Exploring Robust Features for Few-Shot Object Detection in Satellite Imagery

Xavier Bou, Gabriele Facciolo, Rafael Grompone, Jean-Michel Morel, Thibaud Ehret
xavier.bou_hernandez@ens-paris-saclay.fr

CVPR SEATTLE, WA JUNE 17-21, 2024

Code

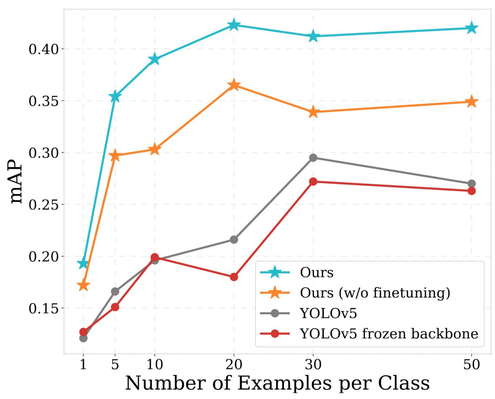
Paper

Context

Object detection typically requires large amounts of annotated data, which is scarce in remote sensing. Open Vocabulary Detection (OVD), which aims to detect objects beyond the set of training classes, has recently gained popularity due to the emergence of Vision-Language Models (VLMs). In this work, we explore how to apply OVD for few-shot detection in remote sensing.

Contributions

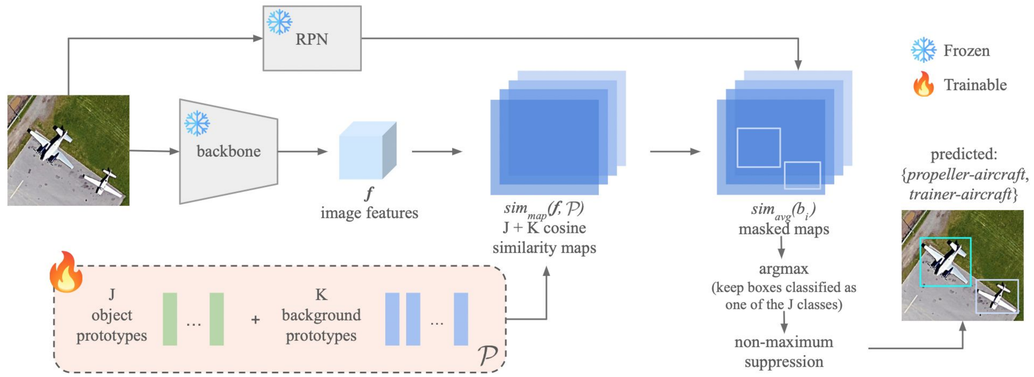
- ➔ We show that **visual features significantly outperform vision-language representations.**
- ➔ By optimizing only class-reference prototypes, we discriminate unseen objects with minimal data, enabling **generalization with very few training parameters**.
- ➔ **We detect any object in satellite data with only a handful of examples**: our approach outperforms popular methods on unseen and fine-grained classes, e.g. YOLOv5 on the SIMD dataset



Methodology

We re-purpose a traditional two-stage object detector for few-shot object detection:

- ➔ **Region Proposal Network (RPN)**: The RPN is pre-trained on an available remote sensing dataset (e.g. DOTA) to generate class-agnostic bounding box proposals for potential objects.
- ➔ **The classification step is replaced by a prototype optimization approach** that learns a class-reference embedding with only a few examples per class.
- ➔ **Region proposals are classified via cosine similarity** between the generated prototypes and the image representations within the proposed bounding box.



Building class-reference prototypes

Using robust representations, we build a set of class-reference and background prototypes to discriminate objects of interest with a handful of examples.

- ➔ **Initialize object prototypes**: For each class, pre-trained representations within an object's bounding box are extracted and averaged, yielding a single embedding vector per class.
- ➔ **Initialization of K background prototypes**: Representations from random image crops (that do not intersect any annotation) are clustered using K-means. The resulting K centroids are used as embeddings for representing the background.
- ➔ **Fine-tune prototypes**: Prototypes are fine-tuned to improve object discrimination. This is achieved by learning a new set of prototypes through optimizing the cross-entropy loss over the available annotated bounding boxes, thus better classifying objects to their respective classes. Since only class-reference embeddings are optimised, the limited amount of training examples does not lead to overfitting.

References

[1] Maxime Oquab *et al.* DINOv2: Learning robust visual features without supervision. TMLR, 2024.
[2] Alec Radford *et al.* Learning transferable visual models from natural language supervision. ICML, 2021.
[3] Muhammad Haroon *et al.* Multisized object detection using spaceborne optical imagery. IEEE JSTARS, 2020.
[4] Ke Li *et al.* Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS JPRS, 2020.
[5] Bingyi Kang *et al.* Few-shot object detection via feature reweighting. ICCV, 2019.
[6] Xinyu Zhang *et al.* Detect every thing with few examples. arXiv preprint arXiv:2309.12969, 2023.

Visual vs. vision-language features

Backbone	Fine-tuned	Architecture	c_{novel}	c_{base}
CLIP		ViT-B/32	0.113	0.201
CLIP		ViT-L/14	0.236	0.306
GeoRSCLIP		ViT-B/32	0.132	0.270
GeoRSCLIP	✗	ViT-L/14	0.161	0.34
RemoteCLIP		ViT-B/32	0.124	0.274
RemoteCLIP		ViT-H/14	0.117	**0.482**
DINOv2		ViT-L/14	**0.306**	0.416
CLIP		ViT-B/32	0.190	0.098
CLIP		ViT-L/14	0.215	0.451
GeoRSCLIP		ViT-B/32	0.097	0.228
GeoRSCLIP	✓	ViT-L/14	0.224	0.420
RemoteCLIP		ViT-B/32	0.116	0.229
RemoteCLIP		ViT-H/14	0.086	**0.452**
DINOv2		ViT-L/14	**0.358**	0.377

DINOv2 features significantly outperform VLM features on uncommon objects, including those specific to remote sensing:

- ➔ **Satellite datasets are overfitted to a small set of general objects**. VLMs tailored for remote sensing are familiar with common objects but perform poorly on rare ones.
- ➔ **Visual features do not need to identify the object specifically**. Similar-looking objects are represented similarly in the feature space.

Quantitative Results

Method	Backbone	5-shot		10-shot		30-shot	
		SIMD	DIOR	SIMD	DIOR	SIMD	DIOR
YOLO	YOLOv5	16.60	4.23	19.57	10.28	29.48	16.99
YOLO	YOLOv5 (frozen)	15.05	5.70	19.94	9.42	27.18	14.90
FSRW	DarkNet-19	11.04	10.20	13.70	15.06	23.77	25.79
DE-ViT	ViT-L/14	20.43	9.12	20.44	8.95	20.06	9.33
Ours	ViT-L/14	**35.44**	9.56	**38.99**	12.51	41.21	12.60
Ours + FSRW	ViT-L/14	29.14	**15.06**	38.61	**18.77**	**41.40**	**26.46**

Qualitative Results (Top: SIMD [3], bottom: DIOR [4])