Person Re-Indentification

GNR 697: R&D Project

by

Team Members:

Darshan Kumar (210010023)

Parth Nawkar (200010044)

Raviraj Shelar (210010053)

under the guidance of

Prof. Biplab Banerjee

April 30, 2024

Department of Aerospace Engineering

Indian Institute of Technology, Bombay

Mumbai 400 076

Contents

Abstract

This project centers on state-of-the-art vision-language models driven by prompts, namely CLIP-REID, CoOp, CoCoOp, and Maple, and their application towards person re-identification. The primary focus was on comprehensively understanding these models, implementing them, and assessing their utility. The majority of efforts were dedicated to gaining insights into these models, running them to achieve comparable performance. Ultimately, a custom sample implementation(**link**) of these models was developed.

Chapter 1

Introduction

Image re-identification (ReID) aims to match the same object across different and non-overlapping camera views. Particularly, it focuses on detecting the same person or vehicle in the surveillance camera networks. ReID is a challenging task mainly due to the cluttered background, illumination variations, huge pose changes, or even occlusions.

1.1 CLIP-ReID:

In recent advancements, pre-trained vision-language models like CLIP have showcased remarkable performance across various downstream tasks, including image classification and segmentation. However, their application to fine-grained image re-identification (ReID) tasks poses unique challenges due to the absence of concrete text descriptions for labels. This issue is addressed by proposing a novel two-stage strategy tailored for enhancing visual representations within the context of ReID. By leveraging the cross-modal description capabilities inherent in CLIP, the proposed approach harnesses learnable text tokens to generate ambiguous descriptions associated with each identity. Through a combination of fixed and optimized stages, this strategy refines the image encoder to accurately represent data in feature embeddings, thus demonstrating promising results across multiple person or vehicle

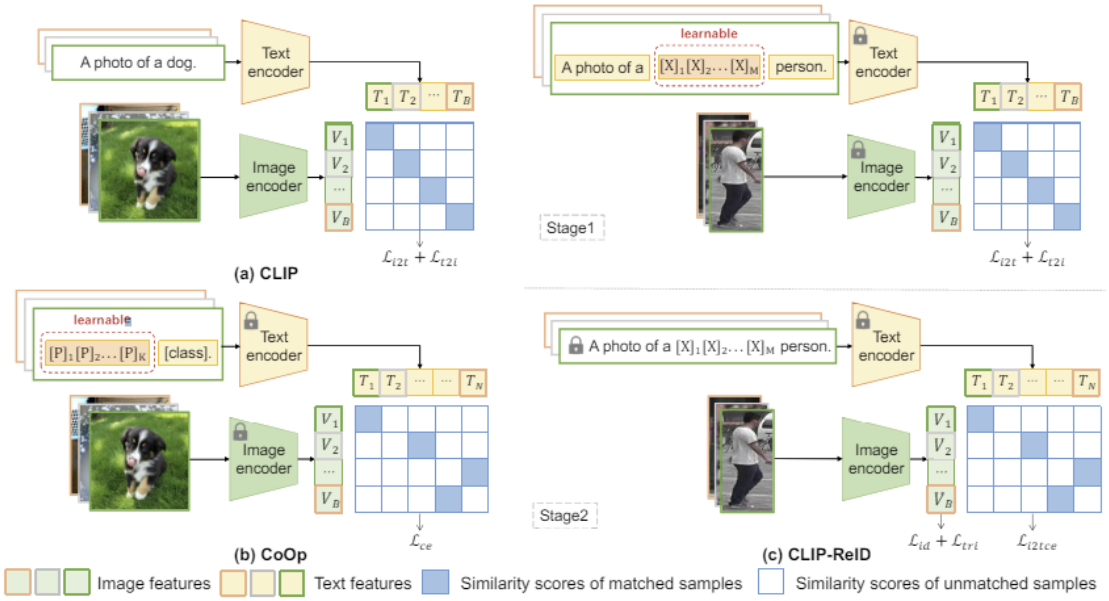
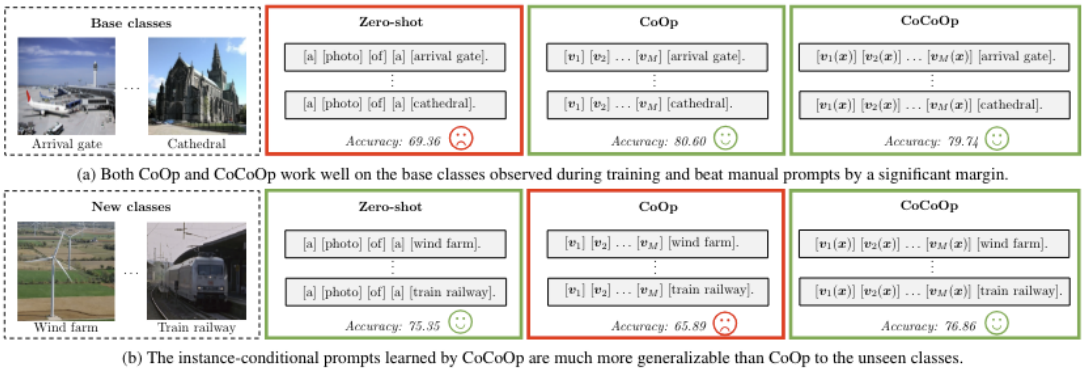


Figure 1.1: : Overview of our approach compared to CLIP and CoOp.

ReID datasets. E.g., given a particular image classification task, the candidate text labels are concrete and can be combined with a prompt, such as "A photo of a", to form the text descriptions. The classification is then realized by comparing image features with text features generated by the text encoder, which takes the text description of categories as input. Note that it is a zero-shot solution without tuning any parameters for downstream tasks but still gives satisfactory results. Based on this, CoOp (Zhou et al. 2021) incorporates a learnable prompt for different tasks. The optimized prompt further improves the performance.

1.2 CoOp

Vision-language models like CLIP utilize prompts to synthesize class-specific weights for classification tasks. Prompt engineering, however, is time-consuming and inefficient, leading researchers to explore prompt learning techniques from NLP for adapting pre-trained models to specific tasks. Context Optimization (CoOp) is one such method that transforms context words in a prompt into learnable vectors,



(a) Both CoOp and CoCoOp work well on the base classes observed during training and beat manual prompts by a significant margin.

(b) The instance-conditional prompts learned by CoCoOp are much more generalizable than CoOp to the unseen classes.

Figure 1.2: Learning generalizable prompts with CoOp and CoCoOp

enabling significant performance boosts with minimal labeled data. Nonetheless, CoOp faces challenges in generalizing its learned context to unseen classes within the same dataset, indicating potential overfitting to training classes.

1.3 CoCoOp

CoCoOp, short for Conditional Context Optimization, addresses the limitations of CoOp by introducing instance-conditional context learning. While CoOp efficiently trains context vectors with minimal labeled data, it struggles with generalizability to unseen classes within the same task. CoCoOp proposes a novel approach where the context is dynamically adapted to each input instance rather than being fixed. This instance-conditional focus aims to reduce overfitting by encompassing the entirety of the task rather than specific class distinctions. Unlike simplistic methods requiring multiple neural networks, CoCoOp employs a parameter-efficient design featuring a Meta-Net. This lightweight neural network generates conditional tokens for each input, which are then combined with context vectors. The resultant prompt, tailored to each instance, facilitates more robust recognition. Through a simple yet effective architecture, CoCoOp achieves improved generalization without exponentially increasing computational costs, marking a significant advancement in vision-language model adaptation.

1.4 MaPLe

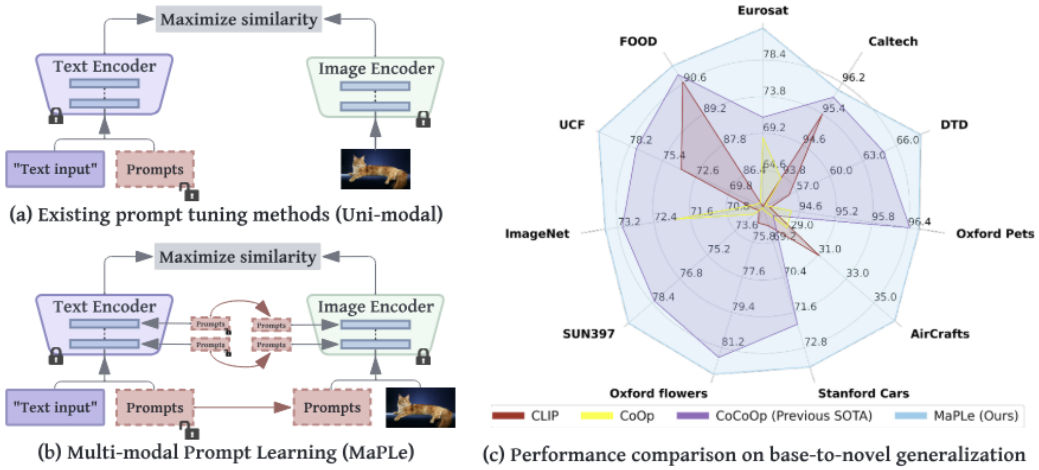


Figure 1.3: Comparison of MaPLe with standard prompt learning methods

Multi-modal Prompt Learning (MaPLe) emerges as a solution to enhance the adaptability and performance of pre-trained vision-language models like CLIP across various downstream tasks. While CLIP demonstrates remarkable generalization capabilities, its reliance on handcrafted text prompts and the challenge of fine-tuning the entire model for specific tasks present significant hurdles. Existing prompt learning approaches primarily focus on adjusting prompts for the text encoder, neglecting the equally important image encoder in CLIP. MaPLe addresses this gap by introducing a comprehensive approach to fine-tune both the text and image encoder representations simultaneously. By leveraging multi-modal prompts, MaPLe ensures optimal alignment between vision and language representations, thereby enhancing model robustness and generalization across diverse tasks. Through extensive experiments spanning base-to-novel generalization, cross-dataset evaluation, and domain generalization, MaPLe outperforms existing methods and demonstrates favorable efficiency during training and inference. Overall, MaPLe represents a significant advancement in prompt learning techniques for pre-trained vision-language models, offering streamlined architectural design and improved performance across a range of tasks.

Chapter 2

Results

2.1 Person ReID using Interaction-And-Aggregation Network [1]

Classification loss	Pairwise loss	MAP
CrossEntropy	Triplet	86.6
CrossEntropy	Contrastive	86.4
CrossEntropy	Cosface	86.2

Table 2.1: Using market-1501 dataset

Classification loss	Pairwise loss	MAP
CrossEntropy	Triplet	57.0
CrossEntropy	Contrastive	56.7
CrossEntropy	Cosface	55.2

Table 2.2: Using MSMT dataset

NOTE:

MAP - Mean Average Precision

ResNet50 is used as the backbone.

2.2 CLIP + CoOp

[4]

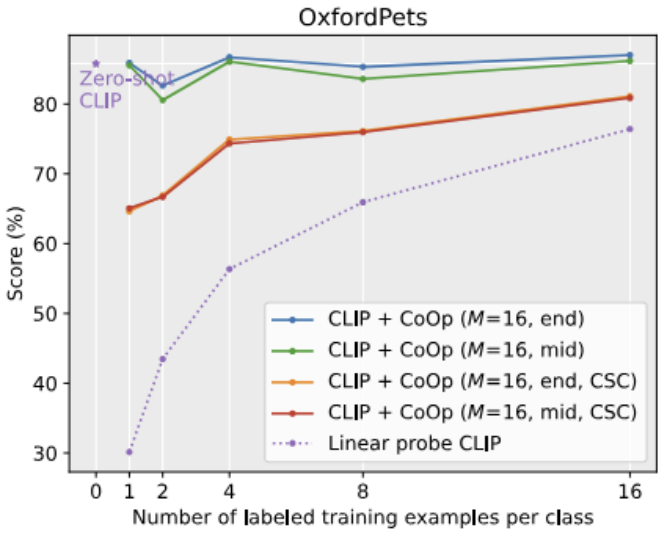


Figure 2.1: Oxford pets

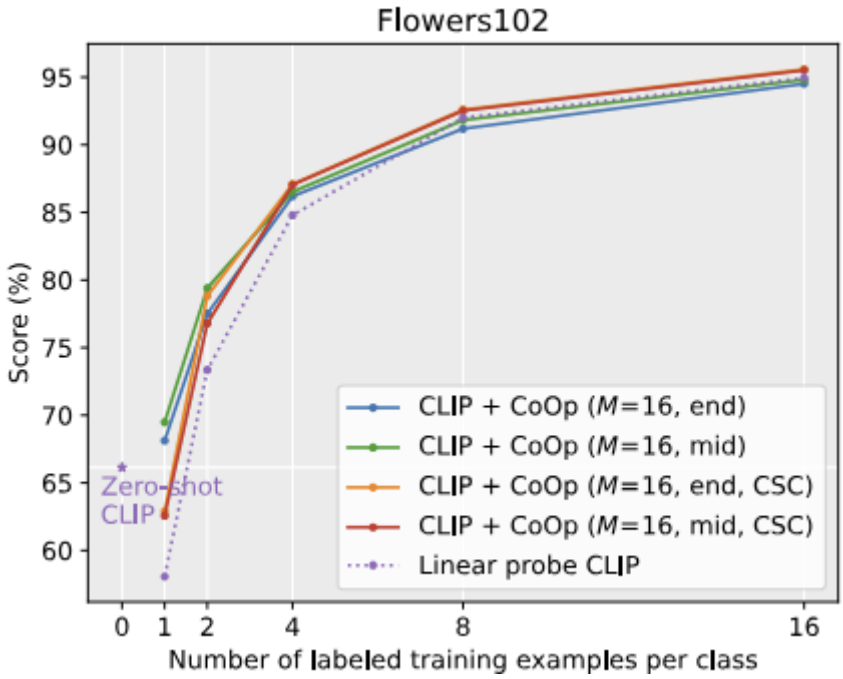


Figure 2.2: Flowers

2.3 CLIP + CoCoOp

[3]

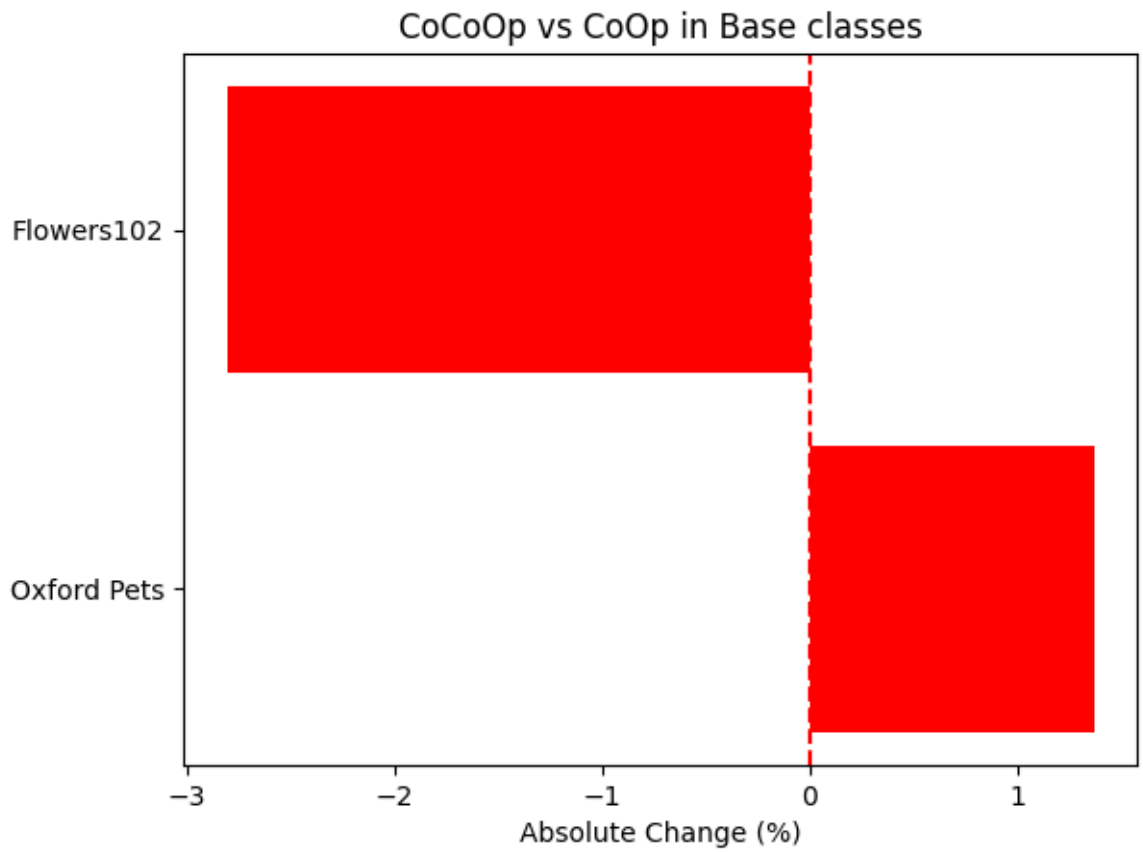


Figure 2.3: Oxford pets

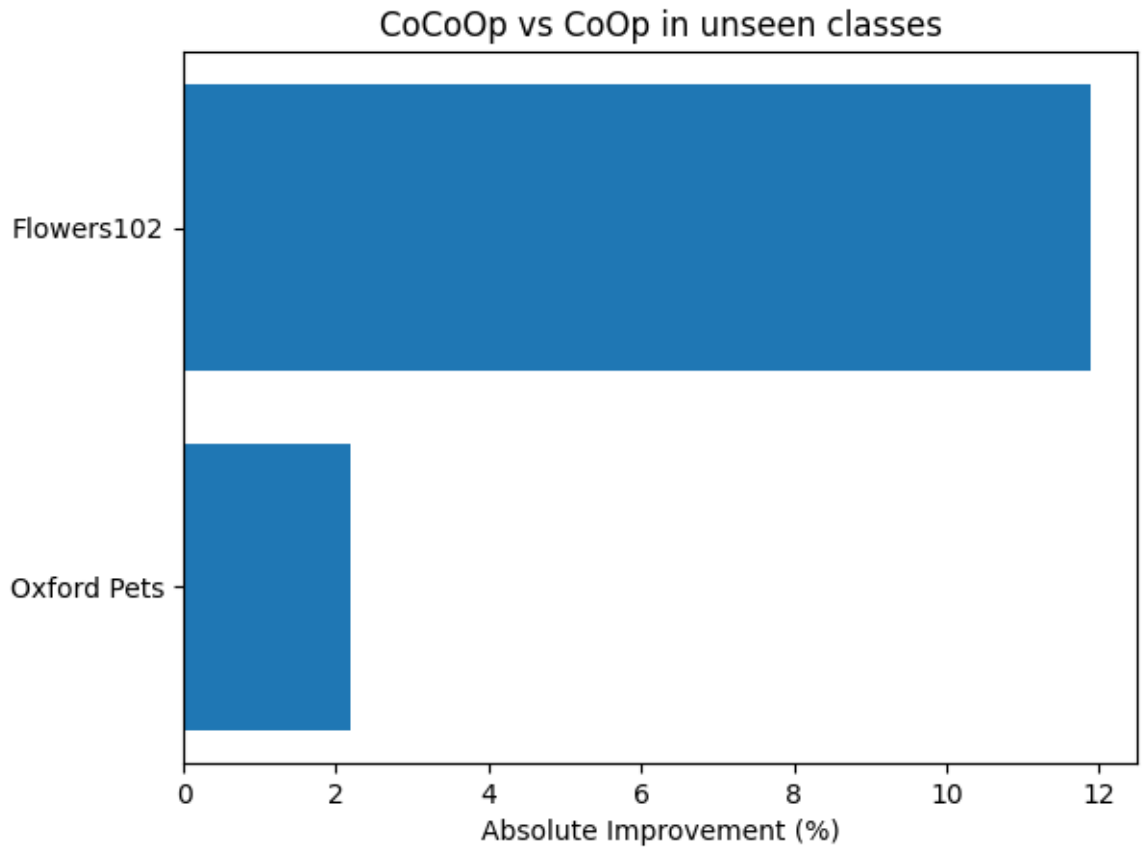


Figure 2.4: Flowers

	ImageNet(Source)	Oxford Pets	Flowers102
CoOp	71.3	89.14	67.32
CoCoOp	70.8	90.01	70.88

Table 2.3: Comparison of prompt learning methods in the cross-dataset transfer setting

Despite maintaining an advantage over CLIP in terms of average performance, CoOp's gains in the base classes are nearly zeroed out by the catastrophic failures in the new classes, highlighting the need to improve generalizability for learning-based prompts. Prompts applied to the 2 target datasets are learned from ImageNet (16 images per class). Clearly, CoCoOp demonstrates better transferability than CoOp.

2.4 MaPLe

[2] [1]

	ImageNet(Source)	Oxford Pets	Flowers102
CoCoOp	70.8	90.01	70.88
MaPLe	70.72	90.32	72.23

Table 2.4: Comparison with state-of-the-art methods on base-to-novel generalization

	ImageNet(Source)	ImageNet V2	ImageNet-S	ImageNet-A	ImageNet-R
CLIP	66.73	60.79	46.2	47.51	73.84
CoOp	71.51	64.2	47.99	49.65	75.3
CoCoOp	71.02	64.07	48.75	50.63	76.18
MaPLe	70.72	64.07	49.15	50.90	76.98

Table 2.5: Comparison with state-of-the-art methods on base-to-novel generalization

	Params	Params(% CLIP)
CoOp	2048	0.002
CoCoOp	35360	0.03
MaPLe	3.55 M	2.85

Table 2.6: Comparison of computational complexity among different prompting methods

In terms of inference speed, Co-CoOp is significantly slower and the FPS (Frames Per Second) remains constant as the batch size increases. In contrast, MaPLe has no such overhead and provides much better inference and training speeds. Further,MaPLe provides better convergence as it requires only half training epochs as compared to Co-CoOp (5 vs 10 epochs). MaPLe adds about 2.85% training parameters on top of CLIP.

Ultimately, a custom sample implementation(**link**) of these models was made in pytorch which calculated the similarity scores.

Chapter 3

Contributions

3.1 Darshan Kumar (210010023)

- **CoCoOp**:
  - Primarily focused on the implementation and experimentation with CoCoOp (Conditional Context Optimization). This involved understanding the theoretical underpinnings of CoCoOp as outlined in the literature and implementing it within the project's framework.
  - Responsible for fine-tuning CoCoOp parameters and evaluating its performance across various datasets and scenarios. This included assessing its generalizability and effectiveness in different re-identification tasks.
- **Sample Pytorch Implementation**:
  - Developed a custom sample implementation of the models using PyTorch. This implementation was utilized to calculate similarity scores and assess the performance of the models in practical scenarios.

3.2 Parth Nawkar (200010044)

- **CLIP**:
  - Took a lead role in understanding and implementing CLIP (Contrastive Language-Image Pre-training). This involved studying the CLIP architecture, its training procedure, and its application in vision-language tasks.
  - Responsible for integrating CLIP into the project pipeline and evaluating its performance in person re-identification tasks. This included experimenting with different prompts and fine-tuning CLIP for optimal performance.
- **CoOp**:
  - Additionally, contributed to the exploration and implementation of CoOp (Context Optimization), a method for adapting pre-trained models to specific tasks. This involved studying CoOp's techniques and integrating them into the project's framework.
- **Sample Pytorch Implementation**:
  - Collaborated in the development of the custom PyTorch implementation. Ensured the implementation accurately represented the functionalities of the models and facilitated the calculation of similarity scores for evaluation purposes.

3.3 Raviraj Shelar (210010053)

- **MaPLe**:
  - Led the efforts related to MaPLe (Multi-modal Prompt Learning), focusing on enhancing the adaptability and performance of pre-trained vision-language models like CLIP.
  - Responsible for studying the MaPLe methodology, implementing it within the project's context, and conducting experiments to evaluate its efficacy across various tasks and datasets.
- **Sample Pytorch Implementation**:
  - Developed a custom sample implementation of the models using PyTorch. This implementation was utilized to calculate similarity scores and assess the performance of the models in practical scenarios.

Bibliography

[1] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019.

[2] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.