static memory network, respectively.

Our DMN is a versatile adaptation approach for vision-language models that handles three task settings, *i.e.*, zero-shot, few-shot, and training-free few-shot adaptations. Considering the inherent variations among different task settings, the implementation of our DMN exhibits subtle differences. For example, in the training-free setting, such

Method	ImageNet	Flower	DTD	Pets	Cars	UCF	Caltech	Food	SUN	Aircraft	EuroSAT	Mean
CLIP-RN50 [44]	58.16	61.75	40.37	83.57	55.70	58.84	85.88	73.97	58.80	15.66	23.69	56.04
DN [75]	60.16	63.32	41.21	81.92	56.55	55.60	87.25	74.64	59.11	17.43	28.31	56.86
TPT [51]	60.74	62.69	40.84	84.49	58.46	60.82	87.02	74.88	61.46	17.58	28.33	57.94
VisDesc [38]	59.68	65.37	41.96	82.39	54.76	58.47	88.11	76.80	59.84	16.26	37.60	58.29
Ensemble [68]	60.32	66.10	40.07	85.83	55.71	61.33	83.94	77.32	58.53	17.10	37.54	58.53
CALIP [14]	60.57	66.38	42.39	86.21	56.27	61.72	87.71	77.42	58.59	17.76	38.90	59.45
DiffTPT [12]*	60.80	63.53	40.72	83.40	60.71	62.67	86.89	**79.21**	62.72	17.60	41.04	59.94
CuPL [43]	61.45	65.44	48.64	84.84	57.28	58.97	89.29	76.94	62.55	19.59	38.38	60.31
SuS-X-SD-C [56]*	61.84	67.72	50.59	85.34	57.27	61.54	89.53	77.58	62.95	19.47	45.57	61.76
CaFo [70]*	62.74	66.54	50.24	**87.49**	58.45	63.67	**90.91**	77.53	63.16	21.06	42.73	62.23
DMN-ZS (Ours)	**63.87**	**67.93**	**50.41**	86.78	**60.02**	**65.34**	90.14	76.70	**64.39**	**22.77**	**48.72**	**63.71**
CLIP-ViTB/16 [44]	66.73	67.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	63.87
Ensemble [68]	68.34	66.99	45.04	86.92	66.11	65.16	93.55	82.86	65.63	23.22	50.42	64.93
TPT [51]	68.98	68.98	47.75	87.79	66.87	68.04	94.16	84.67	65.50	24.78	42.44	65.45
DiffTPT [12]*	70.30	70.10	47.00	88.20	67.01	68.22	92.49	**87.23**	65.74	25.60	43.13	65.91
DMN-ZS (Ours)	**72.25**	**74.49**	**55.85**	**92.04**	**67.96**	**72.51**	**95.38**	85.08	**70.18**	**30.03**	**59.43**	**70.72**

Table 3. Zero-shot classification performance on eleven downstream datasets, where results with * are achieved with external training data.

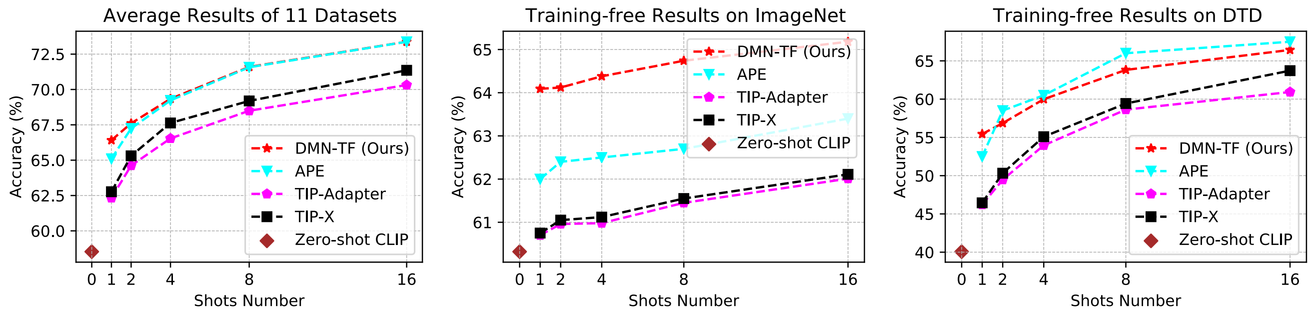


Figure 3. Training-free few-shot results with a ResNet50 backbone. Full results on 11 classification datasets are presented in Fig. A7.

as zero-shot and the training-free few-shot adaptations, we adopt the initialized projection layers in Eq. (6) and conduct memory interaction in the vanilla CLIP feature space, while we finetune these projection layers and explore more efficient feature space for the traditional few-shot setting. To distinguish our results under different task settings, we term the DMN variants with respect to zero-shot, few-shot, and training-free few-shot settings as DMN-ZS, DMN, and DMN-TF, respectively. We summarize these variants in Tab. 2.

4. Experiments

4.1. Experiment Settings

Datasets. We validate our method on 11 classification benchmarks, including ImageNet [9], Flowers102 [39], DTD [5], OxfordPets [42], StandfordCars [25], UCF101 [52], Caltech101 [11], Food101 [2], SUN397 [61], FGVCAircraft [37], and EuroSAT [16]. We also evaluate the robustness of DMN to natural distribution shifts [71, 72] on four ImageNet variants, *i.e.*, ImageNet-V2 [45], ImageNet-A [18], ImageNet-R [17], and ImageNet-Sketch [57].

Settings. We adopt visual encoders of ResNet50 [15] and VIT-B/16 [10] pretrained by CLIP. We follow existing works to conduct the image split in few-shot learning and adopt the textual prompt in [43, 68]. Inspired by [51], we enhance the robust pseudo label estimation in Eq. 7 with view augmentation and confidence selection. We search the optimal prediction weights, *i.e.*, $\alpha_{1\sim3}$, for each downstream task, while illustrate that the fixed weights generalize well within each task setting. We train the DMN with AdamW optimizer [35], where we adopt the cosine annealing learning schedule with the initial learning rate of 1e-4 and set the batch size as 128. We train the model for 20 epochs for most datasets except for the Flower102 and EuroSAT, where 100 epochs are adopted.

4.2. Performance Evaluation

Zero-shot DMN-ZS Results. We first present the experimental results under the zero-shot adaptation setting, where the significance of historical test knowledge becomes particularly pronounced. As illustrated in Tab. 3, our method surpasses its closest competitors that do not involve external training data, such as CALIP and TPT. Specifically, we observe improvements of 3.40% and 5.27% when employing ResNet-50 and ViTB/16 backbones, respectively.

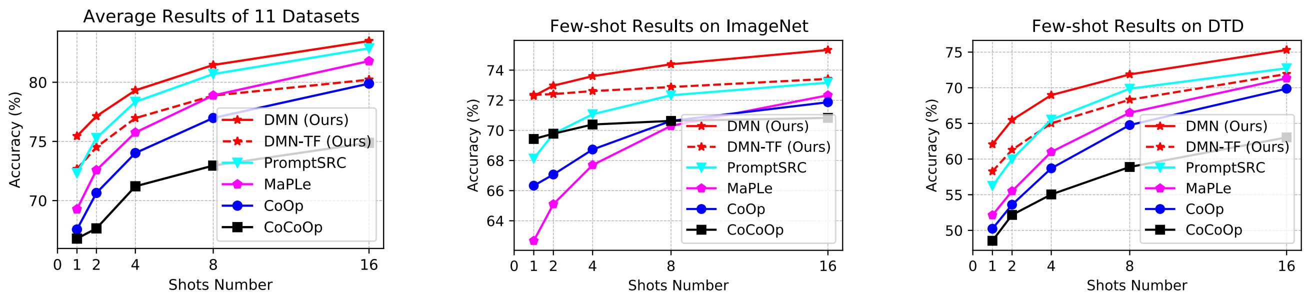


Figure 4. Few-shot performance with ViTB/16 backbone, where the full results on 11 classification datasets are presented in Fig. A8.

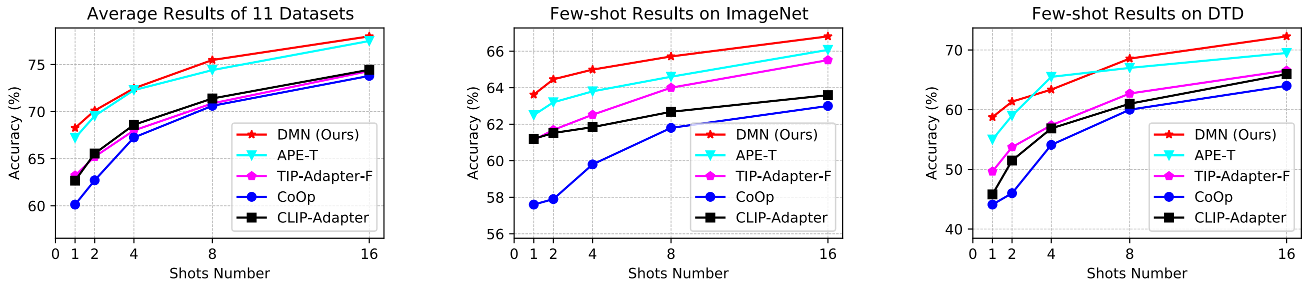


Figure 5. Few-shot performance with ResNet50 backbone, where the full results on 11 classification datasets are presented in Fig. A9.

Method	ImageNet	-A	-V2	-R	-Sketch
CLIP-RN50 [44]	58.16	21.83	51.41	56.15	33.37
Ensemble	59.81	23.24	52.91	60.72	35.48
TPT [51]	60.74	26.67	54.70	59.11	35.09
CALIP [14]	60.57	23.96	53.70	60.81	35.61
DiffTPT [12]	60.80	**31.06**	55.80	58.80	37.10
CoCoOp* [73]	62.81	23.32	55.72	57.74	34.48
CoOp* [74]	63.33	23.06	55.40	56.60	34.67
DMN-ZS (Ours)	**63.87**	28.57	**56.12**	**61.44**	**39.84**
CLIP-ViT-B/16 [44]	66.73	47.87	60.86	73.98	46.09
Ensemble	68.34	49.89	61.88	77.65	48.24
TPT [51]	68.98	54.77	63.45	77.06	47.94
DiffTPT [12]	70.30	55.68	65.10	75.00	46.80
MaPLe* [22]	70.72	50.90	64.07	76.98	49.15
CoCoOp* [73]	71.02	50.63	64.07	76.18	48.75
CoOp* [74]	71.51	49.71	64.20	75.21	47.99
PromptSRC* [23]	71.27	50.90	64.35	77.80	49.55
DMN-ZS (Ours)	**72.25**	**58.28**	**65.17**	**78.55**	**53.20**

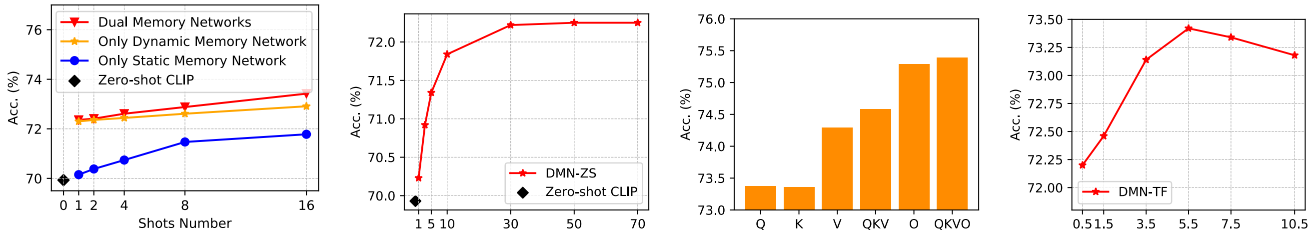
Table 4. Robustness to Natural Distribution Shifts. Results with * are tuned on ImageNet using 16-shot training samples per category, while other methods do not require labeled training data.

Compared to approaches like TPT [51], which necessitate model optimization on test samples, the memory interactions within our DMN do not introduce any test time optimization, substantially accelerating the inference speed, as shown in Tab. 5.

To tackle the zero-shot challenge, some approaches utilize labeled synthetic training samples generated from pre-trained image generation models [56, 70]. By treating these synthetic labeled data like genuine labeled data, the zero-shot problem can be tackled through few-shot approaches. While these strategies offer notable performance gains, the generation of synthetic data and subsequent model optimization come with considerable computational overheads, failing to meet the efficient adaptation requirement in zero-shot setting. In contrast, incorporating historical test knowledge with our dynamic memory network is considerably faster. Interestingly, even when compared to techniques that employ synthetic training data, our approach maintains a distinct advantage, highlighting the superiority of historical test samples over synthetic training data.

Training-free Few-shot DMN-TF Results. We compare our DMN-TF with the training-free few-shot methods of Tip-Adapter [68], Tip-X [56], and the recent APE [77]. As illustrated in Fig. 3, our method achieves a superior advantage with one training sample per category. The advantage gradually diminishes with additional training samples.

Few-shot DMN Results. We compare our method with seven few-shot adaptation methods of CoOp [74], CoCoOp [73], MaPLe [22], PromptSRC [23], CLIP-Adapter [13], Tip-Adapter-F [68], and APE-T [77]. All methods employed for comparison do not utilize external training data. As evidenced by the results averaged over eleven datasets shown in Fig. 4 and Fig. 5, our DMN consistently surpasses competing approaches, maintaining superiority with different backbone architectures and varying numbers of training samples. On individual datasets, although our method occasionally lags behind some competing methods in certain settings, it achieves consistent gains on the acknowledged ImageNet dataset, affirming its effectiveness.



(a) Two memory networks (b) Memory length L (c) Position of projection layers (d) Values of β

Figure 6. Analyses on (a) static and dynamic memory networks, (b) memory length of the dynamic memory, (b) position of projection layers, and (d) values of β in Eq. (4).

Generalization to Natural Distribution Shifts. As illustrated in Tab. 4, our method not only achieves superior performance on traditional ImageNet dataset, but also generalizes well to the samples with natural distribution shifts, validating its robustness.

4.3. Ablation and Analyses

Dynamic Memory Network vs. Static Memory Network. To analyze the roles of dynamic and static memory networks individually, we introduce two degenerated versions of DMN with dynamic or static memory network only. We illustrate the results under the training-free few-shot setting in Fig. 6a. Both dynamic and static memory networks significantly outperform zero-shot CLIP, and the larger the training sample size, the greater the improvement. Results with dynamic memory network surpass those with static memory network, confirming the importance of historical test samples. The optimal results are achieved by combining the advantages of both memory networks, validating their complementarity.

Memory Length. As shown in Fig. 6b, the classification accuracy gradually increases as the memory length increases and saturates when the memory length exceeds 30. In all experiments, we set the memory length to 50.

Position of Projection Layers. We report the results with different projection layers in Fig. 6c, where Q, K, V and O represent the ω_q, ω_k, ω_v and ω_o, respectively. We observe that all these projection layers bring improvement and the output projection, *i.e.*, ω_o, contributes the most to the results. We adopt the QKVO strategy in all experiments.

Values of β. Results with different values of β are illustrated in Fig. 6d. We set β=5.5 in all experiments.

Computation Efficiency. As summarized in Tab. 5, in zero-shot and training-free few-shot settings, our approach does not introduce any learning parameter, maintaining fast inference speed. In classical few-shot learning, our method achieves fast adaptation by introducing a small amount of training computation and learnable parameters.

Due to the limit of space, more analyses on classifier weights, non-linear function $\varphi(\cdot)$, and test data order can be found in the Supplementary Material.

Methods	**Train**	**Test**	**GFLOPs**	**Param.**
Zero-shot				
CLIP [44]	–	10.1ms	0	0
CALIP [14]	–	10.2ms	0	0
TPT [51]	–	436ms	>10	0.01M
DMN-ZS (Ours)	–	10.7ms	0	0
Few-shot				
Tip-Adapter [68]	–	10.4ms	0	0
APE [77]	–	10.4ms	0	0
DMN-TF (Ours)	–	10.7ms	0	0
CoOp [74]	14 h	10.2ms	>10	0.01M
CLIP-Adapter [13]	50 min	10.4ms	0.004	0.52M
Tip-Adapter-F [68]	5 min	10.4ms	0.030	16.3M
APE-T [77]	5 min	10.4ms	0.002	0.51M
DMN (Ours)	5 min	10.7ms	0.033	4.20M

Table 5. Analyses of computation efficiency on zero-shot and 16-shot ImageNet with a ResNet50 backbone. ‘Training’ measures the training time, ‘GFLOPs’ are calculated during training or test-time training with gradient back-propagation, and ‘Param.’ presents the number of learnable parameters. Results are achieved with a NVIDIA RTX A6000 GPU.

5. Conclusion

In this paper, we proposed a versatile adaptation approach, named Dual Memory Networks (DMN), for vision-language models. By leveraging historical test data and few-shot training samples with dynamic and static memory networks, our DMN can handle all the three commonly used task settings: zero-shot, few-shot, and training-free few-shot adaptations, outperforming existing methods designed for single-task scenarios. Notably, the integration of the dynamic memory network, which utilizes historical test knowledge, distinguished our approach from previous research that overlooked this knowledge source. Nonetheless, our approach had some limitations due to the introduction of two external memories. For instance, in the case of 16-shot ImageNet adaptation, the dynamic and static memories occupied storage space of 204.8MB and 65.5MB, respectively. This may pose challenges for its applications to storage-constrained scenarios.

References

[1] Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000. 3

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

[3] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 2, 3

[4] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10337–10346, 2020. 3

[5] 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

[6] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 1

[7] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall· e mini. *HuggingFace. com. https://huggingface. co/spaces/dallemini/dalle-mini (accessed Sep. 29, 2022)*, 2021. 3

[8] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6678–6687, 2019. 3

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

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[12] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 2, 3, 6, 7, 1

[13] 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. *International Journal of Computer Vision*, pages 1–15, 2023. 1, 3, 7, 8

[14] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 746–754, 2023. 1, 2, 3, 6, 7, 8

[15] 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. 6

[16] 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

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 6

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 6

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[20] 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

[21] Geethan Karunaratne, Manuel Schmuck, Manuel Le Gallo, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Robust high-dimensional memory-augmented neural networks. *Nature communications*, 12(1):2468, 2021. 3

[22] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1, 2, 3, 7

[23] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15190–15200, 2023. 3, 7

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[25] 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

[26] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[28] Minghan Li, Shuai Li, Wangmeng Xiang, and Lei Zhang. Mdqe: Mining discriminative query embeddings to segment occluded instances on challenging videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10524–10533, 2023. 3

[29] Minghan Li, Shuai Li, Xindong Zhang, and Lei Zhang. Univs: Unified and universal video segmentation with prompts as queries. *arXiv preprint arXiv:2402.18115*, 2024. 2

[30] Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. A dual weighting label assignment scheme for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9387–9396, 2022. 3

[31] Shuai Li, Minghan Li, Ruihuang Li, Chenhang He, and Lei Zhang. One-to-few label assignment for end-to-end dense detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7350–7359, 2023. 3

[32] Shuai Li, Minghan Li, Pengfei Wang, and Lei Zhang. Opensd: Unified open-vocabulary segmentation and detection. *arXiv preprint arXiv:2312.06703*, 2023. 2

[33] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. *arXiv preprint arXiv:2311.15707*, 2023. 2

[34] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023. 3

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

[36] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 2

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

[38] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 1, 2, 3, 6

[39] 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. IEEE, 2008. 6

[40] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 2, 3

[41] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 3

[42] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[43] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 1, 2, 3, 6

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7, 8

[45] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 6

[46] Zhiyuan Ren, Yiyang Su, and xiaoming Liu. Chatgpt-powered hierarchical comparisons for image classification. *Advances in neural information processing systems*, 2023. 2, 3

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[48] Aditya Sanghi, Rao Fu, Vivian Liu, Karl DD Willis, Hooman Shayani, Amir H Khasahmadi, Srinath Sridhar, and Daniel Ritchie. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18339–18348, 2023. 1

[49] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 3

[50] Cheng Shi and Sibei Yang. Logoprompt:synthetic text images can be good visual prompts for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[51] 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. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 1, 2, 3, 6, 7, 8

[52] 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

[53] Mark G Stokes. 'activity-silent'working memory in prefrontal cortex: a dynamic coding framework. *Trends in cognitive sciences*, 19(7):394–405, 2015. 3

[54] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015. 3

[55] Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability of diffusion models for content consistent super-resolution. *arXiv preprint arXiv:2401.00877*, 2023. 2

[56] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023. 2, 3, 6, 7

[57] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[58] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 3

[59] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 2, 3

[60] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2

[61] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[62] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7293–7302, 2021. 3

[63] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, Peng Wang, and Yanning Zhang. Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia*, 2023. 2

[64] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 1

[65] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 1, 2, 3

[66] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 2, 3

[67] Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. *arXiv preprint arXiv:2312.03502*, 2023. 2

[68] 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. 1, 2, 3, 6, 7, 8

[69] 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

[70] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 2, 3, 6, 7

[71] Yabin Zhang, Bin Deng, Hui Tang, Lei Zhang, and Kui Jia. Unsupervised multi-class domain adaptation: Theory, algorithms, and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2775–2792, 2020. 6

[72] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8035–8045, 2022. 6

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 3, 7

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 7, 8

[75] Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser-Nam Lim. Distribution normalization: An" effortless" test-time augmentation for contrastively learned visual-language models. *arXiv preprint arXiv:2302.11084*, 2023. 2, 3, 6

[76] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 2

[77] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. *arXiv preprint arXiv:2304.01195*, 2023. 2, 3, 7, 8