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

Abstract

Large-scale foundation models, such as CLIP, have demonstrated remarkable success in visual recognition tasks by embedding images in a semantically rich space. Self-supervised learning (SSL) has also shown promise in improving visual recognition by learning invariant features. However, the combination of CLIP with SSL is found to face challenges due to the multi-task framework that blends CLIP's contrastive loss and SSL's loss, including difficulties with loss weighting and inconsistency among different views of images in CLIP's output space. To overcome these challenges, we propose a prompt learning-based model called GOPro, which is a unified framework that ensures similarity between various augmented views of input images in a shared image-text embedding space, using a pair of learnable image and text projectors atop CLIP, to promote invariance and generalizability. To automatically learn such prompts, we leverage the visual content and style primitives extracted from pre-trained CLIP and adapt them to the target task. In addition to CLIP's cross-domain contrastive loss, we introduce a visual contrastive loss and a novel prompt consistency loss, considering the different views of the images. GOPro is trained end-to-end on all three loss objectives, combining the strengths of CLIP and SSL in a principled manner. Empirical evaluations demonstrate that GOPro outperforms the state-of-the-art prompting techniques on three challenging domain generalization tasks across multiple benchmarks by a significant margin. Code is available at `https://github.com/mainaksingha01/GOPro`.

1 Introduction

Vision-language models (VLMs) or foundational models, such as CLIP [42] and ALIGN [27], have recently shown exceptional performance in downstream tasks with zero-shot and few-shot scenarios, by employing image-text pairs contrastively, supported by additional information from hand-crafted text prompts like `"a photo of a [cls]"`. However, prompt engineering can present challenges, and learnable prompting techniques, such as CoOp [56], CoCoOp [55], and CLIP-Adapter [18], have replaced manual prompts, showing better generalization abilities. For example, CoOp employs learnable parameters to create text prompts, but the generation of prompts under the supervision of visual features [29, 55, 57] has gained increasing attention. Nevertheless, *learning generalizable prompts by* *leveraging the pre-trained vision and text backbones of CLIP is still regarded as an open problem, partially due to their overlooking of various image transformations.*

Representation learning is a common approach that involves pre-training a model on a large image dataset, such as ImageNet [32], using a supervised approach, which has shown significant improvements in various downstream tasks. However, self-supervised learning, an unsupervised method, has gained popularity due to its success in language [16, 33] and recent advancements in vision [30, 54]. The main objective of self-supervised learning is to replace laborious supervised pre-training that relies on human annotation by defining an auxiliary task that guides the model towards learning a better embedding space. Recently, contrastive methods [7, 10] have emerged as a powerful approach to self-supervised pre-training, outperforming more ad-hoc approaches such as zig-saw solver or rotation prediction [23], among others. Typically, contrastive SSL approaches consider a pair of augmentations for the input images and aim to learn identical embeddings for them.

Our aim is to investigate the impact of SSL models in leveraging CLIP for complex class and domain generalization tasks. While this approach is not entirely new, the sole existing model in this regard, SLIP [37], has proposed to combine the vision-language contrastive learning of CLIP [42] with a self-supervision head within a multi-task setup. This approach has shown improved performance and demonstrated that SSL could complement CLIP's objective. However, SLIP's full training of the model from scratch can be resource-intensive, and the use of hand-engineered prompts may not be optimal. Moreover, SLIP does not ensure semantic invariance in the prompt space, which can affect generalization performance. Therefore, in combining CLIP with SSL, we need to carefully consider two critical factors. *Firstly, we should leverage the pre-trained CLIP backbone while introducing a small set of learnable parameters to learn an SSL-influenced joint image-text embedding space. Secondly, we should replace ad-hoc prompts with learnable prompts to increase generalizability and jointly ensure a better alignment of image-text features.*

Our proposed GOPRO: To address the research queries mentioned earlier, we present GOPRO, a comprehensive framework that leverages the advantages of contrastive SSL and pre-trained CLIP to generate domain and class generic prompts while enhancing the invariance of embedding space against various image-level geometric and photometric transformations. Our approach ensures that by learning to generate consistent prompts for different augmentations of the original images, better generalization can be achieved.

To accomplish the goals, we propose the introduction of learnable projectors atop CLIP's frozen vision and text encoders. We refer to the text projector as the meta-network, aligning with established literature [55]. To generate augmented views of input images, we leverage popular models such as MoCo v3 [11], and AugMix [24]. On the other hand, in contrast to existing techniques [55, 56] that initialize the prompt learner with hand-crafted tokens like `a photo of a [CLS]`, we propose to learn prompt distributions per class by exploiting image feature distributions. Our hypothesis is that prompts learned in conjunction with visual features offer superior class generalizability. Additionally, we are interested in differentiating between object content features and style features [34] of images, as demonstrated in [4]. However, unlike [4], which suggests learning individual tokens from style features extracted from each layer of CLIP's vision encoder, we propose concatenating content and style information and utilizing a text projection network to learn prompt token embeddings, as this approach is found to be computationally more efficient.

The projectors are meticulously trained with three primary loss objectives to ensure optimal performance. Firstly, we employ a contrastive loss [8] between MoCo v3 augmentations of the input image, enhancing the invariance of the visual projector. Secondly, we fine-tune the visual and text projectors using a contrastive loss applied to the image-prompt embeddings. Lastly, we introduce a consistency loss that compares the prompt embeddings obtained from the actual input image with those of the augmented views. To improve robustness and optimize uncertainty of the shared embedding space, we consider the augmented views of MoCo v3 along with AugMix, as it is found to boost the classification performance [24] given its composition-based more diverse set of image synthesis capabilities. However, while [24] focuses on visual feature consistency with AugMix synthesized images, we propose to leverage AuxMix to enforce semantic consistency together with the weak augmentations from MoCo v3 at the prompt space of CLIP, as the final classification is to be carried out there. We highlight our **major contributions** as,

[-] In this paper, we strategically enhance CLIP's prompt learning by using an SSL objective together with the notion of disentangled image-domain-conditioned prompt learning.

[-] Our key contributions involve updating newly-introduced light-weight vision and text projectors atop frozen CLIP using a combination of visual-space SSL contrastive loss, CLIP's image-text contrastive loss, and a novel prompt consistency loss that takes into account the various views of the images. Furthermore, we propose learning the prompt distributions leveraging the multi-scale visual content and style information extracted from CLIP.

[-] To evaluate the effectiveness of our proposed approach, we conduct extensive experiments across three different settings, including base-to-new class generalization, cross-dataset transfer, and single-source multi-target domain generalization on multiple benchmark datasets (as described in Sec. 4). Our GOPRO method significantly outperforms other state-of-the-art comprehensively in all the cases.

2 Related Works

Vision-language models and prompt learning: In general, multimodal learning has been shown to yield better feature learning than unimodal setups. Tasks such as image captioning [50], image retrieval [3], and visual question answering (VQA) [1] typically require joint visual-semantic supervision. Moving forward, VLMs such as CLIP [42] have recently gained significant attention. VLMs are trained on large-scale image-text pairs in a contrastive manner to align the visual and textual embeddings. VLMs efficiently transfer the learned vision information via prompt-based zero-shot and few-shot downstream tasks.

Prompt learning is a widely used approach in NLP [41] and has recently been applied to visual recognition tasks. The primary aim is to leverage pre-trained language models like BERT [14] to provide useful information for downstream tasks through semantically meaningful textual prompts. In recent years, research has focused on automating prompt generation to eliminate manual intervention. One such method is AutoPrompt [45], which examines tokens with the most significant gradient changes in the label likelihood. Meanwhile, CoOp [56] optimizes prompts by fine-tuning CLIP for few-shot image classification. CoCoOp [55] suggests learning conditional prompts based on image features, which can improve CoOp's generalization capability. Crisply, CoCoOp and ProGrad [57] generate prompts from high-level visual features and optimize the generated context tokens. Besides, prompt distribution learning (PDL) [35] proposes optimizing multiple sets of prompts and APPLeNet [46] has demonstrated significant domain generalization performance using multi-scale features within remote sensing images. On the other hand, AD-CLIP [47] has exhibited notable results in domain adaptation by harnessing visual tokens within the prompt space. While these methods focus on image data, Video Prompt Learning (VPL) [28] proposed leveraging foundation models for video data. Finally, Self-supervised Learning with Inter-modality Prompts (SLIP) [37] proposed supplementing the contrastive learning of CLIP with an SSL objective in a multi-task setup. *However, the SSL objective is applied only to the visual branch and is disjoint from the semantic branch, thus not leveraging the multi-modal aspect of CLIP comprehensively. In contrast, we ensure that the SSL objective improves the learning of both the visual and semantic projectors, resulting in enhanced generalization.*

Our approach to prompt learning differs from the literature [4, 55, 56] in the way we utilize visual information. While we draw inspiration from StyLIP [4] in the idea of disentangling image content and style information, we diverge significantly from [4] in how we leverage the visual information to initialize prompt tokens. In StyLIP, each prompt token is learned solely from style information extracted from a specific layer of CLIP's vision encoder, which limits its ability to handle prompts of varied context lengths. In contrast, our approach combines multi-scale content information with global style information from the final vision encoder layer and subsequently learns prompt tokens through a shared meta-network. This offers more flexibility in the length of prompts, allowing for prompts of different context lengths to be effectively learned and utilized in our approach.

Self-Supervised Learning: Self-supervised learning (SSL) is a technique that aims to learn high-quality visual representations from unlabeled images without additional human supervision. Advancements in SSL have made it possible to narrow the gap between supervised and unsupervised representations, as evaluated in downstream tasks [6, 7, 15, 21]. One popular approach is contrastive learning [20, 52], which aims to embed augmented views of a given image closely in feature space while pushing away other images in the same batch [7] or using a memory bank [21]. Other methods focus on retrieving more informative positive examples during training that exhibit more natural image variation than simple artificial augmentations [2, 39]. Some contrastive variants even report strong performance without negative examples [9, 19]. [13] showed that self-supervised training on ImageNet [32] is still highly effective even when using less than 25% of the unlabeled images during training, outperforming supervised pre-training. *As opposed to these approaches, we are keen on improving the invariance of VLMs by supplementing an SSL task, which is found to be beneficial for domain and class generalization tasks.*

3 Problem Definition and Proposed Methodology

Let $\mathcal{D}_s = \{\mathcal{D}_s^i\}_{i=1}^n$ denote n source domains, each with input data $x^i \in \mathcal{X}^i$ and corresponding label space $y^i \in \mathcal{Y}_{Seen}$. It's important to note that the probability distribution of each domain, $P(\mathcal{D}_s^i)$, may differ for all $i \in 1, \cdots, n$. During training, we use the labels $\mathcal{Y}_{Seen}$ from $\mathcal{D}_s$, while during testing, we use $\mathcal{Y}_{Unseen}$ from a target test domain $\mathcal{D}_t$ with $\mathcal{P}(\mathcal{D}_t) \neq \mathcal{P}(\mathcal{D}_s^i)$, $\forall i \in 1, \cdots, n$. For base-to-new class generalization, we set $\mathcal{Y}_{Seen} \cap \mathcal{Y}_{Unseen} = \emptyset$. In contrast, for domain generalization (DG), we consider single-source DG and assume that the label sets for both domains are identical ($\mathcal{Y}_{Seen} \cap \mathcal{Y}_{Unseen} = \mathcal{Y}_{Seen} \cup \mathcal{Y}_{Unseen}$). Finally, for the across-dataset DG, there may or may not be some overlap between $\mathcal{Y}_{Seen}$ and $\mathcal{Y}_{Unseen}$.

3.1 Explaining the working principles of GOPro

This section provides an elaborate description of the architecture and training methodology of GOPro. Specifically, GOPro utilizes the visual backbone of CLIP, denoted as f_v, and

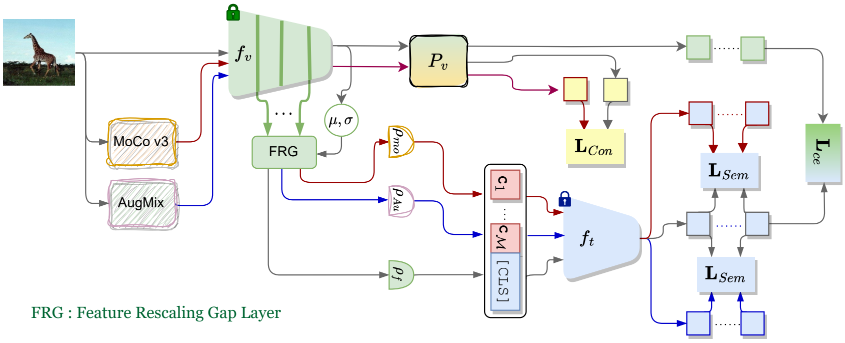

Figure 1: **The design of GOPRO** entails utilizing the fixed image f_v and text encoders f_t of CLIP. In addition, GOPRO incorporates several distinct trainable meta-networks that generate tokens for the original image and augmented images created by MoCo v3 [11] and AugMix [24], denoted as ρ_f, ρ_{mo} and ρ_{Au} respectively. To rescale the features from intermediate layers of f_v, the architecture employs a combination of feature rescaling and the global average pooling (GAP) operation, which we collectively referred to as the FRG layer.

the text encoder, denoted as f_t. Let's assume f_v consists of L encoder layers, and the feature responses for layer $l \in [1, L]$ are denoted as f_v^l. Additionally, we introduce the following learnable units in GOPRO: a vision projector P_v in the visual space, and text projectors ρ_f, ρ_{Au}, and ρ_{mo} for obtaining token embeddings from visual features from the original and the augmented images (MoCo v3 and AuxMix generated), respectively. For simplicity, we consider $\rho = \rho_f = \rho_{Au} = \rho_{mo}$ in our experiments. The goal is to train (ρ, P_v) using image-prompt pairs, such that the learned shared embedding space is generalizable, discriminative, and invariant to image transformations. To achieve this, we propose the following novel considerations: i) A novel prompt learning scheme that leverages multi-scale visual content features extracted from f_v, along with image style information in terms of the instance-wise feature statistics from the L^{th} layer of f_v is introduced. ii) For training P_v using dataset $\mathcal{D}_s$, we deploy the MoCo v3 augmentation technique and use the contrastive formulation of SimCLR [8], following SLIP [37]. iii) To train P_v, we consider the MoCo v3 augmentation x_1 for the original image x and the contrastive loss is realized for (x, x_1). On the other hand, we consider an AugMix-generated image x_2 for x. We then carry out image-text contrastive loss for (x, y), while simultaneously ensuring that $f_t(x, y) \sim f_t(x_1, y) \sim f_t(x_2, y)$.

In this way, we fully leverage the rich representation space of CLIP and smoothly adapt the model for the downstream DG tasks with few training samples.

Image content and style driven prompt generation: In our approach, we aim to generate prompts from the visual features $f_v(x)$ by disentangling the content and style components. To achieve this, we utilize multi-scale content features obtained from different layers of f_v, denoted as $\hat{F}(x) = [\hat{f}_v^1(x); \cdots; \hat{f}_v^L(x)]$ after concatenation. The key idea behind this multi-scale representation is that $\hat{F}(x)$ captures low, mid, and high-level features in its different layers, making it more transferable compared to considering only high-level semantic features [4]. Similarly, we represent the style features using instance-wise feature statistics, namely channel-wise mean and standard deviation, calculated from the L^{th} layer of f_v. Precisely, $\bar{F}(x) = [\vec{\mu}_L(x); \vec{\sigma}_L(x)]$ denotes the style vector. The prompt token initialization for image x is then represented as $F(x) = [\hat{F}(x); \bar{F}(x)]$, rescaled through feature rescaling gap (FRG) layer and it is further adapted to the distribution of $\mathcal{D}$ using ρ. In GOPRO, ρ takes the structure of a single encoder and $\mathcal{M}$ decoders, where $\mathcal{M}$ defines the context length. This way, we learn $\mathcal{M}$ distinct tokens given $F(x)$. The prompt for class y can be derived as: $\mathrm{Pr}_y(x,y) = [c_1; c_2; \cdots; c_{\mathcal{M}}; [CLS_y]]$, where c_m is the output from the m^{th} decoder of ρ, and CLS_y is the semantic embedding for the class y. In the `Supplementary`, we discuss how our approach differs from the existing prompt learning techniques [4, 55, 56].

Visual self-supervised objective: We explore the synergistic combination of the SimCLR SSL objective with augmentations obtained from MoCo v3, as reported in the literature [37], to train P_v. The standard normalized temperature-scaled cross-entropy loss formulation, denoted as $\mathbf{L}_{Con}$ [8], is employed to maximize the `cosine` similarity (δ) between an original image x and its augmented view x_1.

Image-text mapping and prompt consistency objective: Our approach utilizes contrastive learning to map visual and text feature embeddings into a shared embedding space. To compute the class posterior probability of an input x belonging to class y, we employ the following definition, where τ represents the temperature hyper-parameter. We consider the output of P_v as the visual embeddings since this space promotes visual invariance.

$$p(y|x) = \frac{\exp(\delta(P_v(f_v(x)), f_t(\mathrm{Pr}_y(F(x)))/\tau))}{\sum_{k=1}^{|\mathcal{Y}_{Seen}|} \exp(\delta(P_v(f_v(x)), f_t(\mathrm{Pr}_{y_k}(F(x)))/\tau))} \tag{1}$$

Subsequently, the cross-entropy loss ($\mathbf{L}_{ce}$) is computed between the prediction probabilities of each input image and their corresponding class labels as follows:

$$\mathbf{L}_{ce} = \underset{P_v,\rho}{\arg\min} \underset{(x,y)\in\mathcal{P}(\mathcal{D}_s)}{\mathbb{E}} - \sum_{k=1}^{\mathcal{Y}_{Seen}} y_k log(p(y_k|x)_{f_v,f_t}) \tag{2}$$

We emphasize the importance of consistent prompt embeddings obtained from various augmentations applied to the input image. This is crucial in achieving semantic invariance, complementing the visual invariance ensured by $\mathbf{L}_{Con}$. To achieve this, we employ two augmentations per image x: one generated by MoCo v3, which applies geometrical transformations to x, and the other generated by AugMix. AugMix is particularly useful in scenarios where the data distribution encountered during deployment may differ from the training distribution, such as when images are captured with different cameras. AuxMix has been demonstrated to significantly improve generalization performance without necessitating changes to the underlying model [24].

Our approach to achieving semantic consistency involves incorporating distillation losses based on an ℓ_2-norm distance measure. Here, the prompt embedding of the original image x serves as the teacher, while the prompt embeddings of the two augmentations, Moco v3 (x_1) and AugMix (x_2), serve as the students. The loss is defined as,

$$\begin{aligned}\mathbf{L}_{Sem} = \underset{P_v,\rho}{\mathrm{argmin}} \underset{\mathcal{P}(\mathcal{D}_s)}{\mathbb{E}} &||f_t(\mathrm{Pr}_y(\rho(F(x)))) - f_t(\mathrm{Pr}_y(\rho(F(x_1))))||_2 \\ &+ ||f_t(\mathrm{Pr}_y(\rho(F(x)))) - f_t(\rho(\mathrm{Pr}_y(F(x_2))))||_2\end{aligned} \tag{3}$$

Total loss for training and inference: We train the model with respect to all the losses mentioned above, where,

$$\mathbf{L}_{Total} = \mathbf{L}_{Sem} + \mathbf{L}_{ce} + \mathbf{L}_{Con} \tag{4}$$

During inference, we generate all the class prompts given $\mathcal{Y}_{Unseen}$ for a given x_t, and the $y \in \mathcal{Y}_{Unseen}$ maximizing $p(y|x_t)$ is selected.

4 Experimental Evaluations

Dataset descriptions: We evaluate GOPRO on 11 image recognition datasets for base-to-new class generalization and cross-dataset transfer, following the procedure described in CoOp [56]. The datasets include ImageNet [32], Caltech101 [17] for generic object classification, OxfordPets [40], StanfordCars [31], Flowers102 [38], Food101 [5], and FGVCAircraft [36] for fine-grained classification, SUN397 [53] for scene recognition, UCF101 [48] for action recognition, DTD [12] for texture classification, and EuroSAT [22] for satellite imagery recognition. For domain generalization experiments, we employ ImageNet as the source dataset and four other ImageNet variants as target datasets, namely ImageNetV2 [43], ImageNet-Sketch [51] - It consists of 50000 images, 50 images for each of the 1000 ImageNet classes, ImageNet-A [26], and ImageNet-R [25].

Architecture Details: ρ is implemented as a two-layer bottleneck network followed by Linear-ReLU-Linear, where the hidden layer is expanded to the number of context tokens. On the other hand, P_v follows a single-layer MLP structure with a batch normalization layer. f_v and f_t are realized using CLIP's pre-trained transformer backbones.

Training and evaluation protocols: To train GOPRO, we utilize the stochastic gradient descent (SGD) optimizer [44] for 50 epochs and apply scheduling to avoid local minima. During training, we employ 16 shots (samples per class) with a batch size of 4 and ViT-B/16 as the image encoder backbone. The text prompts are initialized using `"a photo of a [CLS]"`, indicating a context length, $\mathcal{M}$=4, as per previous literature [55]. We report the average `top-1` accuracy from three runs of the model.

4.1 Comparisons to the state-of-the-art

Baselines & competitors: In our performance evaluation of GOPRO, we compare it to existing methods from the prompting literature using CLIP. As our baselines, we utilize Zero-shot CLIP [42] and the SSL-based SLIP [37] models, respectively. Additionally, we compare our model with prompt learning techniques, such as CoOp [56], CoCoOp [55], MaPLe [29], and STYLIP [4], using ViT-B/16 backbone.

Base-to-New (B2N) class generalization: Table 1 showcases the experimental results for B2N class generalization averaged over 11 fine-grained and coarse-grained datasets. The harmonic mean (H) between the classification accuracies of the Base and Novel classes is computed. To ensure fairness, we randomly and equally divide the datasets into two groups, defining the base and novel classes for training and testing, respectively. Given that GOPRO is a self-supervised prompt learning method, we pay particular attention to SLIP's self-supervised zero-shot approach. Remarkably, GOPRO achieves superior generalization scores, surpassing SLIP by a significant margin of 14.44% on seen classes and 3.04% on unseen classes across all datasets on average. Furthermore, we compare GOPRO with recent context optimization-based methods, demonstrating its superiority over MaPLe and STYLIP by 2.07% and 1.21% on average, respectively. The detailed results

Table 1: Comparison of GOPRO with state-of-the-art methods on B2N generalization on the average metrics over 11 visual recognition datasets. HM represents the harmonic mean.

Method	Base	Novel	HM
CLIP [42]	69.34	74.22	71.70
SLIP [37]	69.77	74.28	71.96
CoOp [56]	82.69	63.22	71.66
CoCoOp [55]	80.47	71.69	75.83
MaPLe [29]	82.28	75.14	78.55
STYLIP [4]	83.22	75.94	79.41
GOPRO	**84.21**	**77.32**	**80.62**

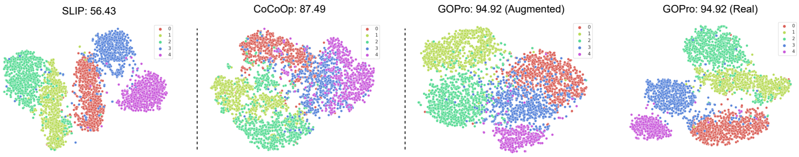

Figure 2: The t-SNE visualizations of visual embeddings from SLIP, CoCoOp and our proposed GOPRO, on the base classes of Eurosat dataset. GOPRO archives better discriminativeness.

are mentioned in the `Supplementary`.

Cross-Dataset (CD) generalization: Table 2 showcases the evaluation results of GOPRO on the CD setup, where the model is trained on the ImageNet dataset (source domain) and zero-shot inference is performed on the remaining ten datasets (target domains). Remarkably, GOPRO surpasses the target classification performance of CLIP and SLIP by substantial margins of 2.63% and 2.55%, respectively. Moreover, GOPRO outperforms MaPLe and STYLIP by 1.56% and 0.48%, respectively, on average. These results demonstrate that GOPRO effectively mitigates the generalization gap for diverse domains and classes.

Table 2: Comparison of GOPRO with the prompt benchmark methods for generalization across datasets. We train the model on ImageNet using 16-shots with CLIP ViT-B/16 and test on 10 other datasets.

Method	Source	Target									
	ImgNet.	C101	Pets	Cars	Flowers	Food	Aircraft	Sun397	DTD	EuroSAT	UCF101	Average
CLIP [42]	66.73	93.31	89.10	65.64	70.73	85.86	24.72	62.58	44.39	48.28	67.72	65.23
SLIP [37]	68.01	93.52	89.23	65.42	70.55	85.92	25.04	62.74	44.16	48.61	67.93	65.31
CoOp [56]	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp [55]	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe [20]	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
STYLIP [4]	72.30	**95.45**	91.60	67.09	72.36	**88.60**	25.21	68.11	47.86	48.22	69.30	67.38
GOPRO	**73.27**	94.81	**92.73**	**68.67**	**72.60**	87.74	**25.85**	**68.70**	**48.04**	**49.43**	**69.98**	**67.86**

Domain generalization (DG): We have conducted experiments to evaluate the generalization performance of GOPRO on a single-source multi-target (SSMT) DG setup. Unlike the CD setting discussed earlier, we only consider the common classes across all datasets, as SSMT is a closed-set setting. The model is trained on the ImageNet dataset and evaluated on its domain variant datasets. Comparison results with state-of-the-art (SOTA) methods and GOPRO are presented in Table 3.

The results demonstrate that GOPRO has outperformed all competitors in the source domain, with a minimum margin of 0.97%. Additionally, GOPRO outperforms other methods in the target domains as well, with minimum margins of 1.07%, 1.09%, and 1.49% in ImageNetV2, ImageNet-A, and ImageNet-R, respectively, except for ImageNet-Sketch, where STYLIP achieves the best performance.

4.2 Ablation Analysis

t-SNE visualization: We present a t-SNE [49] visualization of the image embeddings in Figure 2, generated by the visual features of the original and augmented images. We compare them with SLIP [37] and CoCoOp [55] on the EuroSAT dataset for the B2N generalization

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
CLIP [42]	66.73	60.83	46.15	47.77	73.96
SLIP [37]	68.01	61.12	46.35	47.54	73.88
CoOp [56]	71.51	64.20	47.99	49.71	75.21
CoCoOp [55]	71.02	64.07	48.75	50.63	76.18
MaPLe [29]	70.72	64.07	49.15	50.90	76.98
STYLIP [4]	72.30	64.28	**50.83**	51.14	76.53
GOPRO	**73.27**	**65.35**	50.36	**52.23**	**78.02**

task. The visualization clearly demonstrates that GOPRO has better clustering of each class, while the cluster points of many classes get overlapped in CoCoOp.

Sensitivity to context length for B2N generalization: We have conducted tests on GOPRO using varying prompt tokens ($\mathcal{M}$) ranging from 1 to 16. Instead of manual prompt initialization, we randomly initialize prompts from the context. In Figure 3, we present the average performance of GOPRO on 11 datasets in the B2N generalization task with different context lengths. The results indicate that GOPRO performs well on base classes with eight tokens and new classes with 11 tokens. However, for better overall generalization, it performs best with four tokens, considering the harmonic mean of both. Interestingly, GOPRO achieves almost the same accuracy as random initialization with a context length of 4 in manual initialization of "a photo of a", as shown in Table 1.

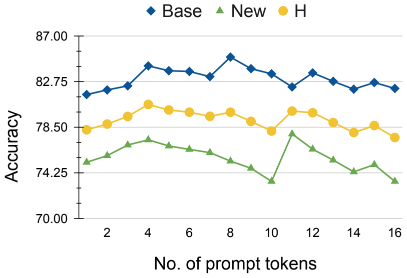

Figure 3: Comparison of results of GOPRO with different numbers of prompt tokens in B2N generalization setup.

Table 4: Ablation study of GOPRO with different losses in B2N generalization setup.

Loss	Base	Novel	HM
$\mathbf{L}_{ce}$	81.34	72.16	76.48
$\mathbf{L}_{ce} + \mathbf{L}_{Con}$	82.15	75.02	78.42
$\mathbf{L}_{ce} + \mathbf{L}_{Sem}(x_1)$	83.23	74.64	78.70
$\mathbf{L}_{ce} + \mathbf{L}_{Sem}(x_2)$	81.65	73.97	77.62
$\mathbf{L}_{ce} + \mathbf{L}_{Sem}(x_1 + x_2)$	83.87	75.15	79.27
$\mathbf{L}_{ce} + \mathbf{L}_{Sem} + \mathbf{L}_{Con}$	**84.21**	**77.32**	**80.62**

Ablation on the loss terms: We have conducted multiple experiments with our proposed model, GOPRO, using various loss terms, as presented in Table 4. The visual contrastive loss, denoted as $\mathbf{L}_{Con}$, is typically utilized to reduce the difference between two different self-supervised views of augmented image features from MoCo v3. Discarding this loss results in a decrease in performance by almost 1.35%.

Furthermore, the employment of $\mathbf{L}_{Sem}$ enhances the efficacies of the semantic space, leading to an additional improvement in the results by 2.2%. We observe that GOPRO experiences a decline in performance by 1.92% and 3% for single augmentations with MoCo v3 (x_1) and AugMix (x_2), respectively when compared to the full GOPRO model. These findings highlight the importance of both losses, along with $\mathbf{L}_{ce}$, which are responsible for the improved performance of GOPRO [1]

[1] More ablations, model complexity analysis, and visualizations are available in the Supplementary material.

5 Takeaways

In this paper, we present a comprehensive analysis of how self-supervised learning can enhance vision-language models. We propose a novel approach called GOPRO that ensures consistency among the augmented views of input images in both the visual and semantic space of CLIP, using innovative loss functions. Furthermore, we introduce a new prompt learning framework in GOPRO that leverages visual features by disentangling content and style information and incorporates them into prompt learning through a learnable encoder-decoder-based text projector. Our experimental results demonstrate that GOPRO outperforms benchmark prompting methods in three challenging domain generalization tasks involving class, domain, and dataset shifts. Additionally, we are excited to explore the potential of GOPRO for more specific applications, such as medical imaging and remote sensing, among others, in the future.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016.

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021.

[3] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.

[4] 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.

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

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[11] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[12] 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.

[13] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14755–14764, 2022.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[16] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*, 2020.

[17] 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.

[18] 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 adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[20] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[22] 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.

[23] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[24] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[25] 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.

[26] 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.

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[28] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 105–124. Springer, 2022.

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

[30] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1920–1929, 2019.

[31] 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.

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

[33] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[34] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[35] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022.

[36] 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.

[37] 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.

[38] 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.

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

[41] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[42] 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.

[43] 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.

[44] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[45] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[46] Mainak Singha, Ankit Jha, Bhupendra Solanki, Shirsha Bose, and Biplab Banerjee. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[47] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. *arXiv preprint arXiv:2308.05659*, 2023.

[48] 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.

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

[50] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[51] 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.

[52] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3): 1–34, 2020.

[53] 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.

[54] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1476–1485, 2019.

[55] 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.

[56] 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.

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