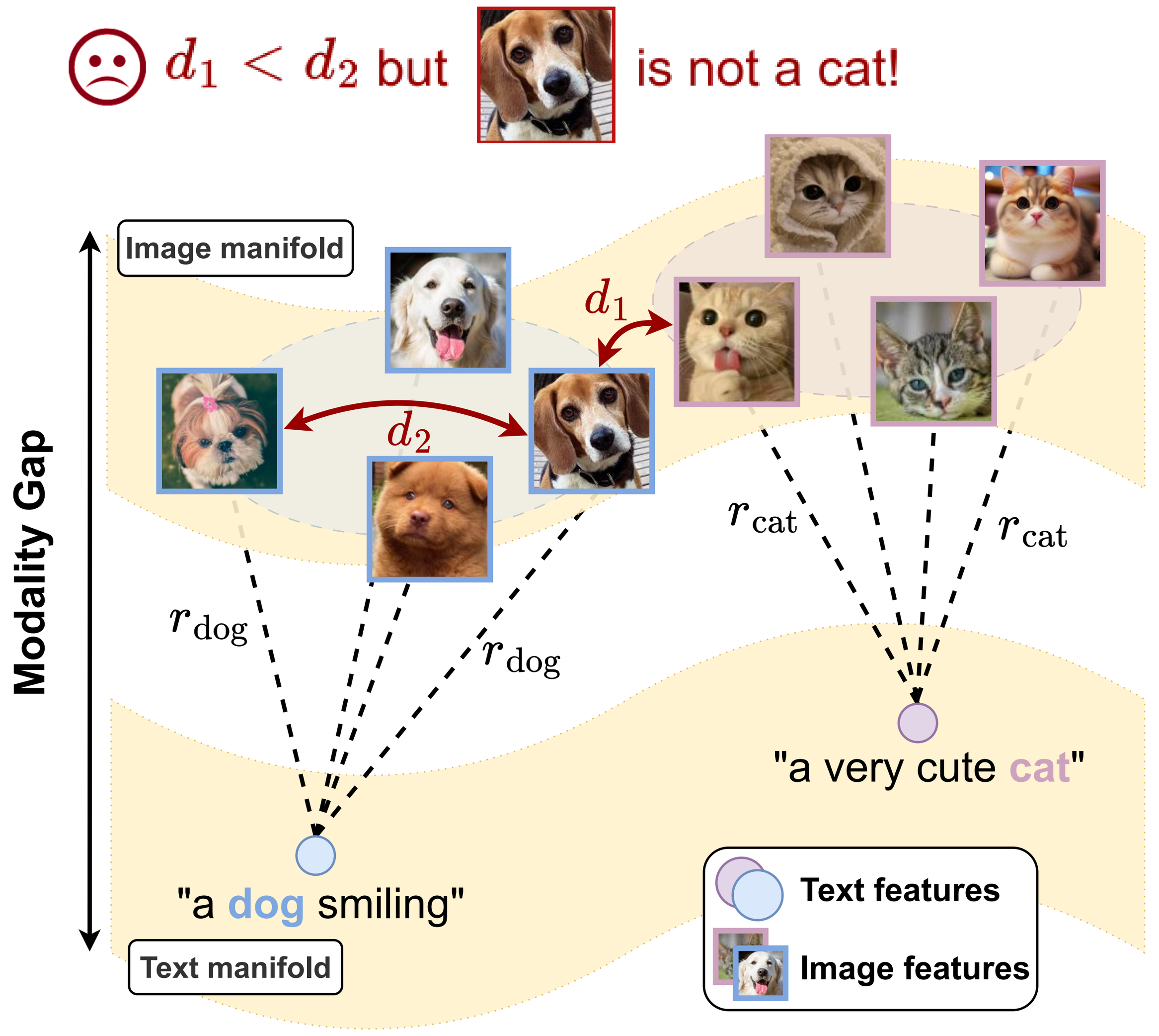
Cross the Gap: Exposing CLIP Intra-modal misalignment Via Modality Inversion

*Marco Mistretta, *Alberto Baldrati, *Lorenzo Agnolucci, Marco Bertini, Andrew D. Bagdanov

Better **STOP** using **CLIP** for image-to-image or text-to-text similarity comparisons. **Intra-modal similarities** *are* suboptimal. Mind the CLIP **intra-modal misalignment!**

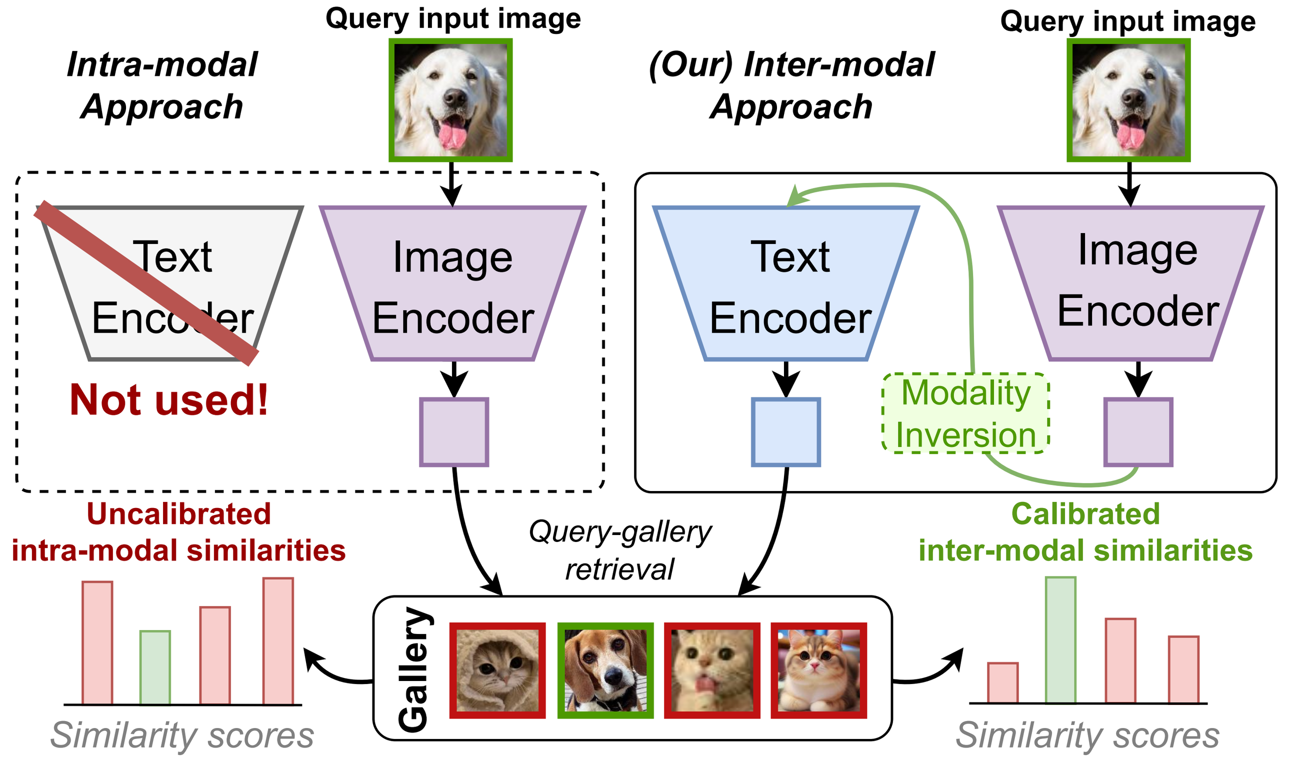
1. Defining CLIP Intra-Modal Misalignment

- VLMs (like **CLIP**) are used off-the-shelf for a variety of applications
- However, CLIP pretraining aligns *only* image-text pairs, and does not ensure that two similar images (or texts) are close to each other
- An image of a dog might end up closer to an image of a cat than to another dog.
- We call this overseen issue **intra-modal misalignment**

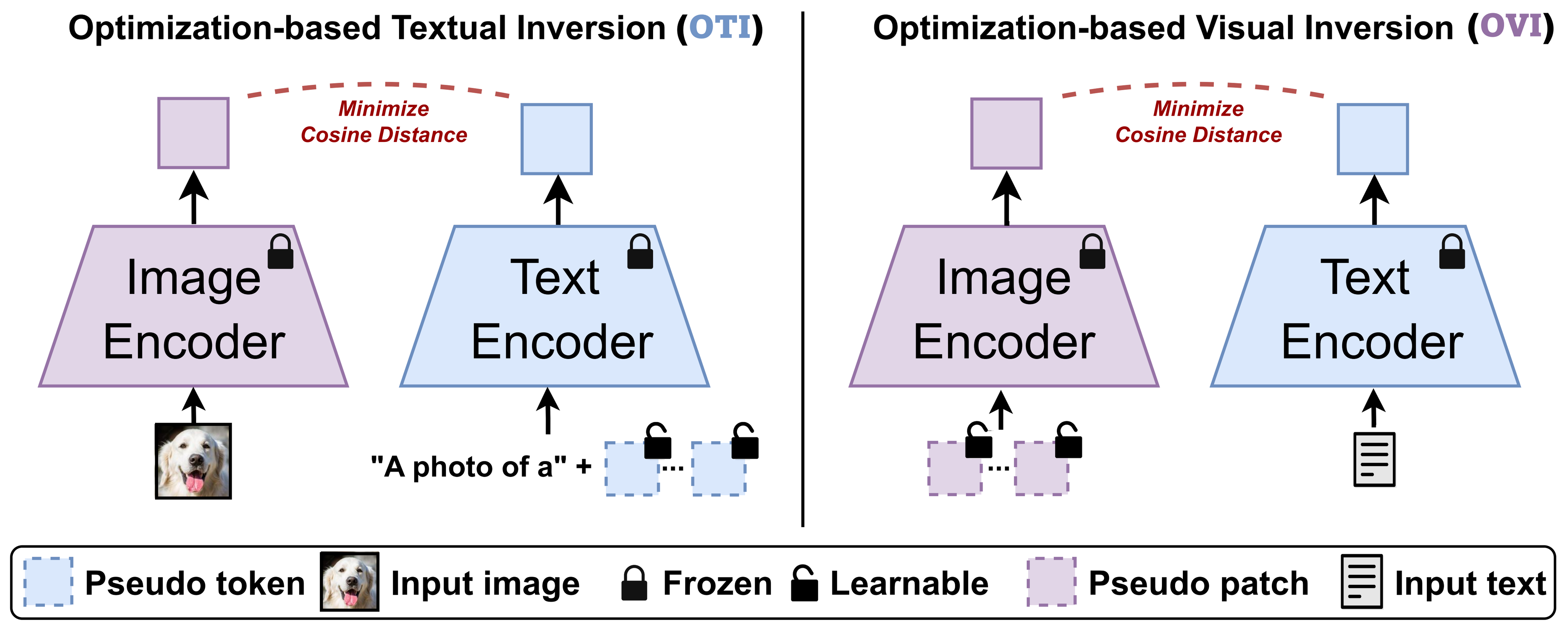


2. How to go from intra-modal to inter modal?

- CLIP features are widely used for **intra-modal comparisons** (e.g., image-to-image retrieval or text-to-text retrieval)
- We argue that common intra-modal methods result in uncalibrated similarities
- We introduce the usage of modality inversion techniques to *approach* any intra-modal task **inter-modally**
- This shouldn't help unless *intra-modal misalignment* is real!
- We show that *inter-modal similarities* outperform intra-modal baselines



3. Proposed modality inversion techniques



Both inversion techniques are single-feature level, freeze the backbones, and optimize a few parameters by minimizing the *cosine distance* with the input feature

4. Approach intra-modal task intermodally

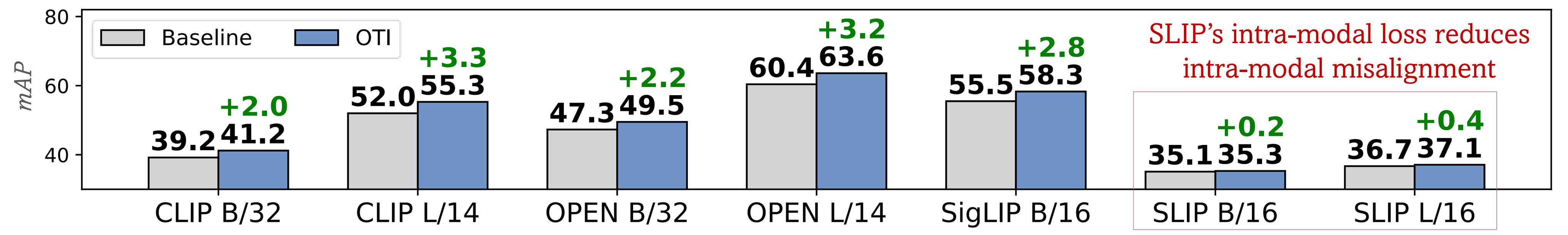
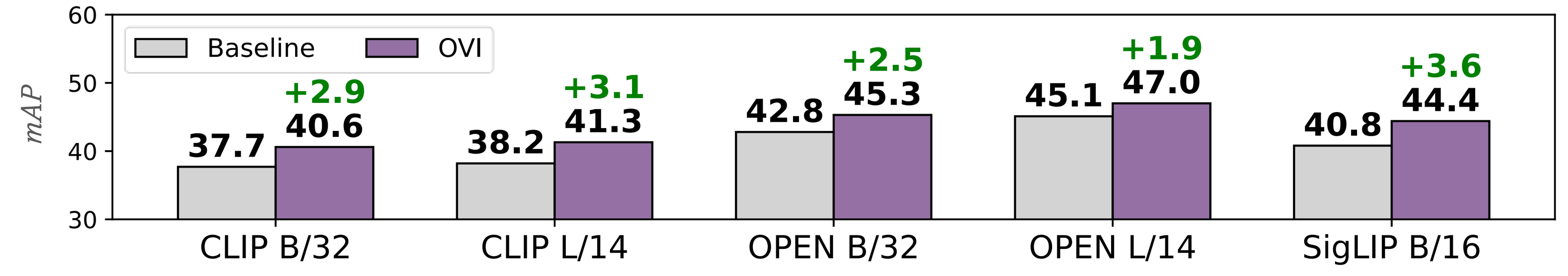


Image-to-image retrieval, **natively intra-modal**, approached *inter-modally* with **OTI**



Text-to-text retrieval, **natively intra-modal**, approached *inter-modally* with **OVI**

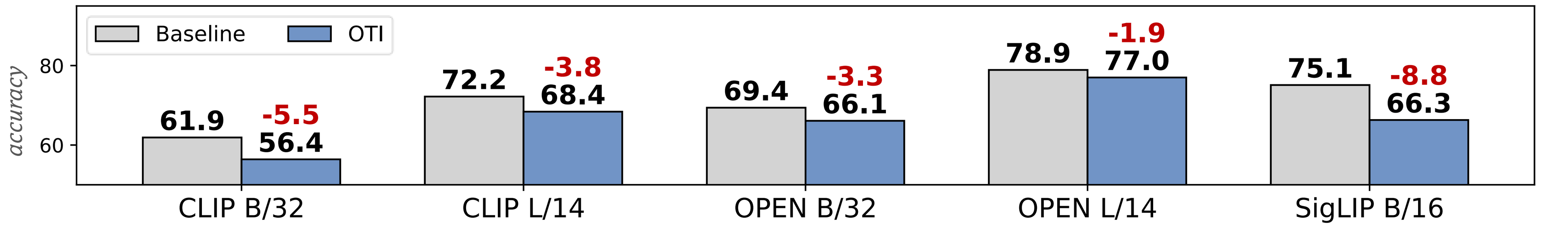
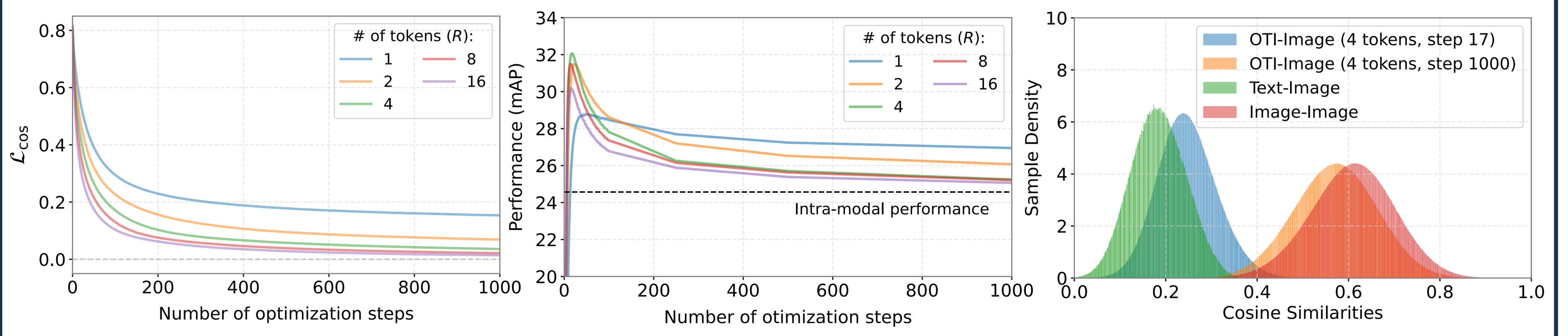


Image classification, **natively inter-modal**, approached *intra-modally* with **OTI**

5. Analyzing the Modality Inversion



Performance peaks early during optimization, before *features drift* toward the native manifold. While a lower number of learnable tokens (*R = 1*) offers more robustness, larger values accelerate convergence and improve peak performance.

6. Role of the modality gap

- We fine-tune CLIP on COCO using different temperatures, preserving and closing the modality gap
- Modality inversion benefit correlates with the magnitude of the modality gap

Fine-tuning Temperature	Inter modal	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	*Average*
$\tau = 1$ *(no gap)*	✗	**15.9**	**23.7**	**29.3**	**46.6**	**19.3**	**27.0**
	✓	14.0	20.4	26.7	43.1	17.4	24.2
$\tau = 0.01$ *(CLIP gap)*	✗	24.0	35.0	43.1	68.6	25.7	39.3
	✓	**24.1**	**35.2**	**44.0**	**70.2**	**27.6**	**40.2**

7. OTI and OVI pseudo-algorithms & Source Code!

Algorithm 1 OTI

1: **Input:** Image I, number of pseudo-tokens R, number of optimization steps S
2: Initialize $v^* = \{v_1^*, v_2^*, \ldots, v_R^*\}$
3: Extract image features: $\psi_I = f_\theta(I)$
4: **for** $s = 1$ to S **do**
5: Form $\overline{Y}_{v^*} = [E_v(\text{"a photo of"}), v^*]$
6: Extract text features: $\psi_T = g_\phi(\overline{Y}_{v^*})$
7: Compute loss: $\mathcal{L}_{\cos} = 1 - \cos(\psi_I, \psi_T)$
8: Update v^* to minimize $\mathcal{L}_{\cos}$
9: **end for**
10: **Output:** OTI-inverted features $\psi_T = g_\phi(\overline{Y}_{v^*})$

Algorithm 2 OVI

1: **Input:** Text Y, number of pseudo-patches P, number of optimization steps S
2: Initialize $w^* = \{w_1^*, w_2^*, \ldots, w_P^*\}$
3: Extract text features: $\psi_T = g_\phi(E_v(Y))$
4: **for** $s = 1$ to S **do**
5: Form input $\bar{I}_{w^*}$ using **??**
6: Extract image features: $\psi_I = f_\theta(\bar{I}_{w^*})$
7: Compute loss: $\mathcal{L}_{\cos} = 1 - \cos(\psi_I, \psi_T)$
8: Update w^* to minimize $\mathcal{L}_{\cos}$
9: **end for**
10: **Output:** OVI-inverted features $\psi_I = f_\theta(\bar{I}_{w^*})$

The code implementing **OTI** and **OVI** with all the different backbones and on all the evaluated datasets is finally available. Feel free to explore, use and contribute!

GitHub