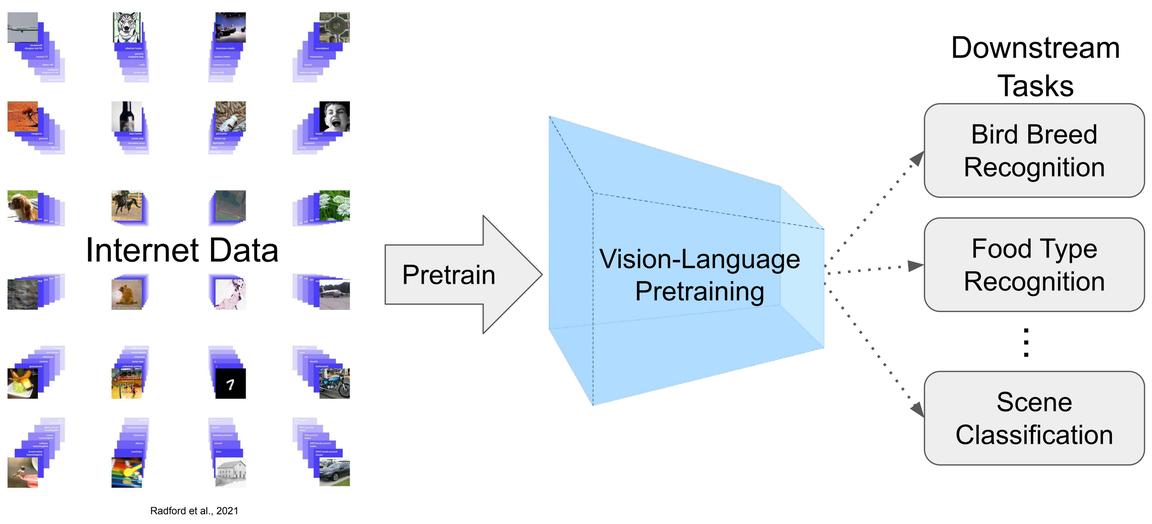
SVL-Adapter: Self-Supervised Adapter for Vision-Language Pretrained Models

Omiros Pantazis[1] Gabriel Brostow[1] Kate Jones[1] Oisin Mac Aodha[2,3]

1 UCL 2 The Alan Turing Institute 3 The University of Edinburgh

Motivation

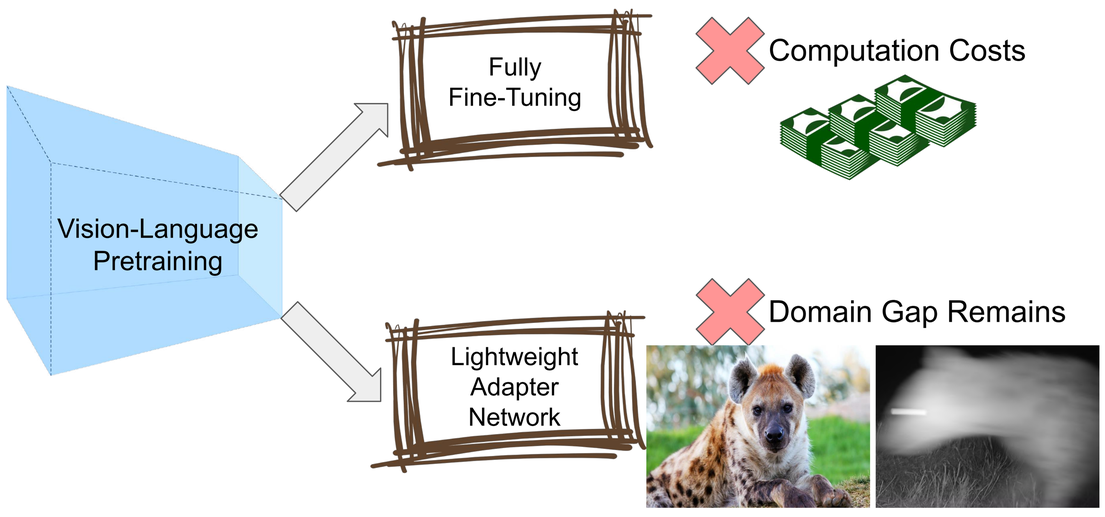
Jointly Learning from Vision & Language



Hundreds of millions image/text pairs available on the web

Vision-Language methods like CLIP exhibit **impressive zero-** and **low-shot transfer**

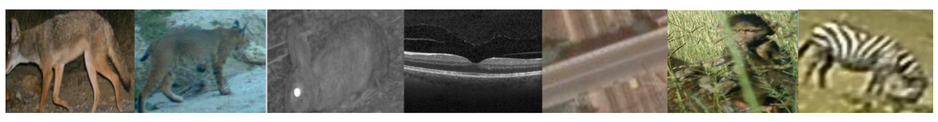
Pitfalls of Vision-Language Transfer



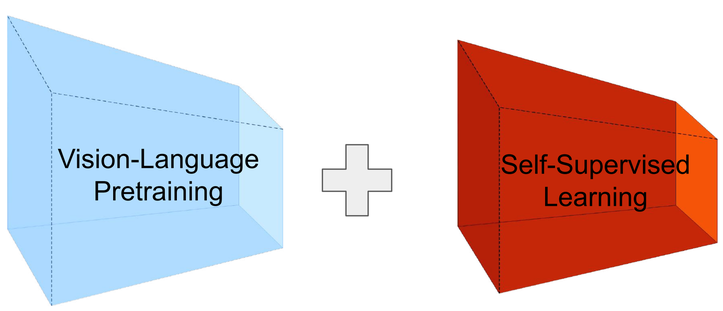
Few-shot learning on top of Vision-Language learnt features **not enough** if **downstream task diverges from internet-style**

Proposition

Utilize **datasets that differ from the content found online** to **test the limits of Vision-Language adaptation**

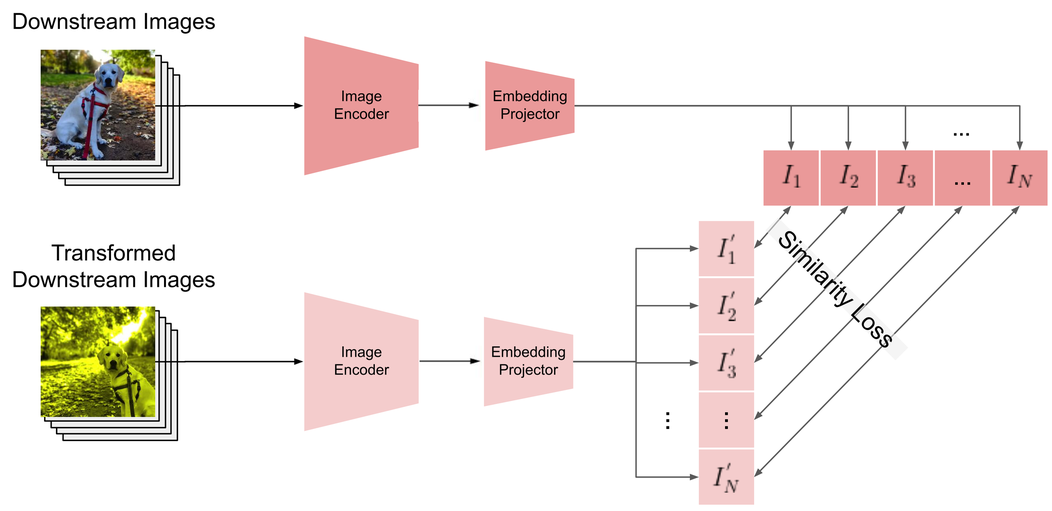


Assist Adaptation by combining **Large-Scale Vision-Language Pretraining** and **Targeted Self-Supervised Learning**



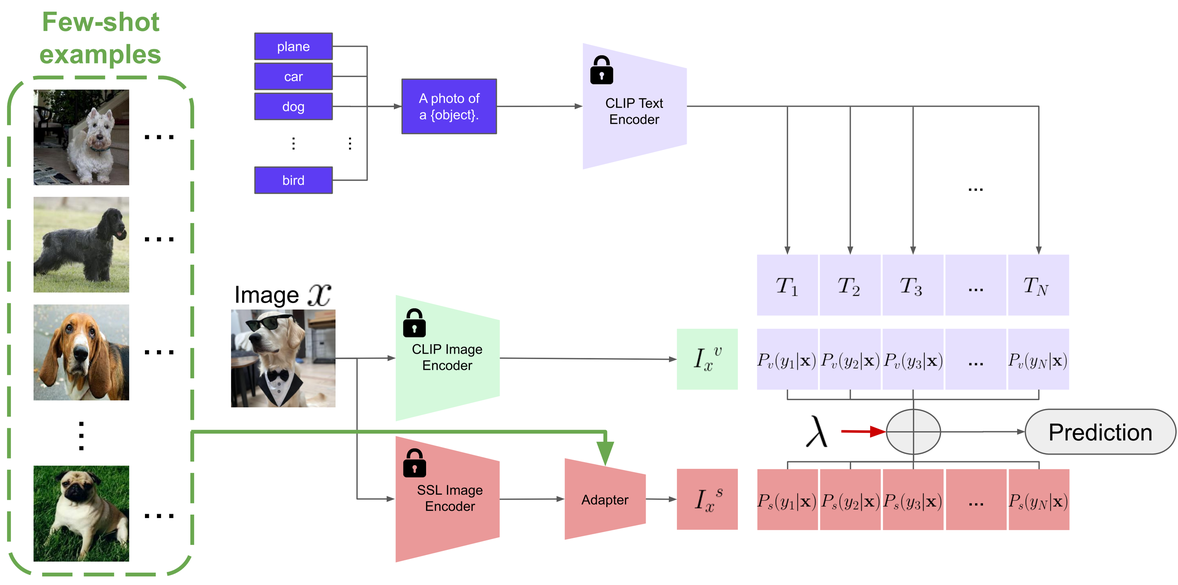
Method: SVL-Adapter

Self-Supervised Learning for Pretraining



We deploy a **Self-Supervised Learning** approach such as SimCLR to learn representations relevant to the downstream task

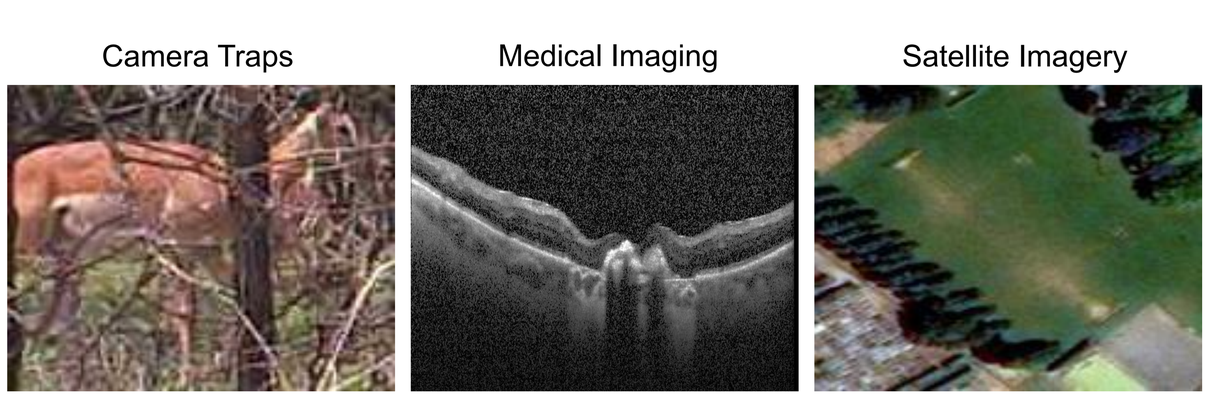
Downstream Adaptation and Fusion with CLIP



- Given few-shots, train an **adapter on top of Self-Supervised features**
- **Combine classfication outputs** of the trained adapter with zero-shot CLIP
- We show how the **blending hyperparameter** that optimally fuses the two outputs can be **selected automatically** (SVL-Adapter*)
- **Zero-shot version of SVL-Adapter** where **CLIP pseudolabels** are utilized as few-shot examples

Evaluation

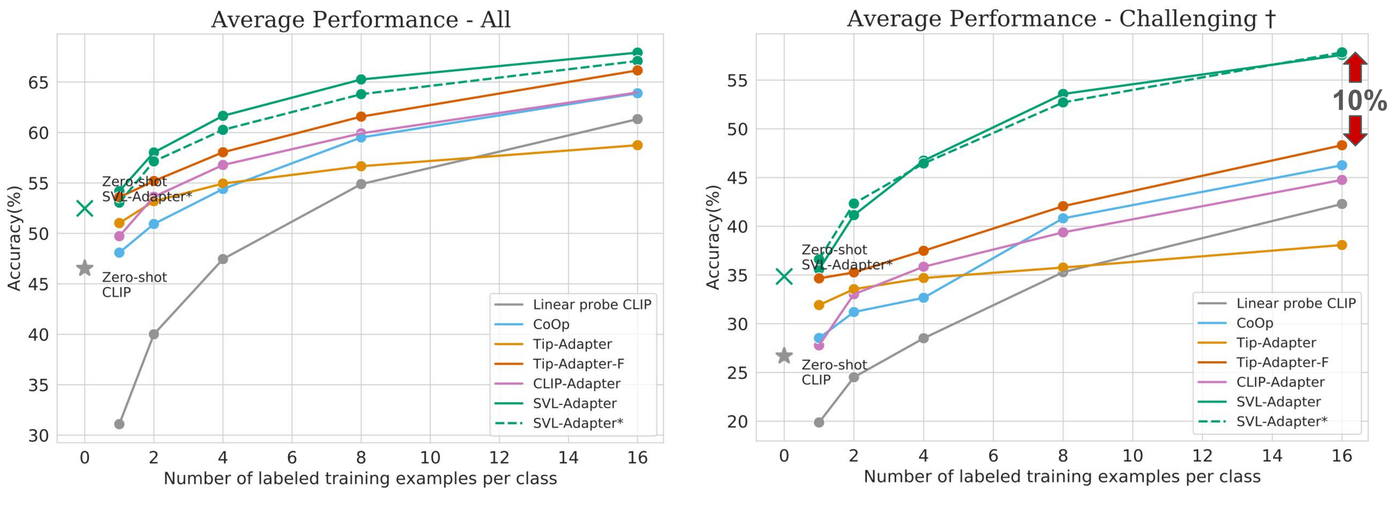
Testbed: Real-World Challenging Datasets



- We evaluate our approach in **10 Standard** and **6 Challenging d**atasets
- We compare with **zero-shot** and **linear probe CLIP** and **state-of-the-art Vision-Language adaptation baselines**

Results

SVL-Adapter outperforms baselines with **significant gains** in the **Challenging tasks**



- **SVL-Adapter consistently better than baselines**, while SVL-Adapter* follows closely
- **Significant gains** of about 10% on average **on challenging tasks**
- **Zero-shot version of SVL-Adapter improves** considerably upon **zero-shot CLIP** predictions

Summary

- Applying Vision-Language adaptation is not straightforward in challenging real-world datasets
- SVL-Adapter: a way to combine large-scale Vision-Language pretraining and targeted Self-Supervised Learning
- Outperforming baselines on zero- and low-shot learning, with significant gains in challenging tasks

More results available in our paper!