Supplementary for STYLIP: Multi-Scale Style-Conditioned Prompt Learning for CLIP-based Domain Generalization

Shirsha Bose[1*] Ankit Jha[2*] Enrico Fini[3] Mainak Singha[2*] Elisa Ricci[3] Biplab Banerjee[2]
[1]Technical University of Munich, Germany [2]Indian Institute of Technology Bombay, India
[3]University of Trento, Italy
shirshabosecs@gmail.com, ankitjha16@gmail.com, enrico.fini@unitn.it
mainaksingha.iitb@gmail.com, e.ricci@unitn.it, getbiplab@gmail.com

We mention the following discussions in the supplementary:

- Detailed dataset descriptions.
- Analysis of computational complexity.
- Detailed results for in-domain base-to-new class generalization over the 11 datasets in Table 2.
- Additional results on cross-domain base-to-new class generalization using *ClipArt* as the source domain for the Office-Home dataset in Table 3.

0.1. Dataset details

We evaluate STYLIP over five benchmark datasets for multi-source and single-source DG, which are described as follows: (1) **Office-Home** [23] - It consists of 15,500 images coming from 65 classes covering four domains, namely, Art, Clipart, Product, and Real. (2) **PACS** [13] - Includes 9991 images consisting of seven classes that are spread across four domains, Artpaint, Cartoon, Sketch, and Photo. (3) **VLCS** [14] - It was prepared by combining images from four image classification datasets, i.e., PASCAL VOC 2007 [4], Caltech [5], LabelMe [21], and Sun [24]. It consists of images from five classes, Bird, Car, Chair, Dog, and Person. (4) **Digits-DG** [28] - This dataset is designed in the combination of handwritten digit recognition datasets, namely, MNIST [12], MNIST-M [6], SVHN [16], and SYN [6]. (5) **DomainNet** [19] - It consists of images from six distinct domains, including real, painting, clipart, quickdraw, infograph, and sketch. Each domain has 48K - 172K images (600K in total) categorized into 345 classes.

We further analyse the performance of STYLIP for cross dataset generalization, where STYLIP is trained on ImageNet [11] and tested on 10 other different datasets, including Caltech101 [5], OxfordPets [18], StanfordCars [10], Flowers102 [17], Food101 [1], FGVCAircraft [15], SUN397 [24], DTD [3], EuroSAT [8] and UCF101 [22].

Computation Complexity. We run our model on NVIDIA RTX 3090 Ti with 24 GB card. Tab. 1 represents the comparison of computational complexity between different prompting techniques (CoOp [27], CoCoOp [26], and MaPLe [9]) in terms of GFLOPS relative to CoOp. MaPLe requires 0.12% more computational overhead than CoOp and CoCoOp, whereas STYLIP needs 0.18% more resources than MaPLe, but STYLIP outperforms state-of-the-art MaPLe on the cross-dataset generalization (average over 11 datasets) approximately by 1.2%.

Table 1. Increase in compute w.r.t. CoOp and CoCoOp.

CoOp [27]	CoCoOp [26]	MaPLe [9]	STYLIP
1×	1×	+0.12%	+0.18%

*equal contribution

References

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

[2] Adrian Bulat and Georgios Tzimiropoulos. Language-aware soft prompting for vision & language foundation models. *arXiv preprint arXiv:2210.01115*, 2022. 2

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[5] 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

Table 2. Comparison with state-of-the-art methods on base-to-new generalization. STYLIP shows better generalization performance over existing methods on 11 different recognition datasets on 16-shots and a context length of four. HM represents the harmonic mean. (In %)

(a) **Average over 11 datasets**

	Base	New	HM
CLIP [20]	69.34	74.22	71.70
CoOp [27]	82.69	63.22	71.66
CoCoOp [26]	80.47	71.69	75.83
LASP [2]	82.70	74.90	78.61
MaPLe [9]	82.28	75.14	78.55
STYLIP-con	82.64	75.39	78.85
STYLIP-sty	82.93	75.67	79.13
STYLIP*	82.30	75.24	78.61
STYLIP	**83.22**	**75.94**	**79.47**

(b) ImageNet

	Base	New	HM
CLIP [20]	72.43	68.14	70.22
CoOp [27]	76.47	67.88	71.92
CoCoOp [26]	75.98	70.43	73.10
LASP [2]	76.20	70.95	73.48
MaPLe [9]	76.66	70.54	73.47
STYLIP-con	76.81	70.74	73.65
STYLIP-sty	76.93	71.05	73.87
STYLIP*	76.34	70.46	73.28
STYLIP	**77.15**	**71.34**	**74.13**

(c) Caltech101

	Base	New	HM
CLIP [20]	96.84	94.00	95.40
CoOp [27]	98.00	89.81	93.73
CoCoOp [26]	97.76	93.81	95.84
LASP [2]	98.10	94.24	96.16
MaPLe [9]	97.74	94.36	96.02
STYLIP-con	97.74	94.83	96.26
STYLIP-sty	97.89	94.78	96.31
STYLIP*	97.45	94.61	96.01
STYLIP	**98.23**	**94.91**	**96.54**

(d) OxfordPets

	Base	New	HM
CLIP [20]	91.17	97.26	94.12
CoOp [27]	93.67	95.29	94.47
CoCoOp [26]	95.20	97.69	96.43
LASP [2]	95.90	97.93	96.90
MaPLe [9]	95.43	97.76	96.58
STYLIP-con	95.66	97.94	96.79
STYLIP-sty	95.82	98.02	96.91
STYLIP*	95.57	97.82	96.68
STYLIP	**95.96**	**98.14**	**97.04**

(e) StanfordCars

	Base	New	HM
CLIP [20]	63.37	**74.89**	68.65
CoOp [27]	**78.12**	60.40	68.13
CoCoOp [26]	70.49	73.59	72.01
LASP [2]	75.17	71.60	73.34
MaPLe [9]	72.94	74.00	73.47
STYLIP-con	73.83	74.15	73.99
STYLIP-sty	74.67	74.35	74.51
STYLIP*	73.16	73.92	73.54
STYLIP	75.19	74.46	**74.82**

(f) Flowers102

	Base	New	HM
CLIP [20]	72.08	**77.80**	74.83
CoOp [27]	**97.60**	59.67	74.06
CoCoOp [26]	94.87	71.15	81.71
LASP [2]	97.0	74.0	83.95
MaPLe [9]	95.92	72.46	82.56
STYLIP-con	96.14	72.75	82.83
STYLIP-sty	96.35	72.91	83.01
STYLIP*	96.02	72.53	82.64
STYLIP	96.54	73.08	**83.19**

(g) Food101

	Base	New	HM
CLIP [20]	90.10	91.22	90.66
CoOp [27]	88.33	82.26	85.19
CoCoOp [26]	90.70	91.29	90.99
LASP [2]	91.20	91.70	91.44
MaPLe [9]	90.71	92.05	91.38
STYLIP-con	90.92	92.23	91.57
STYLIP-sty	90.95	92.30	91.62
STYLIP*	90.84	92.11	91.47
STYLIP	**91.20**	**92.48**	**91.84**

(h) FGVCAircraft

	Base	New	HM
CLIP [20]	27.19	**36.29**	31.09
CoOp [27]	40.44	22.30	28.75
CoCoOp [26]	33.41	23.71	27.74
LASP [2]	34.53	30.57	32.43
MaPLe [9]	37.44	35.61	36.50
STYLIP-con	37.23	35.70	36.45
STYLIP-sty	37.51	35.75	36.61
STYLIP*	37.10	35.58	36.32
STYLIP	**37.65**	35.93	**36.77**

(i) SUN397

	Base	New	HM
CLIP [20]	69.36	75.35	72.23
CoOp [27]	80.60	65.89	72.51
CoCoOp [26]	79.74	76.46	78.27
LASP [2]	80.70	78.60	79.63
MaPLe [9]	80.82	78.70	79.75
STYLIP-con	81.23	78.94	80.07
STYLIP-sty	81.79	79.40	80.58
STYLIP*	80.95	78.80	79.86
STYLIP	**82.12**	**79.95**	**81.02**

(j) DTD

	Base	New	HM
CLIP [20]	53.24	59.90	56.37
CoOp [27]	79.44	41.18	54.24
CoCoOp [26]	77.01	56.00	64.85
LASP [2]	81.40	58.60	68.14
MaPLe [9]	80.36	59.18	68.16
STYLIP-con	80.76	59.44	68.48
STYLIP-sty	81.23	60.94	69.64
STYLIP*	80.22	59.60	68.39
STYLIP	**81.57**	**61.72**	**70.27**

(k) EuroSAT

	Base	New	HM
CLIP [20]	56.48	64.05	60.03
CoOp [27]	92.19	54.74	68.69
CoCoOp [26]	87.49	60.04	71.21
LASP [2]	94.60	77.78	85.36
MaPLe [9]	94.07	73.23	82.35
STYLIP-con	94.45	73.67	82.78
STYLIP-sty	94.57	73.85	82.94
STYLIP*	94.33	73.46	82.60
STYLIP	**94.61**	**74.06**	**83.08**

(l) UCF101

	Base	New	HM
CLIP [20]	70.53	77.50	73.85
CoOp [27]	84.39	56.05	67.46
CoCoOp [26]	82.33	73.45	77.64
LASP [2]	84.77	78.03	81.26
MaPLe [9]	83.00	78.66	80.77
STYLIP-con	84.24	78.93	81.50
STYLIP-sty	84.51	79.05	81.69
STYLIP*	83.37	78.72	80.98
STYLIP	**85.19**	**79.22**	**82.10**

2004 conference on computer vision and pattern recognition workshop, pages 178–178. IEEE, 2004. 1

[6] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature

Table 3. Analysis of the generalization from base to new classes across domains. We show results on Office-Home with *ClipArt* acting as the source domain, while others denote the target. The model is trained (backbone CLIP ViT-B/16) using 16 shots from the base classes. (In %)

Method	Office-Home				
	Base	New			
	Clip Art	Art	Clip Art	Product	Real World
CLIP [20]	78.12	62.01	77.78	87.52	88.02
CoOp [27]	82.60	70.60	82.23	90.44	87.21
CoCoOp [26]	82.64	71.00	83.61	92.12	89.19
CLIP-Adapter [7]	80.00	73.19	83.00	92.11	89.53
DPL [25]	82.20	71.54	82.80	92.37	89.15
ProGrad [29]	82.41	72.00	83.29	92.11	89.585
StyLIP-con	82.67	72.10	84.39	92.17	90.24
StyLIP-sty	83.22	72.60	84.51	92.78	91.25
StyLIP*	83.90	73.48	85.07	92.60	90.77
StyLIP	**84.33**	**74.60**	**87.25**	**93.00**	**91.42**

adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3

[8] 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. 1

[9] 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 (CVPR)*, pages 19113–19122, June 2023. 1, 2

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[13] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 1

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 1

[15] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 1

[16] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 1

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 1

[18] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[19] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 1

[20] 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. 2, 3

[21] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008. 1

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[23] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 1

[24] 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. 1

[25] Xin Zhang, Yusuke Iwasawa, Yutaka Matsuo, and Shixiang Shane Gu. Amortized prompt: Lightweight fine-tuning for clip in domain generalization. *arXiv preprint arXiv:2111.12853*, 2021. 3

[26] 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. 1, 2, 3

[27] 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

[28] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020. 1

[29] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 3