GOPRO: Generate and Optimize Prompts in CLIP using Self-Supervised Learning

Mainak Singha
mainaksingha.iitb@gmail.com
Ankit Jha
ankitjha16@gmail.com
Biplab Banerjee
getbiplab@gmail.com

Indian Institute of Technology Bombay
Mumbai, India

1 Introduction

Our supplementary material includes the following discussions / evaluations:

- Detail description of the dataset used in this paper is provided in Section 2.
- Section 3 discusses the t-SNE [18] visualization in more details for base and new classes separately for SLIP, CoCoOp and GOPRO on Eurosat dataset.
- In section 4, we present the detailed results for base-to-new generalization task and compare the performance of our proposed GOPRO and the referred SOTA prompting techniques.
- We also discuss the performances of the prompting methods with variation of number of shots and prompt initialization strategy in Section 4.
- Comparison of GFlops for the proposed GOPRO with the SOTA methods i.e. CoOp [22], CoCoOp [21], MaPLe [8] and STYLIP [1], discussed in Section 5.

2 Dataset Description

We evaluate GOPRO over 11 benchmark dataset for base-to-new and cross-dataset generalization, which are described as follows: (1) **ImageNet** [10] - Itconsists of approximately 1.2 million labeled images, from 1,000 different object categories, with high-quality and diverse collection of samples. It covers a wide range of object categories, including animals, plants, vehicles, everyday objects, and various concepts. (2) **Caltech101** [4] - It comprises images from 101 object categories, having 40 to 800 images with resolution of 300×200 pixels in each class. The images are of varying sizes and resolutions, capturing different viewpoints, lighting conditions, and object variations. (3) **OxfordPets** [14] - It contains images from 37

different pet breeds, with each breed having a varying number of images. The total number of images in this dataset is around 7,000. The images are typically high-resolution and showcase various poses and backgrounds. (4) **StanfordCars** [9] - It contains images of cars from 196 different classes or car models, ranging from sedans and SUVs to sports cars and luxury vehicles. Each class typically consists of around 100 to 200 images, resulting in a total of approximately 16,185 images. (5) **Flowers102** [13] - It contains images from various flower species, including roses, sunflowers, daisies, tulips, and many more, depicting 102 different categories. Each category typically contains around 40 to 258 images, resulting in a total of approximately 8,189 images in the dataset. (6) **Food101** [2] - It contains images from 101 different food categories, including dishes like sushi, pizza, burgers, desserts, and many more. Each category typically contains around 1,000 images, resulting in a total of approximately 101,000 images in the dataset. The images showcase different cuisines, cooking styles, and food presentations. (7) **FGVCAircraft** [11] - It comprises images from 100 different aircraft classes, covering a wide range of aircraft types including airplanes, helicopters, and drones. Each class typically contains around 100 to 800 images, resulting in a total of approximately 10,200 images in the dataset. The images showcase different viewpoints, lighting conditions, and variations within each aircraft class. (8) **SUN397** [20] - It contains images from 397 different scene categories, encompassing a wide range of indoor and outdoor scenes such as bedrooms, offices, forests, beaches, and city streets. Each category typically contains around 100 to 500 images, resulting in a total of approximately 108,754 images in the dataset. The images showcase different perspectives, lighting conditions, and variations within each scene category. (9) **UCF101** [17] - It contains middle frames of videos from 101 different action categories, covering a wide range of activities such as sports, dance, martial arts, playing musical instruments, and more. Each category typically contains around 100 to 180 frames, resulting in a total of approximately 13,320 images in the dataset. These are captured in different viewpoints, lighting conditions, and variations within each action category. (10) **DTD** [3] - It contains images from 47 different texture classes, including textures such as fabric, wood, metal, brick, and many more. Each class typically contains around 120 to 180 images, resulting in a total of approximately 5,640 images in the dataset. The images capture different scales, lighting conditions, and variations within each texture class., and (11) **EuroSAT** [5] - It contains images captured by the Sentinel-2 satellite, covering different regions of Europe. It consists of 10 different land use and land cover classes, including urban areas, agricultural land, forests, meadows, and more. Each class contains approximately 2,000 high-resolution images, resulting in a total of around 27,000 images in the dataset.

For single-source multi-targe (SSMT) domain generalization, four variants of ImageNet are used. (1) **ImageNetV2** [16] - It contains of 10000 images, 10 images for each of the 1000 ImageNet classes . (2) **ImageNet-Sketch** [19] - It consists of approx 50000 sketch images from 1000 ImageNet categories. (3) **ImageNet-A** [7] - It contains images from 200 ImageNet categories including real-world, unmodified, and naturally occurring samples, with total number of 7,500 images. (4) **ImageNet-R** [6] - It encompasses various artistic renditions and interpretations of ImageNet classes, including artwork, cartoons, deviantart creations, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video game representations. It focuses on 200 specific ImageNet classes and consists of a total of 30,000 images. These images showcase the diverse ways in which the classes can be visually depicted in different artistic mediums.

.

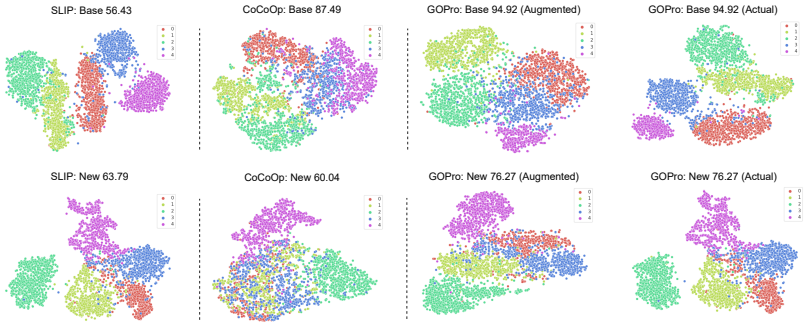

Figure 1: The t-SNE visualizations of visual embeddings from SLIP, CoCoOp and our proposed GOPRO, on the base and new classes of Eurosat dataset. GOPRO archives better discriminativeness.

3 t-SNE Visualization

We have shown a detailed t-SNE [18] visualization of the image embeddings in Figure 1, generated by the visual features of the original and augmented images in both of the base and new classes for the B2N generalization task. . We take SLIP [12] and CoCoOp [21] for comparison on the EuroSAT dataset. The visualization shows that the clusterings are better in base classes, rather than new classes. However, GOPRO gives better clustering of each class, while the cluster points of many classes get overlapped in CoCoOp.

4 Additional Results

Base-to-New (B2N) class generalization: In Table 1, we have shown the detailed results of our proposed GOPRO and other prompting techniques on 11 datasets for B2N generalization task. GOPRO is very much successful to beat others on each and every datasets while considering the harmonic mean (HM) of base and new classes. It is important to notice that GOPRO has shown significant performance in one of the most fine-grained dataset FGVC-Aircraft and outperforms others by at least of 0.38% of margin.

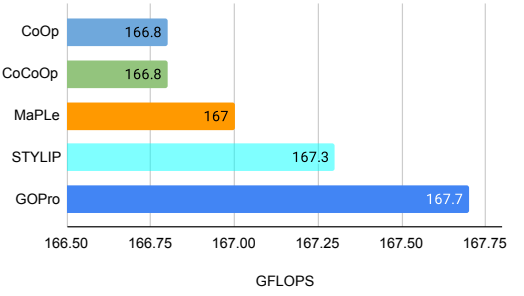

Figure 2: Comparison of the computational complexity of GOPRO among different prompting methods in terms of GFLOPS.

Table 1: Comparison with state-of-the-art methods on base-to-new generalization. GOPRO shows better generalization performance over existing methods on 11 different recognition datasets on 16-shots with context length, $\mathcal{M}$=4. HM represents the harmonic mean.

(a) **Average over 11 datasets**

Method	Base	New	HM
CLIP [ICML]	69.34	74.22	71.70
SLIP [ECCV]	69.77	74.28	71.96
CoOp [IJCV]	82.69	63.22	71.66
CoCoOp [CVPR]	80.47	71.69	75.83
MaPLe [CVPR]	82.28	75.14	78.55
STYLIP [ICCV]	83.22	75.94	79.41
GOPRO	**84.21**	**77.32**	**80.62**

(b) ImageNet

Method	Base	New	HM
CLIP [ICML]	72.43	68.14	70.22
SLIP [ECCV]	72.95	69.76	71.32
CoOp [IJCV]	76.47	67.88	71.92
CoCoOp [CVPR]	75.98	70.43	73.10
MaPLe [CVPR]	76.66	70.54	73.47
STYLIP [ICCV]	77.15	71.34	74.13
GOPRO	**78.56**	**73.22**	**75.80**

(c) Caltech101

Method	Base	New	HM
CLIP [ICML]	96.84	94.00	95.40
SLIP [ECCV]	96.97	94.05	95.49
CoOp [IJCV]	98.00	89.81	93.73
CoCoOp [CVPR]	97.96	93.81	95.84
MaPLe [CVPR]	97.74	94.36	96.02
STYLIP [ICCV]	98.23	94.91	96.54
GOPRO	**98.86**	**95.78**	**97.30**

(d) OxfordPets

Method	Base	New	HM
CLIP [ICML]	91.17	97.26	94.12
SLIP [ECCV]	91.23	97.04	94.05
CoOp [IJCV]	93.67	95.29	94.47
CoCoOp [CVPR]	95.20	97.69	96.43
MaPLe [CVPR]	95.43	97.76	96.58
STYLIP [ICCV]	95.96	98.14	97.04
GOPRO	**96.36**	**98.49**	**97.41**

(e) StanfordCars

Method	Base	New	HM
CLIP [ICML]	63.37	74.89	68.65
SLIP [ECCV]	63.52	74.92	68.75
CoOp [IJCV]	**78.12**	60.40	68.13
CoCoOp [CVPR]	70.49	73.59	72.01
MaPLe [CVPR]	72.94	74.00	73.47
STYLIP [ICCV]	75.19	74.46	74.82
GOPRO	77.59	**75.35**	**76.45**

(f) Flowers102

Method	Base	New	HM
CLIP [ICML]	72.08	77.80	74.83
SLIP [ECCV]	72.17	**77.87**	74.91
CoOp [IJCV]	97.60	59.67	74.06
CoCoOp [CVPR]	94.87	71.15	81.71
MaPLe [CVPR]	95.92	72.46	82.56
STYLIP [ICCV]	96.54	73.08	83.19
GOPRO	**97.73**	77.91	**86.70**

(g) Food101

Method	Base	New	HM
CLIP [ICML]	90.10	91.22	90.66
SLIP [ECCV]	90.14	91.27	90.70
CoOp [IJCV]	88.33	82.26	85.19
CoCoOp [CVPR]	90.70	91.29	90.99
MaPLe [CVPR]	90.71	92.05	91.38
STYLIP [ICCV]	91.20	92.48	91.84
GOPRO	**92.37**	**93.56**	**92.96**

(h) FGVCAircraft

Method	Base	New	HM
CLIP [ICML]	27.19	36.29	31.09
SLIP [ECCV]	27.49	36.11	31.22
CoOp [IJCV]	40.44	22.30	28.75
CoCoOp [CVPR]	33.41	23.71	27.74
MaPLe [CVPR]	37.44	35.61	36.50
STYLIP [ICCV]	37.65	35.93	36.77
GOPRO	**37.89**	**36.44**	**37.15**

(i) SUN397

Method	Base	New	HM
CLIP [ICML]	69.36	75.35	72.23
SLIP [ECCV]	69.35	75.39	72.24
CoOp [IJCV]	80.60	65.89	72.51
CoCoOp [CVPR]	79.74	76.86	78.27
MaPLe [CVPR]	80.82	78.70	79.75
STYLIP [ICCV]	**82.12**	79.95	81.02
GOPRO	81.94	**81.64**	**81.79**

(j) DTD

Method	Base	New	HM
CLIP [ICML]	53.24	59.90	56.37
SLIP [ECCV]	56.71	59.30	57.98
CoOp [IJCV]	79.44	41.18	54.24
CoCoOp [CVPR]	77.01	56.00	64.85
MaPLe [CVPR]	80.36	59.18	68.16
STYLIP [ICCV]	81.57	61.72	70.27
GOPRO	**82.41**	**62.95**	**71.38**

(k) EuroSAT

Method	Base	New	HM
CLIP [ICML]	56.48	64.05	60.03
SLIP [ECCV]	56.43	63.79	59.88
CoOp [IJCV]	92.19	54.74	68.69
CoCoOp [CVPR]	87.49	60.04	71.21
MaPLe [CVPR]	94.07	73.23	82.35
STYLIP [ICCV]	94.61	74.06	83.08
GOPRO	**94.92**	**76.27**	**84.58**

(l) UCF101

Method	Base	New	HM
CLIP [ICML]	70.53	77.50	73.85
SLIP [ECCV]	70.55	77.56	73.89
CoOp [IJCV]	84.69	56.05	67.46
CoCoOp [CVPR]	82.33	73.45	77.64
MaPLe [CVPR]	83.00	78.66	80.77
STYLIP [ICCV]	85.19	**79.22**	82.10
GOPRO	**87.67**	78.91	**83.06**

Sensitivity to the variation in the number of shots: We evaluate the performance of our proposed GOPRO on the base-to-new class generalization task, varying the number of shots from 1 to 16 and taking all training samples from every base class. Table 2 compares our results with state-of-the-art prompting techniques. For this evaluation, we use a context length

($\mathcal{M}$) of 4 and ViT-B/16 as the visual feature backbone, by placing the class token at the end and utilizing a unified context vector. Since CLIP is a zero-shot approach, we exclude it and focus on few-shot-based prompting methods. We present our results as harmonic mean (HM) of base and new classes on average over 11 datasets. Our GOPRO consistently outperforms benchmark prompt learning-based methods by a minimum margin of 0.2%, 0.5%, 2.3%, 1.2% and 1.2% for 1, 2, 4, 8, 16 shots and all training samples, respectively.

Table 2: Comparison of GOPRO with state-of-the-art methods on varying the number of shots for the B2N class generalization task on average of 11 datasets. We choose harmonic mean (H) of base and new classes for comparison, as well as to depict the generalization trade-off.

Method	1-shot	2-shot	4-shot	8-shot	16-shot	All
CoOp	67.14	67.32	68.28	69.33	71.66	0.36in71.89
CoCoOp	70.67	71.94	72.45	74.22	0.36in75.83	75.36
STYLIP	73.98	74.46	75.57	78.86	0.36in79.41	79.25
GOPRO	74.17	74.95	77.84	79.31	**80.62**	80.48

Sensitivity to the prompt initialization strategy: In Table 3, we examine the effectiveness of three distinct prompt initialization strategies for single-source multi-target (SSMT) domain generalization. The results emphasize that manual initialization using `"a photo of a"` surpasses random initialization and no initialization strategies significantly for all the target datasets, except ImageNetV2 and ImageNet-R. However, manual initialization outperforms other strategies in the evaluation of the source domain i.e. ImageNet dataset.

Table 3: Comparison of GOPRO with the prompt benchmark methods for domain generalization across datasets. We train the model on ImageNet using 16-shots with CLIP ViT-B/16 and test on 4 other datasets.

Method	**Source**	**Target**			
	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
random initilization	72.89	**66.03**	49.34	51.96	**78.24**
without initialization	71.56	63.44	49.10	48.62	77.92
manual initialization	**73.27**	65.35	**50.36**	**52.23**	78.02

5 Computational Complexity

We run our model on NVIDIA RTX A6000 GPU with 48 GB card. Fig. 2 represents the comparison of computational complexity between different prompting techniques (CoOp [22], CoCoOp [21], MaPLe [8] and STYLIP [1]) in terms of GFLOPS. MaPLe and STYLIP require 0.12% and 0.3% more computational overhead than CoCoOp respectively, whereas GOPRO needs 0.24%, 0.42% and 0.54% more resources than STYLIP, MaPLe and CoCoOp. However, GOPRO outperforms well the state-of-the-art methods on all of the three generalization tasks i.e. base-to-new, cross-dataset and domain generalization by smart margins.

References

[1] Shirsha Bose, Enrico Fini, Ankit Jha, Mainak Singha, Biplab Banerjee, and Elisa Ricci. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. *arXiv preprint arXiv:2302.09251*, 2023.

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

[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.

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

[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*, 12(7):2217–2226, 2019.

[6] 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*, pages 8340–8349, 2021.

[7] 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 (CVPR)*, pages 15262–15271, June 2021.

[8] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022.

[9] 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. doi: 10.1109/ICCVW.2013.77.

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

[11] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. URL http://arxiv.org/abs/1306.5151.

[12] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 529–544. Springer, 2022.

[13] 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. doi: 10.1109/ICVGIP.2008.47.

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

[15] 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.

[16] 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.

[17] 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. URL http://arxiv.org/abs/1212.0402.

[18] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[19] 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.

[20] 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.

[21] 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.

[22] 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.