FEW-SHOT DUAL-PATH ADAPTATION OF VISION-LANGUAGE FOUNDATION MODELS

Ce Zhang **Simon Stepputtis** **Katia Sycara** **Yaqi Xie**
School of Computer Science, Carnegie Mellon University
{cezhang, sstepput, katia, yaqix}@cs.cmu.edu

ABSTRACT

Leveraging vast datasets on the Internet, large-scale Vision-Language Models (VLMs) demonstrates great potential in learning open-world visual concepts, and exhibit remarkable performance across a wide range of downstream tasks through efficient fine-tuning. In this work, we propose a simple yet effective fine-tuning approach called DualAdapter, which for the first time investigates the inference capabilities of VLMs along both positive and negative directions. Unlike conventional approaches that solely rely on positive adapter-style fine-tuning, DualAdapter uniquely incorporate negative text descriptions and image samples, enabling fine-tuning from a dual perspective. During the few-shot adaptation process, our DualAdapter explicitly enhances correct alignments while simultaneously minimizing incorrect associations. Our rigorous evaluation across 15 datasets reveals that DualAdapter significantly surpasses existing state-of-the-art methods in terms of both adaptation efficiency and robustness to distribution shifts.

1 INTRODUCTION

Extensive pre-trained Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021) and CoCa (Yu et al., 2022), provide a new paradigm for multi-modal learning and generalizable visual recognition (Radford et al., 2021). Recently, researchers have demonstrated the versatility and effectiveness of these VLMs in interpreting complex visual and textual inputs, as evidenced by their superior performance in a variety of vision-language tasks, *e.g.*, visual reasoning (Shu et al., 2022; Zhang et al., 2023), and visual question answering (Zhou et al., 2022c; Duan et al., 2022).

To transfer well-learned knowledge from VLMs to downstream datasets, a variety of efficient fine-tuning approaches from two main categories have been developed: prompt tuning methods and adapter-style methods. (1) Prompt tuning methods are designed to create adaptive input prompts, which update the textual classifier for the specific downstream task. For instance, CoOp (Zhou et al., 2022b) firstly introduces the prompt tuning method to fine-tune CLIP. Building on this, CoCoOp (Zhou et al., 2022a) enhances the generalizability of CoOp by learning prompts conditioned on each input image. (2) Adapter-style methods, on the other hand, directly modulate the textual or/and visual features produced by CLIP's encoders (Zhang et al., 2023; 2024). Notable approaches include Tip-Adapter (Zhang et al., 2022) and TaskRes (Yu et al., 2023), which focus on adjusting the visual and textual features for enhanced task-specific performance, respectively.

In this work, we propose DualAdapter to further explore the potential of adapter-style fine-tuning for VLMs. Unlike prior methods that solely rely on standard text descriptions ("*A photo of a* {`CLASS`}") and few-shot image samples to encourage positive class predictions, we introduce negative prompts ("*A photo of no* {`CLASS`}") and negative image pseudo-samples, which enables a reverse prediction problem. The idea behind this approach is also straightforward: we aim to enhance the model's ability to discern not just what an image is, but also what it is not. By designing our DualAdapter to adapt VLMs in both positive and negative directions, our method achieves state-of-the-art performance across 11 few-shot learning datasets, surpassing the second-best by 1.92% in 16-shot average accuracy.

Our key contributions are as follows: (1) We explore and exploit the negative inference capabilities of VLMs, and for the first time adopt a dual-path inference approach for adapting CLIP. (2) We introduce DualAdapter, a novel framework that incorporates positive and negative adapters across both vision and language modalities, ensuring efficient and effective adaptation. (3) Through extensive

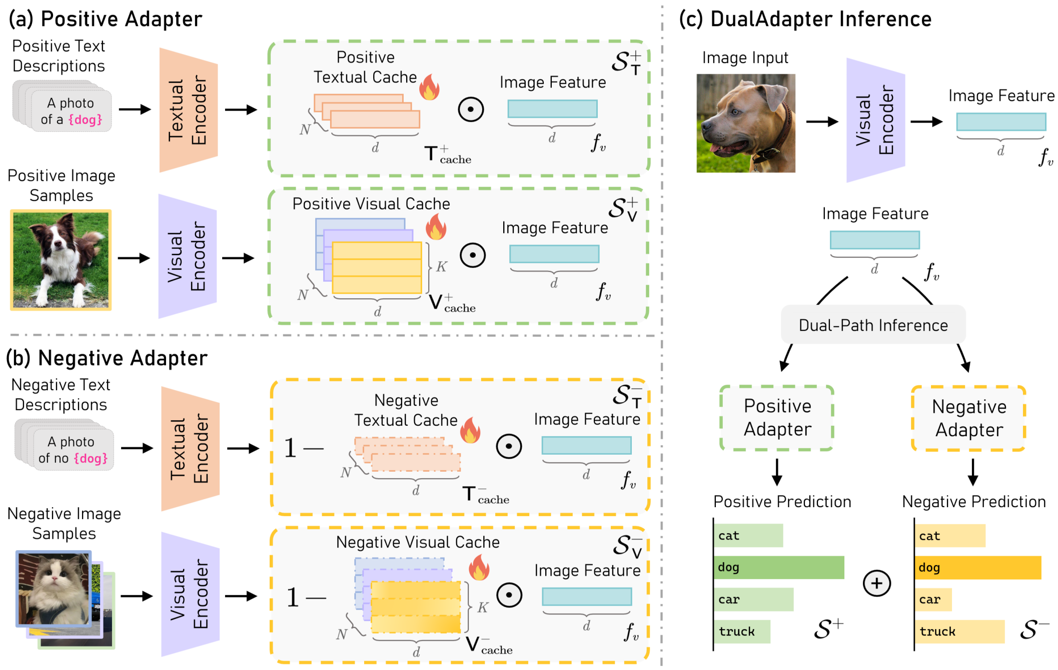

Figure 1: **An overview of our proposed DualAdapter**. The positive/negative adapter caches features from positive/negative text descriptions and positive/negative image samples. Given an image to be classified, the classification logit for a specific class increases when the image feature closely aligns with the features in the positive cache and diverges from those in the negative cache.

experiments, we demonstrate that our DualAdapter significantly improves adaptation performance and achieves superior generalizability across out-of-domain datasets.

2 Method

2.1 A Revisit of CLIP

In this work, we employ CLIP's pretrained visual encoder $\mathcal{F}_{\mathsf{V}} : \mathbb{R}^{h\times w\times 3} \to \mathbb{R}^d$ and textual encoder $\mathcal{F}_{\mathsf{T}} : \mathbb{R}^{m\times d_t} \to \mathbb{R}^d$ to map the images and textual descriptions into a unified d-dimensional embedding space. Consider an N-class classification task, CLIP conducts zero-shot predictions by evaluating the similarity between the image feature and N class-specific text features as follows:

$$f_v = \mathcal{F}_{\mathsf{V}}(\mathcal{I}), \quad f_{t_i}^+ = \mathcal{F}_{\mathsf{T}}(\mathcal{T}_i^+), \quad \mathbb{P}(y = y_i|x) = \frac{\exp\left(\mathrm{sim}\left(f_{t_i}^+, f_v\right)/\tau\right)}{\sum_{t'} \exp\left(\mathrm{sim}\left(f_{t'}^+, f_v\right)/\tau\right)}, \tag{1}$$

where $\mathcal{I}$ denotes the input image in $\mathbb{R}^{h\times w\times 3}$, and $\mathcal{T}_i^+$ represents the m-word sentence embedding of the class descriptor prompt "*A photo of a* $\{\mathtt{CLASS}_i\}$" in $\mathbb{R}^{m\times d_t}$. The term τ refers to the temperature parameter in the softmax function, and $\mathrm{sim}(\cdot,\cdot)$ computes the cosine similarity. Given that both the image and text features are L2-normalized ($\|f_t\|_2 = \|f_v\|_2 = 1$), the cosine similarity is effectively a dot product, *i.e.*, $\cos(f_t, f_v) = f_v^\top f_t$.

To streamline this process, a weight matrix can be precomputed and stored in a textual cache, which concatenates the textual features associated with each class, denoted as $\mathsf{T}_{\mathtt{cache}}^+ = [f_{t_1}^+\, f_{t_2}^+ \cdots f_{t_N}^+]^\top \in \mathbb{R}^{N\times d}$. Subsequently, we can efficiently obtain the logit $\mathcal{S}$ and the final prediction $\mathbb{P}(y|\mathcal{I})$ via vectorized computation:

$$\mathcal{S} = f_V {\mathsf{T}_{\mathtt{cache}}^+}^\top \in \mathbb{R}^{1\times N}, \quad \mathbb{P}(y|\mathcal{I}) = \mathsf{Softmax}\left(\mathcal{S}\right). \tag{2}$$

2.2 Our Proposed DualAdapter

We propose a novel framework DualAdapter, as illustrated in Figure 1, to enable a more efficient dual-path adaptation of VLMs. We introduce the each component of our DualAdapter in detail below.

Positive Textual Adapter. To adapt the VLMs to downstream tasks, we introduce a group of learnable parameters $\mathcal{R}_{\mathsf{T}}^{+} \in \mathbb{R}^{N \times d}$. These parameters are added element-wise to the text cache in a residual form, updating the positive text cache. Using this updated text cache, we then calculate the logit $\mathcal{S}_{\mathsf{T}}^{+}$ given the input image feature f_v. This process can be formally denoted as:

$$\mathsf{T}_{\mathsf{cache}}^{+} \leftarrow \mathsf{Normalize}\left(\mathsf{T}_{\mathsf{cache}}^{+} + \mathcal{R}_{\mathsf{T}}^{+}\right), \quad \mathcal{S}_{\mathsf{T}}^{+} = f_v {\mathsf{T}_{\mathsf{cache}}^{+}}^{\top} \in \mathbb{R}^{1 \times N}. \tag{3}$$

Note that both f_v and $\mathsf{T}_{\mathsf{cache}}^{+}$ are L^2-normalized, thus the cosine similarity simplifies to a dot product.

Negative Textual Adapter. Recall that the positive textual cache, denoted as T_{cache}^{+}, is constructed based on the class descriptor prompt "*A photo of a* $\{\mathit{CLASS}_i\}$". In a corresponding manner, we introduce negative prompts in the form of "*A photo of no* $\{\mathit{CLASS}_i\}$", and extract the text embeddings $f_{t_i}^{-} = \mathcal{F}_{\mathsf{T}}(\mathcal{T}_i^{-})$ using CLIP's textual encoder. For all N classes, we store the negative text embeddings $\{f_{t_i}^{-}\}_{i=1}^{N}$ in a cache matrix $\mathsf{T}_{\mathsf{cache}}^{-} \in \mathbb{R}^{N \times d}$. We similarly incorporate a learnable residual $\mathcal{R}_{\mathsf{T}}^{-} \in \mathbb{R}^{N \times d}$ to refine the text embedding throughout task-specific training.

The intuition behind this approach is straightforward: if an image is associated with a particular class, its feature representation should align closely with the positive prompt embeddings and diverge from those of the negative prompts. Specifically, the logit $\mathcal{S}_{\mathsf{T}}^{-}$ for the negative textual adapter is given by:

$$\mathsf{T}_{\mathsf{cache}}^{-} \leftarrow \mathsf{Normalize}\left(\mathsf{T}_{\mathsf{cache}}^{-} + \mathcal{R}_{\mathsf{T}}^{-}\right), \quad \mathcal{S}_{\mathsf{T}}^{-} = \delta_{\mathsf{T}}\left(1 - f_v {\mathsf{T}_{\mathsf{cache}}^{-}}^{\top}\right) \in \mathbb{R}^{1 \times N}, \tag{4}$$

where δ_{T} is a fixed scaling parameter that adjusts $\mathcal{S}_{\mathsf{T}}^{-}$ to match the mean value of $\mathcal{S}_{\mathsf{T}}^{+}$.

Positive Visual Adapter. Given an N-class K-shot training dataset, we utilize the NK annotated images to classify the input image from a visual perspective. Utilizing the pre-trained visual encoder of CLIP, we first extract the image features $\{f_v^{+}\}_{i=1}^{NK}$ and store them in a positive visual cache $\mathsf{V}_{\mathsf{cache}}^{+} \in \mathbb{R}^{NK \times d}$. To update the training features during the training stage, we introduce a set of learnable parameters $\mathcal{R}_{\mathsf{V}}^{+} \in \mathbb{R}^{N \times d}$, which are broadcast to $\mathbb{R}^{NK \times d}$ and added to the positive visual cache: $\mathsf{V}_{\mathsf{cache}}^{+} \leftarrow \mathsf{Normalize}\left(\mathsf{V}_{\mathsf{cache}}^{+} + \mathcal{R}_{\mathsf{V}}^{+}\right)$. Given an image feature f_v to be classified, we calculate its image-image affinities A^{+} with all the training images following Zhang et al. (2022), then multiplied by their corresponding one-hot labels $L \in \mathbb{R}^{NK \times N}$ to obtain the classification logit:

$$A^{+} = \exp\left(-\beta\left(1 - f_v {\mathsf{V}_{\mathsf{cache}}^{+}}^{\top}\right)\right) \in \mathbb{R}^{1 \times NK}, \quad \mathcal{S}_{\mathsf{V}}^{+} = \alpha A^{+} L \in \mathbb{R}^{1 \times N}, \tag{5}$$

where α represents a balance factor and β denotes a modulating hyper-parameter.

Negative Visual Adapter. Drawing inspirations from contrastive learning (Khosla et al., 2020), we generate some pseudo-negative prototypes from the few-shot training set. More specifically, for class i, we consider the K-shot images from the remaining $N-1$ classes as negative samples. To mitigate individual biases, we randomly select one image from each of the for each of the $N-1$ classes and compute the average of their extracted features to represent the pseudo-negative prototypes. In this way, we can get a total of K pseudo-negative prototypes for each of the N classes, thereby constructing a negative visual cache $\mathsf{V}_{\mathsf{cache}}^{-} \in \mathbb{R}^{NK \times d}$. The cache is further refined using a set of learnable parameters $\mathcal{R}_{\mathsf{V}}^{-} \in \mathbb{R}^{N \times d}$: $\mathsf{V}_{\mathsf{cache}}^{-} \leftarrow \mathsf{Normalize}\left(\mathsf{V}_{\mathsf{cache}}^{-} + \mathcal{R}_{\mathsf{V}}^{-}\right)$.

Following the same intuition with the negative textual adapter, we consider the reverse classification problem and calculate the logit as:

$$A^{-} = \delta_{\mathsf{V}} \exp\left(-\beta f_v {\mathsf{V}_{\mathsf{cache}}^{-}}^{\top}\right) \in \mathbb{R}^{1 \times NK}, \quad \mathcal{S}_{\mathsf{V}}^{-} = \alpha A^{-} L \in \mathbb{R}^{1 \times N}, \tag{6}$$

where δ_{V} is another fixed scaling parameter that adjusts A^{-} to match the mean value of A^{+}.

DualAdapter Inference. To derive the final classification scores, we aggregate the outputs from both the positive and negative adapters across textual and visual modalities. This is formalized as:

$$\mathcal{S}_{\mathsf{final}} = \lambda\left(\mathcal{S}_{\mathsf{T}}^{+} + \mathcal{S}_{\mathsf{V}}^{+}\right) + (1 - \lambda)\left(\mathcal{S}_{\mathsf{T}}^{-} + \mathcal{S}_{\mathsf{V}}^{-}\right). \tag{7}$$

Here, λ serves as a tuning hyper-parameter to balance the contribution of positive and negative adapter logits. During the training process, the set of learnable parameters $\{\mathcal{R}_{\mathsf{T}}^{+}, \mathcal{R}_{\mathsf{T}}^{-}, \mathcal{R}_{\mathsf{V}}^{+}, \mathcal{R}_{\mathsf{V}}^{-}\}$ is updated through gradient descent, leveraging a cross-entropy loss function.

3 EXPERIMENTS

To validate the effectiveness of our proposed DualAdapter, we evaluate our proposed method on two standard benchmarking tasks: few-shot learning and domain generalization, respectively. We compare our proposed method with the following state-of-the-art methods: zero-shot and linear probe CLIP (Radford et al., 2021), CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), CLIP-Adapter (Gao et al., 2023), Tip-Adapter-F (Zhang et al., 2022), TPT (Shu et al., 2022), TaskRes (Yu et al., 2023), and GraphAdapter (Li et al., 2023).

3.1 IMPLEMENTATION DETAILS

Following previous works (Zhou et al., 2022a; Zhang et al., 2022), we adopt ResNet-50 (He et al., 2016) backbone as the visual encoder of CLIP in our experiments by default. We adopt prompt ensembling, leveraging textual prompts from both CLIP (Radford et al., 2021) and CuPL (Pratt et al., 2023) to enhance model performance. Our DualAdapter is trained using the AdamW optimizer with a cosine scheduler. The batch size is set to 256. For $\mathcal{R}_{\mathsf{T}}^{+}$ and $\mathcal{R}_{\mathsf{V}}^{+}$, the learning rate is set to 0.0001, while for $\mathcal{R}_{\mathsf{T}}^{-}$ and $\mathcal{R}_{\mathsf{V}}^{-}$, the learning rate is set to 0.0005. Additionally, our model is trained for 200 epochs on the EuroSAT Helber et al. (2019) dataset, and for 20 epochs on all other datasets. All experiments are conducted on a single NVIDIA RTX 6000 Ada GPU.

3.2 FEW-SHOT LEARNING

Following previous literature on efficient fine-tuning of the CLIP model (Zhou et al., 2022b; Zhang et al., 2022), we comprehensively evaluate our method on 11 well-known image classification benchmarks: ImageNet (Deng et al., 2009), Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), StandfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food-101 (Bossard et al., 2014), FGVC Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014), SUN397 (Xiao et al., 2010), EuroSAT (Helber et al., 2019), and UCF101 (Soomro et al., 2012).

In Figure 2, we compare the few-shot learning performance of our proposed DualAdapter with other state-of-the-art methods on 11 image classification datasets. In the top-left sub-figure, we also present the average classification accuracy across all 11 datasets. The results indicate that our proposed DualAdapter consistently outperforms other methods across various few-shot learning protocols by large margins (*e.g.*, by an average of 1.92% in the 16-shot setting).

3.3 ROBUSTNESS TO NATURAL DISTRIBUTION SHIFTS

We follow CoOp (Zhou et al., 2022b) to investigate the generalization capability of our proposed method on 4 variant datasets of ImageNet: ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a). In Table 1, we compare the performance results of our proposed DualAdapter and other methods using ResNet-50 visual backbone in the presence of distribution shifts. We can observe that our proposed DualAdapter outperforms other state-of-the-art methods on both source and target domains, showcasing its remarkable generalizability.

Table 1: **Performance comparison on robustness to distribution shifts**. All the models are trained on 16-shot ImageNet and directed tested on the OOD target datasets. The best results are in **bold** and the second are <u>underlined</u>.

Method	Source	Target				
	ImageNet	-V2	-Sketch	-A	-R	Avg.
Zero-Shot CLIP$_{21}$	60.33	53.27	35.44	21.65	56.00	41.59
Linear Probe CLIP$_{21}$	56.13	45.61	19.13	12.74	34.86	28.09
CoOp$_{22}$	63.33	55.40	34.67	23.06	56.60	42.43
CoCoOp$_{22}$	62.81	55.72	34.48	23.32	57.74	42.82
TPT$_{22}$	60.74	54.70	35.09	**26.67**	59.11	43.89
TaskRes$_{23}$	64.75	56.47	35.83	22.80	60.70	43.95
GraphAdapter$_{23}$	<u>64.94</u>	<u>56.58</u>	<u>35.89</u>	23.07	<u>60.86</u>	<u>44.10</u>
DualAdapter (Ours)	**66.52**	**57.87**	**36.38**	<u>25.73</u>	**61.12**	**45.28**

3.4 ABLATION STUDIES

In Table 2, we conduct a systematic analysis of the impacts of various components within our Dual-Adapter framework. More specifically, we assess the performance of four distinct DualAdapter variants, each configured to allow two adapters to be updated while keeping the others fixed. We have the

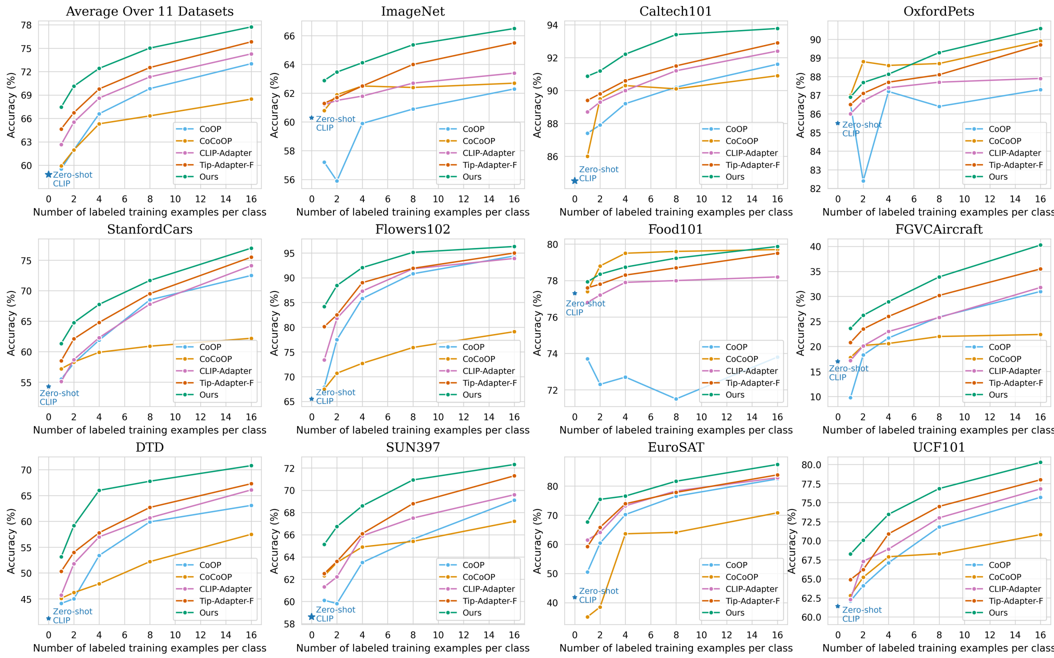

Figure 2: **Performance comparisons on few-shot learning on 11 image classification datasets**. For each dataset, we report the accuracy on 1-/2-/4-/8-/16-shot settings.

Table 2: **Ablation studies for different variants of** DualAdapter. We evaluate the few-shot adaptation capabilities of four DualAdapter variants on ImageNet (Deng et al., 2009).

#	Method	$\mathcal{R}_{\mathsf{T}}^{+}$	$\mathcal{R}_{\mathsf{T}}^{-}$	$\mathcal{R}_{\mathsf{V}}^{+}$	$\mathcal{R}_{\mathsf{V}}^{-}$	1-shot	2-shot	4-shot	8-shot	16-shot
1	DualAdapter$^{\mathsf{T}}$	✓	✓	✗	✗	62.86	63.36	64.01	65.23	66.34
2	DualAdapter$^{\mathsf{V}}$	✗	✗	✓	✓	62.21	62.37	62.68	63.72	65.30
3	DualAdapter^{+}	✓	✗	✓	✗	62.83	63.31	63.95	65.13	66.27
4	DualAdapter^{-}	✗	✓	✗	✓	62.65	63.07	63.60	64.36	65.12
5	DualAdapter	✓	✓	✓	✓	**62.89**	**63.47**	**64.12**	**65.37**	**66.52**

following main observations: (1) Compared to zero-shot CLIP, all four variants demonstrate a performance improvement of approximately 5%$\sim$6% (from 60.33%) with 16-shot samples, indicating that each variant can operate effectively; (2) Relatively, the textual variant (DualAdapter$^{\mathsf{T}}$) and the positive variant (DualAdapter^{+}) demonstrate superior efficiency over the visual and negative counterparts.

4 Conclusion

In this work, we propose DualAdapter for effectively adapting vision-language models to downstream datasets. We innovatively design both positive and negative adapters spanning visual and textual modalities. Based on this, we further introduce a set of learnable residual parameters to learn task-specific knowledge efficiently with limited training data. Our extensive empirical evaluation across 15 diverse datasets demonstrates that DualAdaptor outperforms the state-of-the-art methods in both few-shot learning and domain generalization tasks.

Acknowledgement

This work has been funded in part by the Army Research Laboratory (ARL) under grant W911NF-23-2-0007 and W911NF-19-2-0146, and the Air Force Office of Scientific Research (AFOSR) under grants FA9550-18-1-0097 and FA9550-18-1-0251.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014. 4

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. 4

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. 4, 5

Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15651–15660, 2022. 1

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178–178, 2004. 4

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*, pp. 1–15, 2023. 4

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. 4

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a. 4

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b. 4

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*, pp. 4904–4916, 2021. 1

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020. 3

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 554–561, 2013. 4

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. In *Advances in Neural Information Processing Systems*, 2023. 4

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729. IEEE, 2008. 4

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. 4

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*, pp. 15691–15701, 2023. 4

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. 1, 4

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019. 4

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14274–14289, 2022. 1, 4

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. 4

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*, volume 32, pp. 10506–10518, 2019. 4

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. 4

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 1

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*, pp. 10899–10909, 2023. 1, 4

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Negative yields positive: Unified dual-path adapter for vision-language models. *arXiv preprint arXiv:2403.12964*, 2024. 1

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pp. 493–510. Springer, 2022. 1, 3, 4

Yi Zhang, Ce Zhang, Zihan Liao, Yushun Tang, and Zhihai He. Bdc-adapter: Brownian distance covariance for better vision-language reasoning. In *British Machine Vision Conference*, 2023. 1

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a. 1, 4

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, 2022b. 1, 4

Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16485–16494, 2022c. 1