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] - It consists 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 $300 \times 200$ 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. (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-target (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 16-datasets with context length, $M=4$. HM represents the harmonic mean.

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
(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.

<table>
<thead>
<tr>
<th>Method</th>
<th>1-shot</th>
<th>2-shot</th>
<th>4-shot</th>
<th>8-shot</th>
<th>16-shot</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoOp</td>
<td>67.14</td>
<td>67.32</td>
<td>68.28</td>
<td>69.33</td>
<td>71.66</td>
<td>0.36</td>
</tr>
<tr>
<td>CoCoOp</td>
<td>70.67</td>
<td>71.94</td>
<td>72.45</td>
<td>74.22</td>
<td>0.36in75.83</td>
<td>75.36</td>
</tr>
<tr>
<td>STYLIP</td>
<td>73.98</td>
<td>74.46</td>
<td>75.57</td>
<td>78.86</td>
<td>0.36in79.41</td>
<td>79.25</td>
</tr>
<tr>
<td>GOPRO</td>
<td>74.17</td>
<td>74.95</td>
<td>77.84</td>
<td>79.31</td>
<td>80.62</td>
<td>80.48</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>ImageNetV2</th>
<th>ImageNet-Sketch</th>
<th>ImageNet-A</th>
<th>ImageNet-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>random initialization</td>
<td>72.89</td>
<td>66.03</td>
<td>49.34</td>
<td>51.96</td>
<td>78.24</td>
</tr>
<tr>
<td>without initialization</td>
<td>71.56</td>
<td>63.44</td>
<td>49.10</td>
<td>48.62</td>
<td>77.92</td>
</tr>
<tr>
<td>manual initialization</td>
<td>73.27</td>
<td>65.35</td>
<td>50.36</td>
<td>52.23</td>
<td>78.02</td>
</tr>
</tbody>
</table>

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 Co-CoOp. 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


