Zero-Shot Composed Image Retrieval with Textual Inversion

*Alberto Baldrati, *Lorenzo Agnolucci, Marco Bertini, Alberto Del Bimbo

University of Florence – Media Integration and Communication Center (MICC)

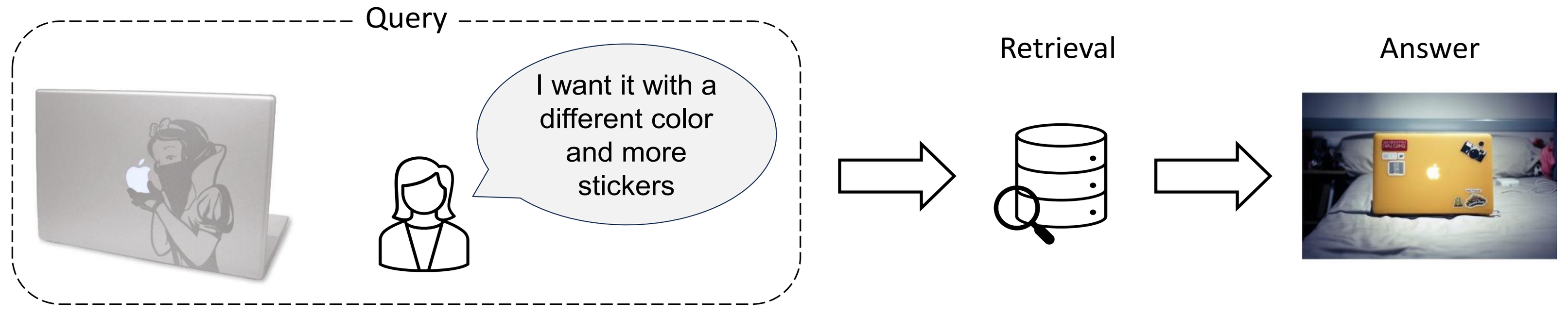
[name].[surname]@unifi.it

MICC Media Integration and Communication Center

ICCV23 PARIS

Introduction

Composed Image Retrieval (CIR)



Current Limitations

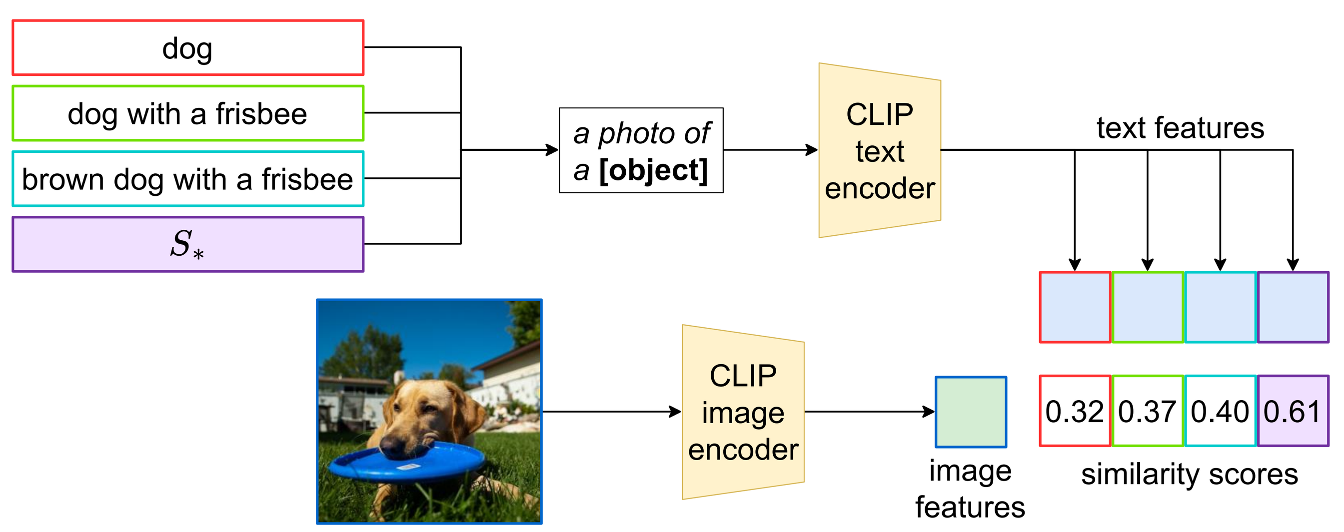
- Existing methods for CIR rely on supervised learning, which requires expensive and time-consuming manual data labeling
- Existing CIR datasets contain several false negatives, *i.e.*, images that could be potential ground truths for the query but are not labeled as such

Contributions

- We propose **SEARLE**, a CLIP-based [3] method that addresses CIR in a zero-shot manner, thus without requiring a labeled training dataset
- We introduce **CIRCO**, an open-domain benchmarking dataset for CIR with multiple annotated ground truths and reduced false negatives

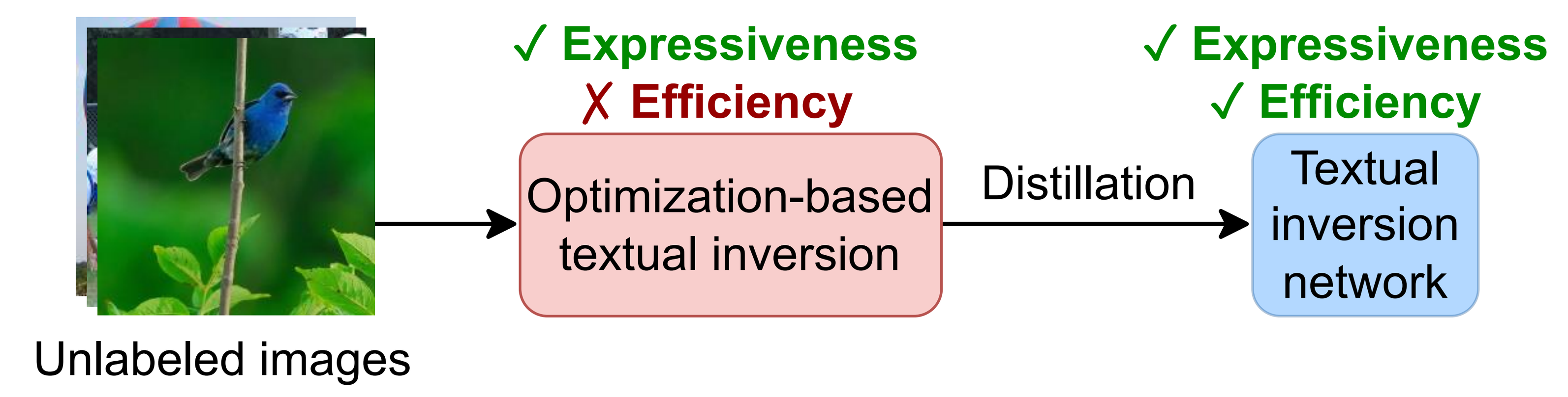
Textual Inversion

- The term **textual inversion** [4] refers to the process of mapping an image into a pseudo-word token residing in the CLIP token embedding space
- Textual inversion consists of expanding CLIP vocabulary by defining a new pseudo-word **S_*** which encapsulates the visual information of the image

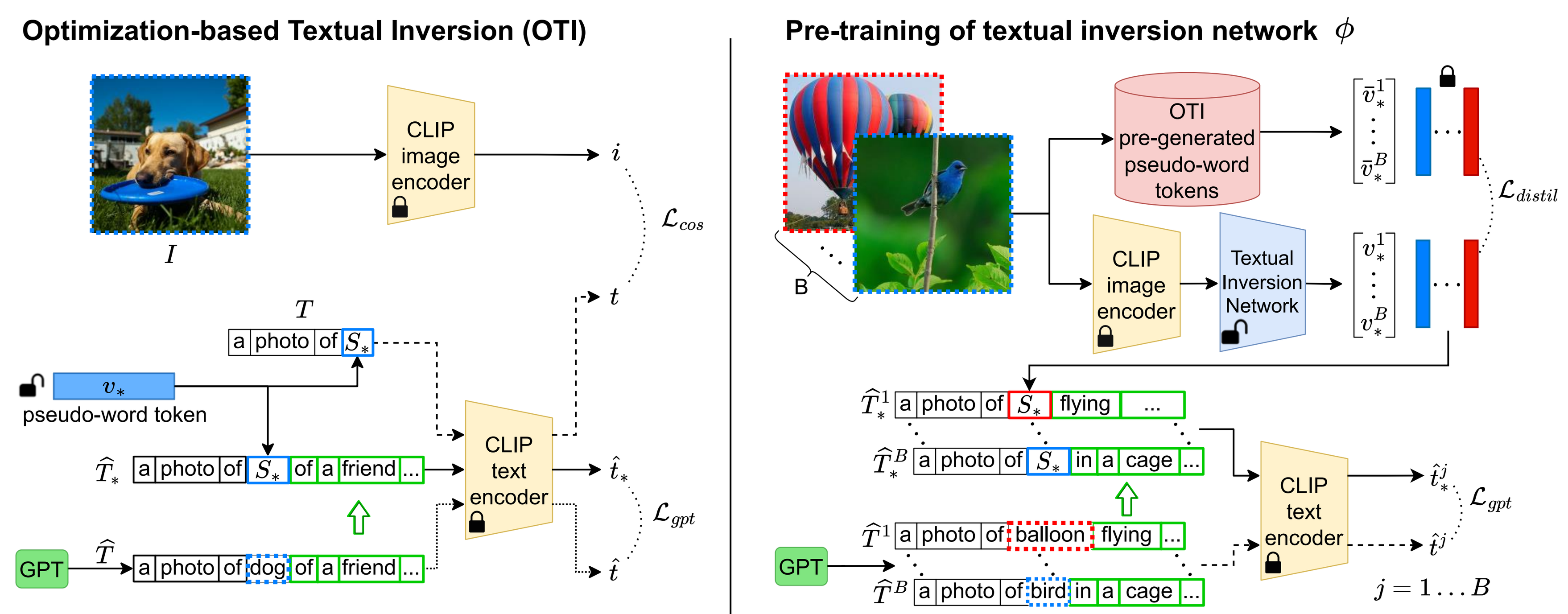


SEARLE Training

Overview



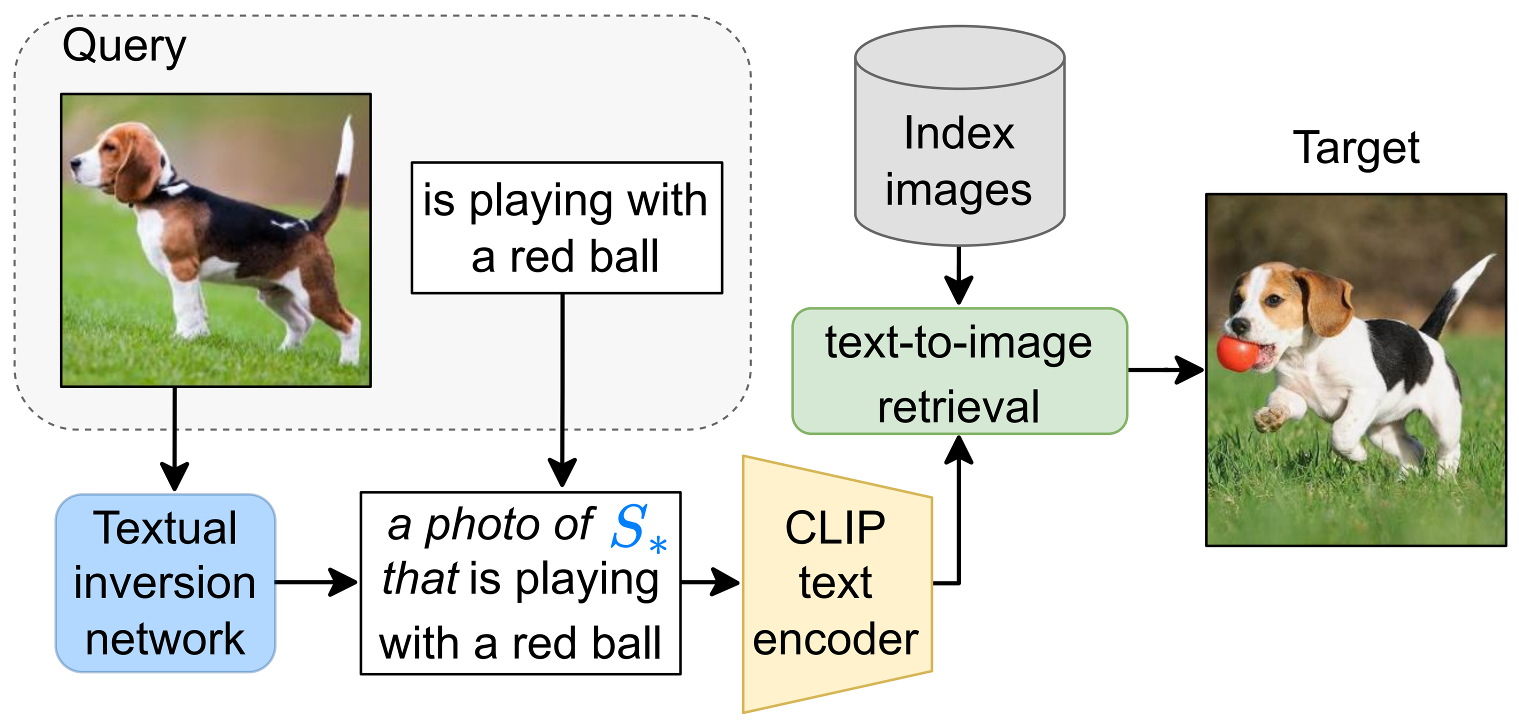
Breakdown



$$\mathcal{L}_{OTI} = \lambda_{cos}\mathcal{L}_{cos} + \lambda_{OTIgpt}\mathcal{L}_{gpt}$$

$$\mathcal{L}_{\phi} = \lambda_{distil}\mathcal{L}_{distil} + \lambda_{\phi gpt}\mathcal{L}_{gpt}$$

SEARLE Inference



References

[1] Cohen, Niv, et al. ""This is my unicorn, Fluffy": Personalizing frozen vision-language representations." ECCV2022

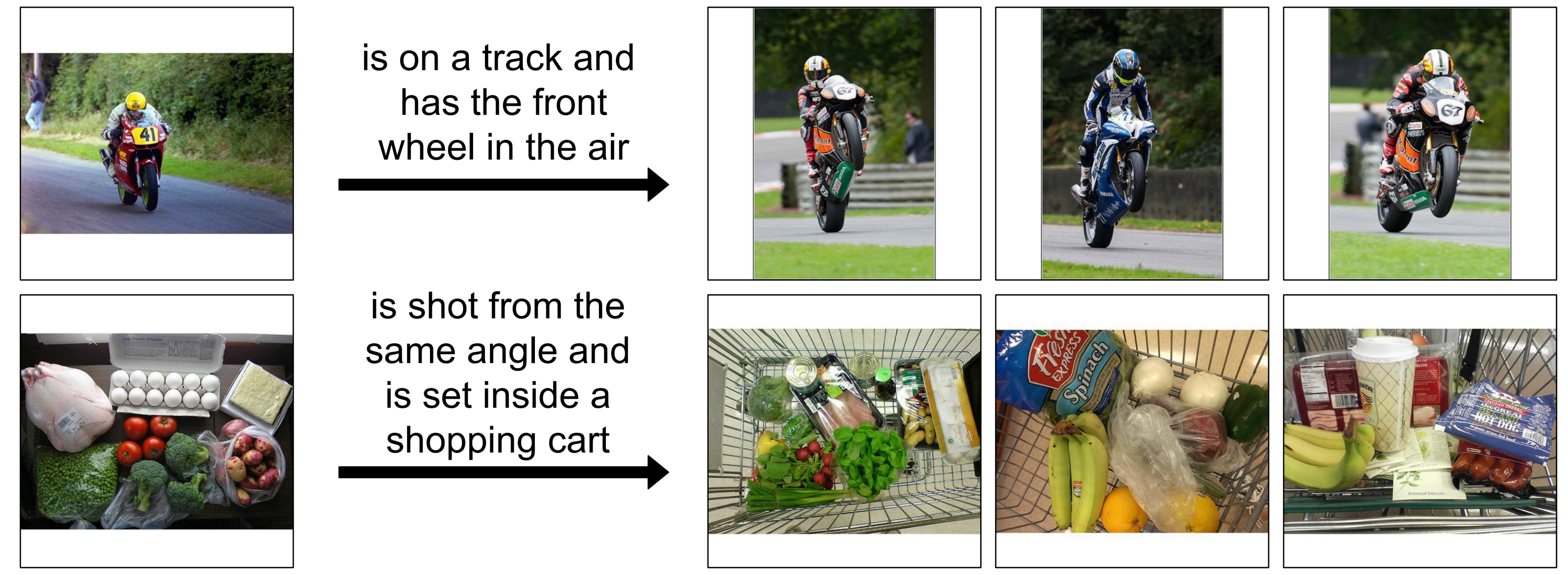
[2] Saito, Kuniaki, et al. "Pic2word: Mapping pictures to words for zero-shot composed image retrieval." CVPR2023

[3] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML2021

[4] Gal, Rinon, et al. "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion." ICLR2022

CIRCO Dataset

CIRCO is the first CIR dataset with multiple annotated ground truths and reduced false negatives



Results

Backbone	Method	CIRR R@1	CIRR R@5	CIRR R@10	FashionIQ R@10	FashionIQ R@50	CIRCO mAP@10
B/32	Image-only	6.89	22.99	33.68	5.90	13.37	1.60
	Text-only	21.81	45.22	57.42	18.70	36.84	2.67
	Image + Text	11.71	35.06	48.94	14.78	29.60	3.25
	Captioning	12.46	35.04	47.71	13.98	28.62	5.77
	PALAVRA [1]	16.62	43.49	58.51	19.76	37.25	5.32
	SEARLE-OTI	**24.27**	53.25	66.10	22.44	42.34	7.83
	SEARLE	24.00	**53.42**	**66.82**	**22.89**	**42.53**	**9.94**
L/14	Pic2Word [2]	23.90	51.70	65.30	24.70	43.70	9.51
	SEARLE-XL-OTI	**24.87**	52.31	**66.29**	**27.61**	**47.90**	11.03
	SEARLE-XL	24.24	**52.48**	**66.29**	25.56	46.23	**12.73**

SEARLE is trained on just **3%** of Pic2Word data

Conclusions

- Existing approaches for CIR are limited by their reliance on expensive labeled datasets
- SEARLE achieves state-of-the-art results on CIR without the need for a labeled training dataset
- We introduce CIRCO, the first CIR dataset with multiple annotated ground-truths and reduced false negatives

arXiv GitHub DEMO