A. Implementation Details

We implement DAPT based on the open source from CoOp [17] and VPT [6], using the PyTorch [11] library. Before training, the learnable vectors for the text prompt are initialized with a zero-mean Gaussian distribution following the CoOp. In contrast, the learnable vectors for the visual prompt are initialized with the Xavier uniform initialization scheme following the VPT. In all experiments, the length of the learnable vector is set to 16 in both the text and visual prompt. In the case of linear probe CLIP [12] and zero-shot CLIP [12], we set the initialization of the text prompt as "`a photo of a [CLASS].`" In the few-shot learning experiments, we follow the approach of Zhou *et al.* [17] and conduct random sampling three times for each dataset. We report the average after testing three times for all experiments, including DAPT, CoOp [17], VPT [6], and linear probe CLIP. As observed in the case of VPT, there is variance in the results of the visual prompt depending on hyperparameters such as learning rate. Therefore, we conduct a grid search for learning rate in the range of {0.002, 0.02, 0.2, 2.0, 20.0}, following the approach of Jia *et al.* [6].

B. Additional Analyses

In this section, we further analyze DAPT with various experiments.

B.1. Analysis of Hyperparameter β_t and β_v.

The hyperparameter β_t and β_v adjust the strength of inter-dispersion loss and intra-dispersion loss, respectively. Due to the characteristics of each dataset, it may have different optimal values. We depict the accuracy according to β_t and β_v in Figure A1 to investigate the hyperparameters.

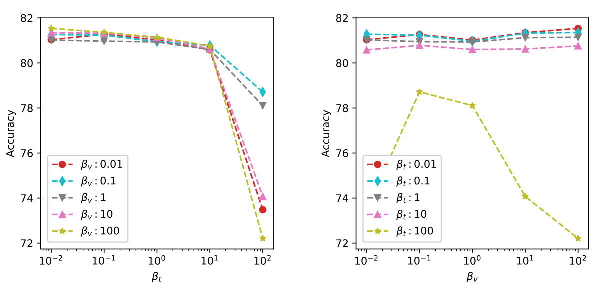


Figure A1: Exploration of hyperparamters.

As described in Figure A1, DAPT shows consistent performance with a wide range of hyperparameters. To sum up, DAPT is robust to the choice of hyperparameters, except $\beta_t = 100$. Table A3 summarizes optimal hyperparameters for each dataset.

B.2. t-SNE Visualization

Figure A2 presents t-SNE [14] visualization of image embeddings in zero-shot CLIP [12] and DAPT. All plots show that DAPT properly increases the distance between different classes as well as minimizes the intra-class variance. Especially the results on OxfordPets [10], Flowers102 [9], and UCF101 [13] demonstrate that DAPT helps embeddings form compact clusters and increase the distance between different classes.

B.3. Detailed Experimental Results

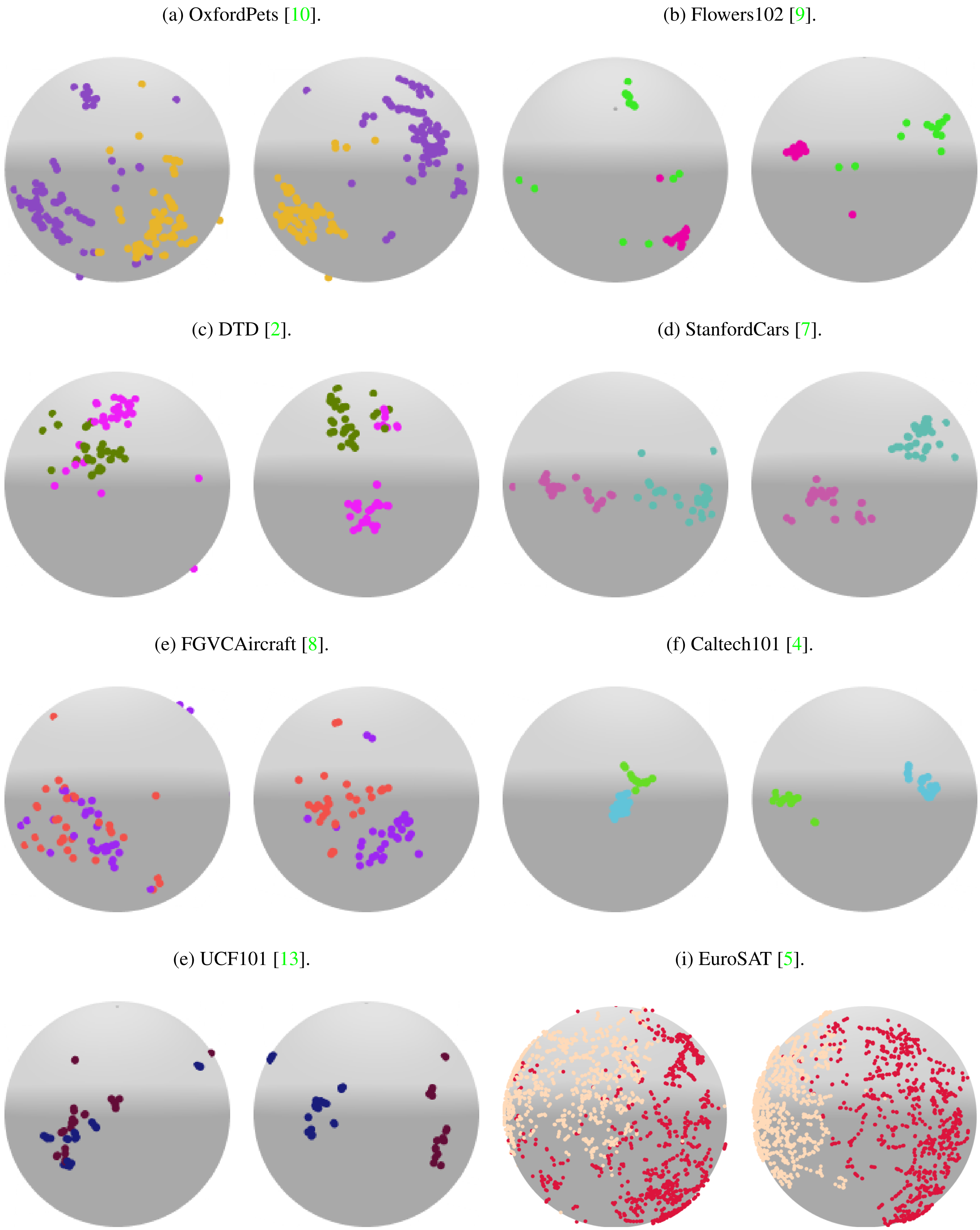
In all experiments, we ran three times with randomly sampled data in each run and noted average values. For compelling results, we provide accuracy with standard deviation in 16-shots image classification on 11 datasets in Table A1. On average, DAPT achieved the best performance in 10 benchmarks.

Dataset	LP-CLIP	CoOp	VPT	DAPT
OxfordPets	86.49$\pm$0.06	91.91$\pm$0.42	92.04$\pm$0.58	**92.27**$\pm$0.40
Flowers102	**97.51**$\pm$0.13	96.79$\pm$0.35	91.48$\pm$0.15	97.06$\pm$0.25
FGVCAircraft	45.51$\pm$0.08	43.96$\pm$0.74	34.92$\pm$0.16	**46.37**$\pm$1.00
DTD	69.58$\pm$0.73	69.98$\pm$0.18	61.47$\pm$0.37	**71.38**$\pm$1.62
EuroSAT	87.24$\pm$0.23	85.58$\pm$1.63	90.67$\pm$1.44	**92.65**$\pm$0.86
StanfordCars	80.67$\pm$0.46	82.62$\pm$0.13	70.59$\pm$1.00	**83.03**$\pm$0.34
Food101	83.14$\pm$0.45	84.31$\pm$0.17	86.03$\pm$0.27	**86.55**$\pm$0.10
SUN397	73.03$\pm$0.74	74.69$\pm$0.24	70.33$\pm$0.16	**75.99**$\pm$0.12
Caltech101	95.51$\pm$0.31	95.68$\pm$0.20	95.35$\pm$0.15	**95.82**$\pm$0.07
UCF101	82.35$\pm$0.36	82.15$\pm$1.42	79.99$\pm$0.69	**84.53**$\pm$0.55
ImageNet	67.42$\pm$0.26	71.93$\pm$0.10	69.31$\pm$0.05	**72.20**$\pm$0.18

Table A1: 16-shots image classification on 11 datasets.

C. Generalization From Base to New Classes

CoOp [17] demonstrated exemplary performance in the few-shot learning using text prompts, but it has a weak generalizability problem regarding unseen classes, as discussed in CoCoOp [16]. As shown in Table 2, we prove that DAPT has significant performance gain in generalization. However, we supplement more experiments to prove that DAPT has superior generalization performance compared with baselines. In all experiments, we evaluate not only original classes but also unseen classes. Following Zhou *et al.* [16], we divide the dataset into base classes and new classes, then train on 16 samples of the base class before testing on the new class. Similarly to the few-shot learning setting, we report the average of three times. The result for 11 datasets and the overall average is presented in Table A2. The experimental results show that the accuracy for the new class is higher than CoOp in most datasets. The harmonic mean of the base and new class demonstrates superior performance for seven datasets.



(a) OxfordPets [10].

(b) Flowers102 [9].

(c) DTD [2].

(d) StanfordCars [7].

(e) FGVCAircraft [8].

(f) Caltech101 [4].

(e) UCF101 [13].

(i) EuroSAT [5].

Figure A2: **t-SNE [14] visualization of image embeddings.** In each dataset, the left hypersphere represents zero-shot CLIP, and the right hypersphere represents DAPT.

(a) Average over 11 datasets.

	Base	New	H
CLIP	69.53	74.34	71.85
CoOp	82.71	62.84	71.42
DAPT	84.20	63.71	72.54

(b) OxfordPets [10].

	Base	New	H
CLIP	91.33	97.15	94.15
CoOp	93.37	95.43	94.39
DAPT	94.00	72.43	81.82

(c) Flowers102 [9].

	Base	New	H
CLIP	71.70	77.45	74.46
CoOp	97.82	59.79	74.21
DAPT	98.16	61.37	75.53

(d) FGVCAircraft [8].

	Base	New	H
CLIP	27.67	35.87	31.24
CoOp	40.66	24.44	30.53
DAPT	45.54	19.74	27.54

(e) DTD [2].

	Base	New	H
CLIP	53.24	60.87	56.80
CoOp	79.59	40.30	53.51
DAPT	82.06	53.42	64.71

(f) EuroSAT [5].

	Base	New	H
CLIP	56.93	63.92	60.22
CoOp	92.21	50.70	65.43
DAPT	95.05	43.02	59.23

(g) StanfordCars [7].

	Base	New	H
CLIP	63.93	74.99	69.02
CoOp	77.70	59.39	67.32
DAPT	79.69	57.46	66.77

(h) Food101 [1].

	Base	New	H
CLIP	90.08	91.13	90.60
CoOp	88.40	85.87	87.11
DAPT	89.57	89.82	89.69

(i) SUN397 [15].

	Base	New	H
CLIP	69.46	75.56	72.38
CoOp	80.64	65.43	72.24
DAPT	81.87	74.80	78.18

(j) Caltech101 [4].

	Base	New	H
CLIP	97.22	94.21	95.69
CoOp	98.19	86.17	91.79
DAPT	98.24	87.74	92.69

(k) UCF101 [13].

	Base	New	H
CLIP	70.89	78.42	74.47
CoOp	84.80	55.62	67.17
DAPT	85.09	71.46	77.68

(l) ImageNet [3].

	Base	New	H
CLIP	72.40	68.12	70.19
CoOp	76.44	68.11	72.04
DAPT	76.97	69.54	73.07

Table A2: Comparison of CLIP, CoOp, and DAPT in the base-to-new generalization setting.

Hyperparameters	OxfordPets	Flowers102	FGVCAircraft	DTD	EuroSAT	StanfordCars	Food101	SUN397	Caltech101	UCF101	ImageNet
β_t	0.1	0.01	0.01	0.01	10.0	0.1	0.01	0.01	0.01	0.1	0.01
β_v	10.0	10.0	100.0	100.0	100.0	100.0	10.0	100.0	10.0	0.1	0.01
Learning rate	0.02	0.002	2.0	20.0	20.0	0.002	20.0	20.0	0.2	20.0	2.0

Table A3: Hyperparameters on 11 datasets in few-shot learning.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 3

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2, 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[4] 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 *CVPRW*, 2004. 2, 3

[5] 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*, 2019. 2, 3

[6] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1

[7] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 2, 3

[8] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 2, 3

[9] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 1, 2, 3

[10] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 1, 2, 3

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[12] 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 *ICML*, 2021. 1

[13] 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. 1, 2, 3

[14] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 1, 2

[15] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 3

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1

[17] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 1