

Appendix: Adaptive Multi-Modality Prompt Learning

Anonymous Authors

A: RELATED WORK

Vision-Language Models

Vision-language models (VLMs), an important research direction in the field of deep learning, are dedicated to establishing a tight connection between images and natural language for better understanding and processing of multi-modality information. The development of VLMs stems from the urgent need to integrate visual and linguistic capabilities, and this integration provides a new paradigm for tasks such as image understanding, automatic image annotation and visual question answering. Unlike traditional unimodal models, VLMs process image and text data by learning together, making the model more capable of understanding the semantic content in the image. For instance, CLIP [13] employs a visual-language contrastive learning approach for joint pre-training on diverse datasets, enabling the model to comprehend images and text within a unified embedding space. BLIP [9] introduces a novel vision-language pre-training framework that, through caption bootstrapping, effectively utilizes noisy web data. BLIP-2 [8] presents a streamlined vision-language pre-training strategy for BLIP, leveraging frozen image encoders and language models.

Prompt Learning

With the in-depth exploration of the field of natural language processing, prompt learning has become a highly prominent research direction in recent years. Prompt learning aims to guide models in generating more accurate and targeted outputs by designing effective prompt information. The key idea of this approach is to improve model performance by directing its attention to specific information. In text-related tasks, previous works have skillfully constructed prompts to guide models in targeted text generation, thereby enhancing task performance. For instance, PET [14] combines pre-trained language models with cloze-style reformulations, assigning soft labels to unlabeled data. Furthermore, the extension of this concept has also offered new perspectives for image-related tasks, enhancing model performance in multi-modality scenarios by designing prompts suitable for image data. For instance, VPT [6] introduces an efficient alternative to full fine-tuning for large-scale Transformer models in computer vision, achieving significant performance gains and outperforming full fine-tuning in various scenarios while reducing storage costs. VP [1] adds noise to every patch so that it can reduce the influence of meaningless patches, but it may influence the meaningful patches for prompt learning.

Prompt Learning in VLMs

In complex application scenarios, the relationships between images and language are often difficult to mine, which poses a challenge to the performance of VLMs. Recent research work aims to enable models to better understand and capture these complex relationships by introducing prompt learning. Specifically, the introduction of prompt learning allows models to focus on task-relevant information in a targeted manner, helping to more accurately model

the interactions between images and language. This includes the design of effective prompting strategies applicable to VLMs, as well as insights into how to fully utilize the potential of prompt learning in multi-modality tasks. For instance, CoCoOp [19] enhances VLMs adaptation by introducing input-conditional tokens, addressing over-fitting issues, and demonstrating improved performance on unseen classes and domain generalization. MaPLe [7] dynamically adjusting both vision and language branches, improving alignment, and achieving improved performance across diverse downstream tasks. In this paper, we follow prior works on prompt learning [7, 19, 20], utilizing CLIP as the backbone for multi-modality prompt learning.

B: EXPERIMENTS

Experimental Setting

We evaluate our AMMPL with 7 comparison methods in terms of one in-sample task (*i.e.*, few-shot learning) and two out-of-sample tasks (*i.e.*, generalization from base-to-novel classes and cross-data evaluation) on 9 benchmark datasets.

The used datasets include four fine-grained datasets (*i.e.*, OxfordPets [12], Flowers102 [11], Food101 [2], and FGVC Aircraft [10]), one generic-objects dataset, *i.e.*, Caltech101 [4], one satellite-image dataset, *i.e.*, EuroSAT [5], one texture dataset, *i.e.*, DTD [3], one action recognition dataset, *i.e.*, UCF101 [15], and one scene recognition dataset, *i.e.*, Sun397 [17]. Datasets-specific details are shown in Table 7. The comparison methods include three single-modality PL methods (*i.e.*, DLP [7], VPT [6], and VP [1]), one non-interactive multi-modality PL method, *i.e.*, IVLP [7], and three interactive multi-modality PL methods, *i.e.*, CoCoOp [19], DPT [18], and MaPLe [7]. We list the details of the comparison methods as follows:

- **DLP** introduces learnable tokens in each Transformer [16] block of the text encoder until a specific depth is reached (*i.e.*, 5-layer). This innovative method allows the model to adapt and refine its representations by incorporating learnable tokens at various levels within the text encoder architecture.
- **VPT** introduces an efficient alternative to full fine-tuning for large-scale Transformer models in computer vision. This groundbreaking methodology not only streamlines the training process but also significantly optimizes the utilization of computational resources.
- **VP** adds noise to every patch so that it can reduce the influence of meaningless patches, but it may influence the meaningful patches for prompt learning.
- **IVLP** combines deep vision and language prompts separately but lacks synergy between the branches during the learning of task-relevant context prompts.
- **CoCoOp** enhances VLMs adaptation by introducing input-conditional tokens, addressing over-fitting issues, and demonstrating improved performance on unseen classes and domain generalization.
- **DPT** proposes a dual-modality prompt tuning paradigm, simultaneously adapting text and visual prompts, with a

Method	Shot	Source (OxfordPets)	Target							
			Caltech101	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	Sun397	UCF101
CoCoOp	1	91.93(0.5)	84.97(3.5)	36.27(3.4)	37.30(8.7)	15.23(3.3)	53.40(8.3)	67.67(9.6)	50.00(6.2)	56.97(4.6)
MaPLe	1	83.83(5.8)	86.10(2.3)	29.87(2.4)	43.80(7.1)	11.50(6.3)	52.10(9.1)	75.13(3.2)	52.07(4.5)	52.77(7.4)
AMMPL	1	91.10(0.9)	87.53(0.9)	35.87(2.4)	44.70(4.1)	20.90(1.8)	57.90(4.0)	80.15(2.8)	54.80(1.5)	61.20(2.5)
CoCoOp	8	93.17(0.3)	88.47(2.9)	35.47(0.5)	40.20(5.8)	17.60(1.5)	60.47(3.4)	79.70(5.4)	56.87(3.1)	59.87(0.6)
MaPLe	8	92.53(0.8)	88.20(2.9)	41.37(3.3)	35.00(6.1)	18.17(1.9)	58.83(9.8)	76.33(9.0)	55.50(5.2)	58.07(3.1)
AMMPL	8	92.93(1.2)	88.63(1.4)	39.43(1.3)	44.10(3.7)	22.46(2.0)	60.90(3.0)	80.89(1.3)	54.83(1.2)	62.30(1.2)
CoCoOp	16	93.47(0.3)	88.70(1.3)	37.63(3.0)	39.20(8.3)	16.97(2.7)	61.33(1.7)	74.73(3.7)	55.20(1.3)	59.40(0.6)
MaPLe	16	92.50(0.5)	86.93(4.2)	39.00(3.0)	38.77(9.1)	15.63(7.5)	58.40(8.9)	74.90(9.9)	56.43(9.8)	60.40(6.4)
AMMPL	16	93.50(0.3)	88.93(1.2)	40.07(1.8)	42.93(4.9)	22.65(1.5)	60.40(3.2)	78.35(5.1)	55.63(2.1)	61.89(2.1)

Table 1: Classification accuracy (mean and standard deviation) over 3 runs of all interactive multi-modality methods (with ViT-B/16) in terms of cross-data evaluation with different shot numbers (i.e., 1-shot, 8-shot, and 16-shot) on all datasets.

(a) Caltech101				(b) DTD				(c) EuroSAT			
Combo	Base	Novel	HM	Combo	Base	Novel	HM	Combo	Base	Novel	HM
C1	97.77(0.3)	92.97(0.6)	95.31(0.4)	C1	77.83(1.2)	54.50(3.0)	64.11(1.7)	C1	87.30(1.9)	57.97(1.6)	69.67(1.7)
C3	97.80(0.1)	93.00(0.1)	95.34(0.1)	C3	77.30(0.5)	54.57(1.2)	63.97(0.7)	C3	85.63(2.7)	60.33(4.5)	70.79(3.4)
C1+C2	97.97(0.1)	93.07(0.7)	95.46(0.2)	C1+C2	79.17(0.6)	56.63(6.5)	66.03(1.1)	C1+C2	91.38(1.2)	59.38(6.5)	71.98(2.0)
C1+C2+C3	97.99(0.1)	94.59(0.1)	96.25(0.1)	C1+C2+C3	78.33(1.5)	58.43(3.1)	66.93(2.0)	C1+C2+C3	94.10(2.0)	67.39(6.3)	78.54(3.0)
(d) FGVCAircraft				(e) Flowers102				(f) Food101			
Combo	Base	Novel	HM	Combo	Base	Novel	HM	Combo	Base	Novel	HM
C1	35.14(0.6)	33.83(1.3)	34.47(0.8)	C1	95.43(0.2)	70.30(1.4)	80.96(0.4)	C1	89.07(0.4)	90.90(0.6)	89.98(0.5)
C3	34.37(0.5)	32.70(1.0)	33.51(0.6)	C3	94.97(1.2)	71.43(1.4)	81.53(1.3)	C3	90.67(0.2)	91.27(0.6)	90.96(0.3)
C1+C2	35.33(1.8)	30.29(9.3)	32.62(3.0)	C1+C2	95.37(0.3)	72.17(2.3)	82.16(0.5)	C1+C2	89.38(0.3)	91.34(0.5)	90.35(0.4)
C1+C2+C3	35.69(1.6)	35.91(1.3)	35.80(1.4)	C1+C2+C3	94.90(1.1)	74.61(1.3)	83.54(1.2)	C1+C2+C3	90.90(0.1)	92.10(0.2)	91.50(0.1)
(g) OxfordPets				(h) Sun397				(i) UCF101			
Combo	Base	Novel	HM	Combo	Base	Novel	HM	Combo	Base	Novel	HM
C1	94.73(0.1)	97.33(1.3)	96.01(0.2)	C1	78.47(0.3)	76.60(0.5)	77.52(0.4)	C1	81.47(1.2)	70.13(3.8)	75.38(1.8)
C3	95.20(0.4)	97.89(0.1)	96.52(0.2)	C3	81.27(0.5)	78.90(0.7)	80.07(0.6)	C3	81.27(0.5)	73.77(2.5)	77.34(0.9)
C1+C2	94.80(0.2)	97.49(0.7)	96.13(0.3)	C1+C2	79.00(0.2)	76.73(0.1)	77.85(0.1)	C1+C2	82.33(0.3)	72.60(2.4)	77.16(0.5)
C1+C2+C3	96.11(0.3)	98.03(0.1)	97.31(0.1)	C1+C2+C3	81.02(0.3)	78.49(0.3)	79.73(0.3)	C1+C2+C3	82.58(0.5)	76.72(1.2)	79.54(0.7)

Table 2: Classification accuracy (mean and standard deviation) over 3 runs of AMMPL with different components in generalization from base-to-novel classes at 16-shot learning on all datasets. Note that, “Base”, “Novel”, and “HM”, respectively, indicate the classification accuracy of the base classes, the novel classes, and the harmonic mean.

Combo	Source (Food101)	Target							
		Caltech101	DTD	EuroSAT	FGVCAircraft	Flowers102	OxfordPets	Sun397	UCF101
C1	83.13(0.8)	86.50(1.4)	33.27(2.5)	39.37(5.3)	12.53(2.6)	58.17(5.2)	78.00(4.9)	52.60(1.6)	58.60(3.0)
C3	84.90(0.5)	85.80(0.9)	34.80(4.2)	41.27(2.2)	11.40(4.5)	51.83(1.5)	82.70(4.0)	48.27(5.0)	55.37(8.8)
C1+C2	84.23(1.3)	87.63(1.8)	34.27(3.9)	39.20(3.9)	15.13(4.2)	58.37(3.7)	80.83(2.5)	51.60(3.0)	58.83(8.8)
C1+C2+C3	85.17(0.7)	87.23(1.3)	35.30(1.7)	45.90(1.4)	15.53(3.0)	59.70(3.0)	79.03(4.0)	54.53(0.4)	60.55(3.6)

Table 3: Classification accuracy (mean and standard deviation) over 3 runs of AMMPL with different components in cross-data evaluation with 1-shot learning on all datasets and the bold number represents the results in the whole column.

Mean	Caltech101	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	OxfordPets	Sun397	UCF101
0.60	89.97(1.1)	46.23(1.3)	28.07(9.7)	11.60(9.3)	73.73(2.2)	84.80(0.1)	92.38(0.5)	68.13(0.2)	58.43(9.4)
0.80	92.67(1.0)	48.13(1.3)	37.97(7.9)	19.00(5.3)	73.90(3.4)	84.80(1.1)	91.63(0.8)	68.20(0.4)	71.20(0.4)
0.90	92.47(1.0)	48.07(2.7)	43.30(4.2)	27.60(0.6)	74.30(1.4)	85.30(1.1)	91.20(0.9)	67.80(0.6)	71.33(0.2)
0.93	93.10(2.3)	51.93(1.9)	53.30(7.2)	28.07(0.6)	78.83(0.7)	84.93(1.1)	91.37(0.7)	68.03(0.5)	70.80(1.7)
0.95	93.33(0.1)	49.37(1.4)	59.27(1.2)	28.90(0.4)	75.57(1.1)	85.30(1.1)	91.03(1.1)	68.33(0.1)	72.40(1.1)

Table 4: Classification accuracy (mean and standard deviation) of AMMPL with different mean values of the initialized probability matrix at 1-shot on all datasets and the bold number represents the best results in the whole column.

(a) Caltech101				(b) DTD				(c) EuroSAT			
Mean	Base	Novel	HM	Mean	Base	Novel	HM	Mean	Base	Novel	HM
0.60	96.47(0.5)	93.40(0.2)	94.91(0.3)	0.60	73.27(4.0)	55.53(5.5)	63.17(4.6)	0.60	74.40(5.4)	43.53(9.9)	54.92(7.0)
0.80	97.67(0.1)	93.40(0.4)	95.48(0.2)	0.80	76.53(0.8)	55.80(1.2)	64.54(1.0)	0.80	83.30(8.3)	66.10(3.9)	73.71(5.3)
0.90	97.67(0.2)	94.10(1.4)	95.85(0.4)	0.90	77.53(0.5)	55.90(1.5)	64.96(0.8)	0.90	85.37(1.5)	58.50(5.1)	69.42(2.3)
0.93	97.80(0.4)	94.30(0.9)	96.02(0.6)	0.93	76.93(1.1)	54.23(6.3)	63.62(1.9)	0.93	90.70(1.8)	66.00(5.1)	76.40(2.7)
0.95	97.80(0.3)	94.43(0.6)	96.09(0.4)	0.95	78.40(1.2)	57.23(4.3)	66.16(1.9)	0.95	93.77(1.1)	69.40(4.6)	79.77(1.8)

(d) FGVCAircraft				(e) Flowers102				(f) Food101			
Mean	Base	Novel	HM	Mean	Base	Novel	HM	Mean	Base	Novel	HM
0.60	14.00(9.9)	25.17(9.9)	17.99(9.9)	0.60	91.13(0.8)	71.50(3.0)	80.13(1.3)	0.60	86.87(0.3)	91.63(0.2)	89.19(0.2)
0.80	28.77(6.5)	30.32(8.2)	29.52(7.3)	0.80	93.27(1.3)	72.80(1.7)	81.77(1.5)	0.80	89.40(0.1)	91.63(0.6)	90.50(0.2)
0.90	35.43(6.6)	30.60(9.2)	32.84(7.7)	0.90	93.53(0.9)	72.13(0.2)	81.45(0.3)	0.90	90.10(0.1)	91.73(0.3)	90.91(0.2)
0.93	33.50(1.0)	35.03(0.8)	34.25(0.9)	0.93	93.60(0.4)	73.97(0.3)	82.64(0.3)	0.93	90.90(0.1)	92.07(0.2)	91.48(0.1)
0.95	34.50(0.4)	32.43(2.1)	33.43(0.7)	0.95	94.80(0.5)	72.20(1.6)	81.97(0.8)	0.95	90.13(0.1)	91.57(0.5)	90.84(0.2)

(g) OxfordPets				(h) Sun397				(i) UCF101			
Mean	Base	Novel	HM	Mean	Base	Novel	HM	Mean	Base	Novel	HM
0.60	92.33(0.4)	97.13(0.7)	94.67(0.5)	0.60	76.60(0.2)	76.83(0.4)	76.71(0.3)	0.60	76.53(1.0)	74.77(2.4)	75.64(1.4)
0.80	94.50(0.6)	97.47(0.5)	95.96(0.5)	0.80	78.00(0.3)	77.27(0.4)	77.63(0.3)	0.80	79.37(0.7)	74.23(1.6)	76.71(1.0)
0.90	95.23(0.3)	96.93(0.8)	96.07(0.4)	0.90	78.37(0.1)	77.53(0.1)	77.94(0.1)	0.90	80.87(1.0)	74.47(2.0)	77.53(1.3)
0.93	95.43(0.1)	97.37(0.5)	96.39(0.2)	0.93	79.03(0.1)	77.27(0.4)	78.14(0.2)	0.93	82.53(0.4)	74.13(1.5)	78.10(0.6)
0.95	96.07(0.3)	98.10(0.1)	97.07(0.2)	0.95	81.01(0.4)	78.50(0.6)	79.74(0.5)	0.95	80.57(0.9)	75.93(1.8)	78.18(1.2)

Table 5: Classification accuracy (mean and standard deviation) over 3 runs of AMMPL with different mean values of the initialized probability matrix at 16-shot learning on all datasets. Note that, “Base”, “Novel”, and “HM”, respectively, indicate the classification accuracy of the base classes, the novel classes, and the harmonic mean.

		Target							
Mean	Source (Food101)	Caltech101	DTD	EuroSAT	FGVCAircraft	Flowers102	OxfordPets	Sun397	UCF101
0.60	84.80(0.1)	84.67(1.5)	32.97(1.5)	37.07(4.3)	12.30(6.5)	52.07(8.1)	74.53(8.4)	51.13(1.3)	61.70(2.3)
0.80	84.80(1.1)	89.33(2.7)	34.60(3.0)	40.53(6.3)	13.10(0.6)	58.97(6.3)	80.60(3.9)	54.03(2.9)	57.90(5.2)
0.90	85.30(1.1)	88.83(2.5)	38.20(4.8)	42.80(2.8)	14.77(4.9)	62.47(1.8)	80.93(0.5)	52.80(2.9)	60.13(2.3)
0.93	84.93(1.1)	87.07(1.1)	35.43(4.4)	46.37(7.0)	15.53(2.7)	55.97(3.1)	78.97(0.8)	51.90(2.3)	57.90(3.9)
0.95	85.30(1.1)	88.07(1.7)	32.30(1.5)	41.53(5.0)	14.77(3.2)	57.70(8.1)	76.73(1.1)	50.73(4.0)	60.03(4.6)

Table 6: Classification accuracy (mean and standard deviation) over 3 runs of AMMPL with different mean values of the initialized probability matrix in cross-data evaluation with 1-shot learning on all datasets.

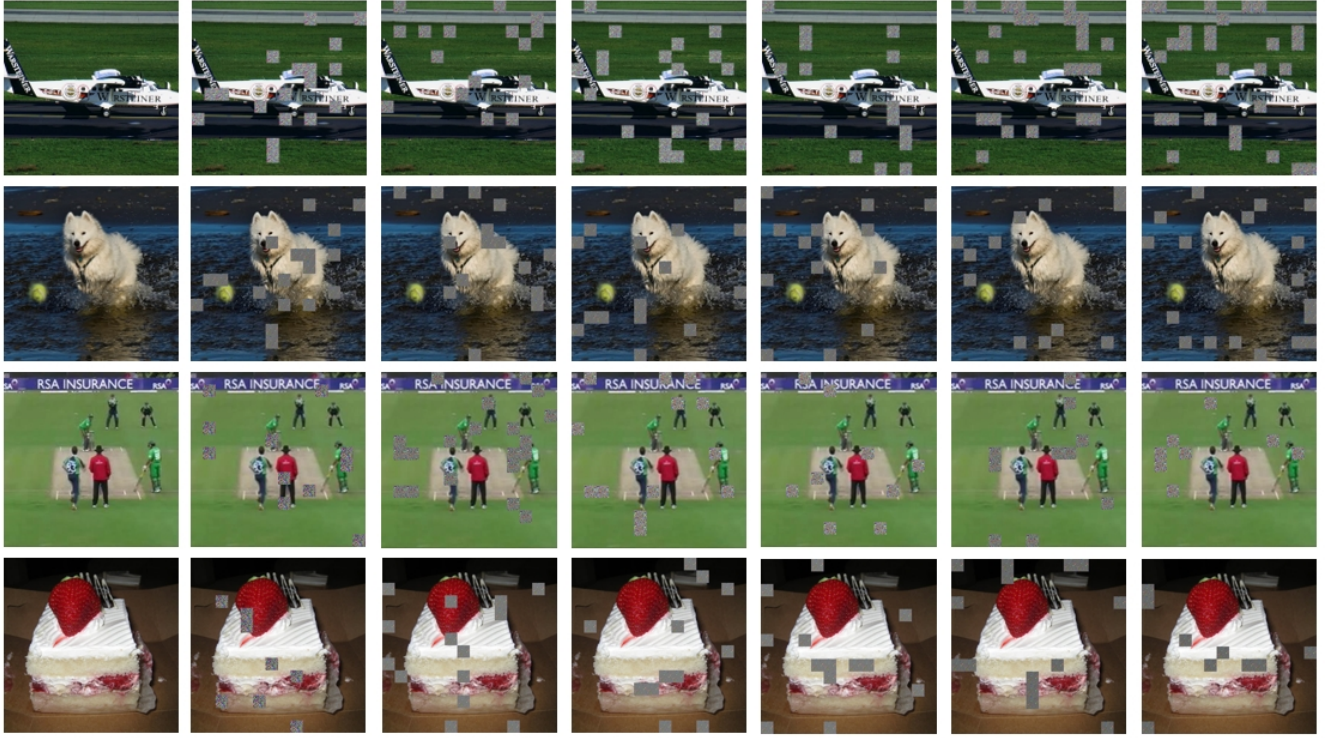


Figure 1: Visualization of masked patches by our method with increased iteration (to add more detail to Figure 3 in the main text). Note that, the first column is the training images, and the iterations increase from the 2nd column to the 7th column.

class-aware visual prompt tuning scheme for improved concentration on target visual concepts.

- **MaPLE** dynamically adjusting both vision and language branches, improving alignment, and achieving improved performance across diverse downstream tasks. This method involves real-time adaptation, where the model intelligently fine-tunes its vision and language components based on the specific requirements of the given task.

Next, we list the details of the two types of downstream tasks (*i.e.*, in-sample generalization task and out-of-sample generalization task) as follows:

- **Few-shot Learning.** To assess the performance of our proposed AMMPL on in-sample generalization task. We train the model using 1, 2, 4, 8, and 16 shots, and then evaluate its performance on a test set with the same classes as the training samples.
- **Base-to-Novel Generalization.** To preliminarily assess the performance of our proposed AMMPL on out-of-sample generalization tasks. We divide the dataset into base and novel classes (*i.e.*, no intersection between the two classes). Models are trained only in the base class and evaluated in the base and novel classes, respectively. Moreover, we employ a harmonic mean to comprehensively assess the generalization of the two types of tasks.
- **Generalization from Base-to-Novel Classes.** To further assess the performance of our method on out-of-sample generalization tasks, we directly evaluated our trained models on

other datasets. Specifically, we employed settings with 1, 8, and 16 shots to train the models on OxfordPets and Food101 datasets. Then, we performed cross-data evaluations on eight remaining datasets.

Cross-data Evaluation

We conduct a cross-data evaluation to further evaluate the out-of-sample generalization of the proposed AMMPL. This requires model training on one dataset and model evaluation on other datasets. Specifically, we first train all interactive PL methods (*i.e.*, our AMMPL, CoCoOp and MaPLE) on the source dataset (*i.e.*, either Food101 or OxfordPets) with different shot numbers (*i.e.*, 1, 8, and 16) and then test these methods on the remaining 8 datasets. We report the results in Table 1 for the source dataset OxfordPets.

REFERENCES

- [1] Hyojin Bahng, Ali Jahani, Swami Sankaranarayanan, and Phillip Isola. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274* (2022).
- [2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101—mining discriminative components with random forests. In *ECCV*. 446–461.
- [3] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *CVPR*. 3606–3613.
- [4] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*. 178–178.
- [5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226.

Datasets	Classes	Train	Val	Test
Sun397	397	15,880	3,970	19,850
Flowers102	102	4,093	1,633	2,463
Food101	101	50,500	20,200	30,300
UCF101	101	7,639	1,898	3,783
Caltech101	100	4,128	1,649	2,465
FGVCAircraft	100	3,334	3,333	3,333
DTD	47	2,820	1,128	1,692
OxfordPets	37	2,944	736	3,669
EuroSAT	10	13,500	5,400	8,100

Table 7: Description of the datasets.

- [6] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *ECCV*. 709–727.
- [7] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. MaPLe: Multi-modal Prompt Learning. In *CVPR*.
- [8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
- [9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. 12888–12900.

- [10] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* (2013).
- [11] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. 722–729.
- [12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *CVPR*. 3498–3505.
- [13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. 8748–8763.
- [14] Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676* (2020).
- [15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision* 2, 11 (2012).
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
- [17] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*. 3485–3492.
- [18] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, Peng Wang, and Yanning Zhang. 2023. Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia* (2023).
- [19] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.
- [20] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision* (2022).