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.

**Motivation**

- 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.
- We should replace ad-hoc prompts with learnable prompts to increase generalizability and jointly ensure a better alignment of image-text features.

**Contributions**

The present study investigates the following objectives:

- 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. Our GOPro method significantly outperforms other state-of-the-art comprehensively in all the cases.

**Formulation of Metric Objectives**

\[
\begin{align*}
L_{cross} &= \min_{P_{atp}, P_{ctx}} \mathbb{E}_{(x,y)} \sum_{t=1}^{S} |p_t(y|x) - p_t(y|x)^t| \quad (1) \\
L_{sem} &= \min_{P_{atp}, P_{ctx}} \mathbb{E}_{(x,y)} \left(\|p_t(x|y) - p_t(x|y)^t\|_2 + \|p_t(x,y) - p_t(x,y)^t\|_2\right) \quad (2) \\
L_{total} &= L_{sem} + L_{cross} + L_{con} \quad (3)
\end{align*}
\]

**Results and discussion**

Table 2. Comparison of GOPro with the prompt benchmark methods for domain generalization across datasets. We train the model on ImageNet with 16 shots with CLIP-ViT-B/16 and test on 4 other datasets.

Table 3. Ablation study of GOPro with different losses in B2N generalization setup.

**Conclusions**

- 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.
- 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.
- 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] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In Computer Vi-

https://github.com/mainaksingha01/GOPro

British Machine Vision Conference (BMVC) 2023

mainaksingha.iitb@gmail.com