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

Xavier Bou[1] Gabriele Facciolo[1] Rafael Grompone von Gioi[1] Jean-Michel Morel[2]
Thibaud Ehret[1]
[1]Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, France
[2]City University of Hong Kong, Department of Mathematics, Kowloon, Hong Kong
xavier.bou_hernandez@ens-paris-saclay.fr

Abstract

The goal of this paper is to perform object detection in satellite imagery with only a few examples, thus enabling users to specify any object class with minimal annotation. To this end, we explore recent methods and ideas from open-vocabulary detection for the remote sensing domain. We develop a few-shot object detector based on a traditional two-stage architecture, where the classification block is replaced by a prototype-based classifier. A large-scale pre-trained model is used to build class-reference embeddings or prototypes, which are compared to region proposal contents for label prediction. In addition, we propose to fine-tune prototypes on available training images to boost performance and learn differences between similar classes, such as aircraft types. We perform extensive evaluations on two remote sensing datasets containing challenging and rare objects. Moreover, we study the performance of both visual and image-text features, namely DINOv2 and CLIP, including two CLIP models specifically tailored for remote sensing applications. Results indicate that visual features are largely superior to vision-language models, as the latter lack the necessary domain-specific vocabulary. Lastly, the developed detector outperforms fully supervised and few-shot methods evaluated on the SIMD and DIOR datasets, despite minimal training parameters.

1. Introduction

Object detection in remote sensing data is a crucial problem for earth observation applications such as intelligent monitoring, urban planning and precision agriculture [18]. It consists of locating and assigning labels to objects of interest in an image. In recent years, fully supervised object detection has shown impressive performances with methods like YOLO [29], Faster RCNN [30] or Mask-RCNN [10]. However, these methods require large amounts of annotated data which are not currently available in remote sensing.

More recently, the emergence of large-scale vision-language models (VLMs) introduced the problem of Open-Vocabulary Detection (OVD), which aims to detect objects beyond the set of training classes. OVD methods try to exploit the knowledge of these pre-trained VLMs to perform detection conditioned by their image and/or text embeddings [7, 23, 35, 41]. Conveniently, this leads to a drastic reduction of the annotation cost and enables the use of text prompts, for zero-shot, or a reference image, for one-shot, to specify novel target classes.

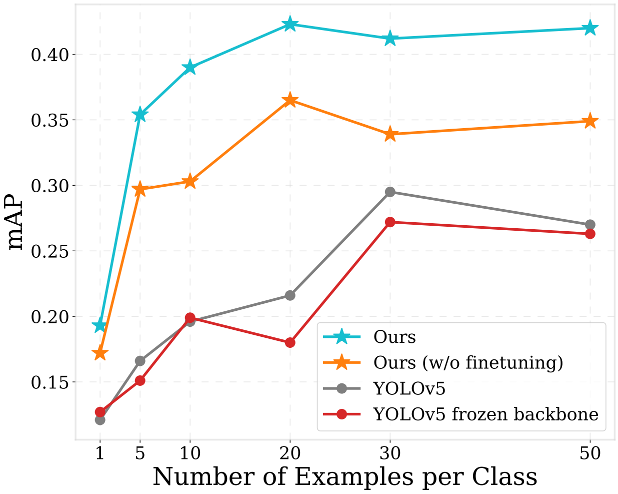

Figure 1. Performance (mAP) of the proposed detector with DINOv2 features on the SIMD dataset, compared to YOLOv5 for different amounts of available examples per class. Robust visual features largely outperform state-of-the-art supervised methods when annotated data is limited.

These methods show impressive capabilities detecting rare object classes that are uncommon amongst popular datasets. Yet, the amount of aerial or satellite image examples used during training is minimal. Hence, due to the large gap between natural images and optical remote sensing data, the performance of current OVD methods on the latter is quite poor. In addition, OVD methods require delicate prompt engineering to find the most suitable wording for desired object classes. It has been indeed shown that a slight word change can negatively or positively affect detection performance [5, 26]. For this reason, some works have proposed approaches to automatically find the most suitable adjectives [42], text tokens [26] or embeddings [13] to improve classification using a few examples, framing the problem as Few-Shot Object Detection (FSOD). Furthermore, DE-ViT [26] recently introduced the idea that CLIP text embeddings are not discriminative enough, proposing a framework using purely visual DINOv2 features [25].

In this work, we revisit the ideas of OVD and FSOD based on large-scale pre-trained models and explore their capabilities for FSOD in remote sensing. To this end, we re-purpose a two-stage object detector for FSOD, where the classification step is replaced by an OVD-inspired classifier that uses feature embeddings as reference classes. Furthermore, the limited class examples are used to fine-tune the class reference embeddings in automatic prompt-engineering mode, learning the difference between target objects and background classes.

We compare the performance of our detector with other FSOD and fully supervised methods. In addition, we explore several robust features for both visual and vision-language models, including RemoteCLIP [21] and GeoRSCLIP [39], which are specifically tailored for remote sensing applications. Our results indicate that visual features are more suitable for remote sensing detection, as image-text approaches seem to be limited by the granularity of their image captions. Furthermore, DINOv2 representations show an impressive ability to discriminate similar types of rare classes, such as types of aircraft or vehicles. The proposed detector outperforms all other evaluated methods on the SIMD and DIOR datasets. Figure 1 illustrates the detection advantage of the proposed framework over a fully supervised approach for minimal annotations.

2. Related works

Closed-vocabulary object detection is a traditional problem in image understanding that attempts not only to identify objects in images but to precisely estimate their location as well [40]. In recent years, advances in deep neural networks have led to great success in the field, and such approaches have consolidated as the state of the art over classical algorithms [34]. Current-day object detectors can generally be divided into two-stage and one-stage detectors. Two-stage object detectors proceed in two steps. First, a region proposal network (RPN) extracts region-of-interest (RoI) features and generates class-agnostic bounding box proposals, which are later classified by a sub-network [6, 10, 30]. On the other hand, one-stage detectors avoid the time-consuming proposal generation and work directly over a dense sampling of locations or anchors [11, 20, 29]. While one-stage detectors achieve faster inference, this can come at the cost of performance. Despite recent progress, closed-vocabulary object detection methods are limited to detecting only classes seen during training, requiring significant annotation and training effort to extend them to new categories.

Few-Shot Object Detection aims to detect objects in images with limited annotated data to handle the absence of objects in common large-scale datasets. The general underlying idea is training a detector on a set of base classes with a large amount of annotated bounding boxes, and then adapting the classification step to perform for new, unseen classes using only a few examples [16]. Kang *et al.* [12] proposed a feature re-weighting module (FSRW) that quickly learns to use general meta-features to detect novel classes. Snell *et al.* [31] proposed a simple approach to few-shot classification called prototypical networks, where each class is represented by a mean vector of the embedded support points belonging to its class. We build on this idea in our study.

Open-Vocabulary Detection tries to detect objects beyond the set of categories seen during training, i.e. reducing the need for re-training. It has recently gathered attention in the literature, propelled by the development of large-scale vision-language models such as CLIP [27], which learns relationships from both image and text with a shared embedding space. OVD was introduced by Zareian *et al.* [36], who provided a framework to learn both weakly supervised and zero-shot capabilities. Then, Vild [7] proposed to use CLIP to match text embeddings of novel classes to corresponding representations in image crops, so that unseen objects are supported at inference. RegionCLIP [41] generates location pseudo-labels from image-caption pairs, which are then used to align region-text pairs in the feature space via self-supervised learning. OWL-ViT [23, 24] scaled these methods with a simple transformer-based architecture in two steps. First, a vision-language model is trained, which is then fine-tuned to perform object detection conditioned by CLIP embeddings. As image and text share a common embedding representation, OWL-ViT can perform detection conditioned by an image instead of text, thus enabling one-shot detection. More recently, Kaul *et al.* [13] developed a framework to build multi-modal OVD classifiers. They pretrain a class-agnostic RPN and train an aggregator network that utilizes CLIP embeddings from a few examples to generate a classifier vector. DE-ViT [38], on the other hand, diverges from reliance on large vision-language models, opting for purely visual DINOv2 [25] features instead. Class reference vector prototypes are built for each class using a few examples, which are then used for the classification of region proposals. A distinctive work by Parisot *et al.* [26] illustrates the significant impact of prompt engineering on OVD, and derives an approach to fine-tune learnable text tokens, thus reaching the most suitable category prompt for a set of examples. While some recent works require a few examples and thus diverge from OVD towards few-shot detection, they reveal great potential for the detection of rare objects or concepts. Moreover, commonly used large-scale vision-language and vision models are tailored for natural images, as the number of remote sensing data seen during training is minimal. Our work evaluates their performance on remote-sensing data.

Object Detection in Remote Sensing is a fundamental problem in the field of aerial image analysis, assuming an important role in a number of applications due to the increasing availability of satellite images [1]. The success of deep learning-based object detection in natural images inspired the earth observation community to drive efforts to object detection in optical remote sensing data. As a consequence, several benchmarks were proposed [8, 18, 28, 33], and various works developed tailored approaches for the characteristics of satellite imagery, e.g. rotation-invariance or small object detection [2, 17, 43]. Nevertheless, most remote sensing datasets present shortcomings, such as a low number of categories, class imbalance or insufficient image diversity and variation [18]. These constraints obstruct the progress of traditional deep learning-based object detectors, as they require large amounts of well-annotated, curated data. Additionally, some works have explored few-shot methods in this domain. Deng *et al.* [19] introduces a reweighting module that re-calibrates feature maps from a set of labelled support images on a YOLO architecture. Zhang *et al.* [37] proposes a few-shot approach focusing on avoiding catastrophic forgetting of the base classes. Wolf *et al.* [32] designed a double-head architecture to prevent knowledge loss of base classes, paired with a sampling and pre-processing strategy to better exploit base class annotations. Cheng *et al.* [3] introduced a prototypical approach based on ResNet101 [9] features that adapts a two-stage architecture to detect and classify objects based on support images. More recently, Lu *et al.* [22] proposed to fuse visual features with text description features for each category, reducing the classification confusion of novel classes.

Nevertheless, few works attempt few-shot detection beyond common, general classes, e.g. *airplane* or *vehicle*. Furthermore, despite some works have adapted image-text pre-training strategies to the remote sensing domain [21, 39], no attempts have been made to use these tools and ideas for detection in optical remote sensing data. Consequently, our study on robust representations for few-shot detection in optical remote sensing, akin to recent works on OVD and FSOD for natural images, is relevant and of interest to the community.

3. Methodology

In this section, we first frame the problem we are addressing. Subsequently, a comprehensive breakdown of our detector is provided, including the architecture overview, how class reference prototypes are built, the classification step, and how prototypes are fine-tuned to improve their classification capabilities.

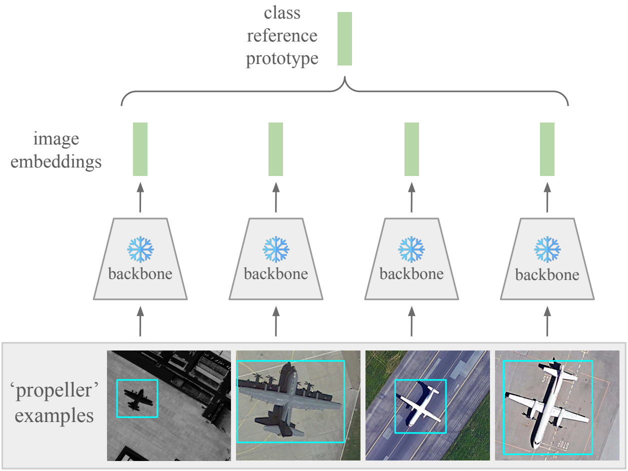

Figure 2. Building a class reference prototype for the aircraft category *propeller* with four examples. The frozen pre-trained backbone is used to extract image representations. Then, patches overlapping box annotations are averaged into one single vector. Lastly, all four embeddings are combined into a reference vector via averaging and normalization.

Problem setup. We consider the problem of detecting objects in optical remote sensing data with limited annotations. We assume to have two remote sensing datasets: a large dataset containing general, common object classes $\mathcal{D}^{train}$, and a dataset containing a very limited number of training instances $\mathcal{D}^{target}$ per class. The objects in $\mathcal{D}^{target}$ can be grouped into base classes c_{base} and novel classes c_{novel}, where the first correspond to those contained in $\mathcal{D}^{train}$, and the latter belong to object classes that have never been seen by the model before. Our goal is to detect c_{novel} objects from the limited examples available.

The detector takes an image $I \in \mathbb{R}^{3\times H\times W}$ as input and predicts the location and class of present objects. For the i^{th} predicted object in an image, bounding box coordinates $b_i \in \mathbb{R}^4$ and a class prediction $\hat{c}_i \in \mathcal{C}$ are generated, where $\mathcal{C}$ is a set of classes such that $\mathcal{C} = c_{\text{base}} \cup c_{\text{novel}}$.

Architecture overview. Following the approach of DE-ViT [38] and Kaul *et al.* [13], we opt for a standard two-stage architecture and use a Faster-RCNN [30] as region proposal network. For classification, we extract robust image features using a pre-trained backbone, which consists of a Transformer encoder architecture [4] that transforms an image $I \in \mathbb{R}^{3\times H\times W}$ into a high-dimensional feature representation f:

$$Backbone(I) = f \in \mathbb{R}^{\frac{H}{ps}\times\frac{W}{ps}\times D}, \tag{1}$$

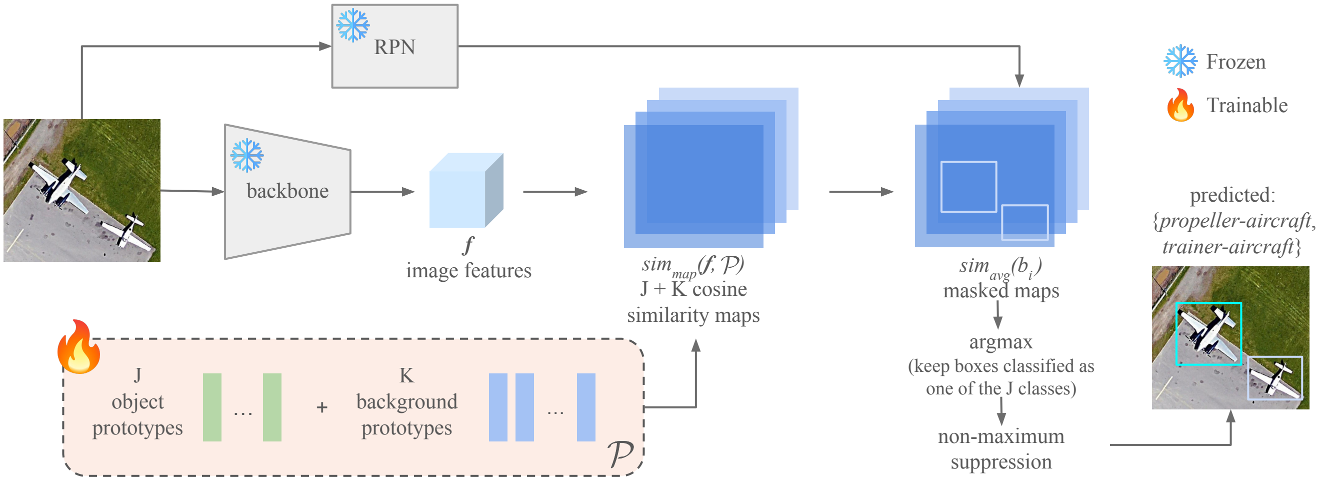

Figure 3. General diagram of our detector. An input image is fed to the RPN to generate region proposals, as well as to the backbone to extract high-level representations. Then, cosine similarity maps are generated using the features and pre-computed prototypes. For each region proposal, the mean average similarity with each prototype is computed, and the proposal is then classified as the most similar prototype class. Lastly, we discard boxes classified as a background prototype and apply non-maximum suppression.

where H and W are the image height and width, and p_s and D the transformer patch size and feature dimensionality, respectively. Class reference prototypes are then built to represent object categories, and region proposals are classified via cosine similarity between prototypes and image features. This process is detailed in the following sections.

3.1. Building prototypes

Object prototypes. We want to build a set of reference prototype embeddings $\mathcal{P} = \{p_1, p_2, \dots, p_J\}$ that represent target classes $\mathcal{C} = \{c_1, c_2, \dots, c_J\}$ and can be used to classify region proposals. Hence, for an object category c_j containing N_j examples, its representative prototype p_j is computed as

$$p_j = \frac{\hat{p}_j}{\|\hat{p}_j\|_2}, \quad \text{where} \quad \hat{p}_j = \frac{\Sigma_{n=1}^{N_j} \sum_{(l,h)\in b_n} (f_n)_{l,h}}{N_j X_n Y_n}, \quad (2)$$

b_n is the bounding box annotation of the n^{th} object, $(f_n)_{l,h}$ is the feature representation of image n at position (l, h), and X_n and Y_n are and the width and height of the object ground truth bounding box b_n, respectively. Figure 2 depicts the process of building a prototype for a specific class.

Background prototypes are additionally built to reduce the number of false alarms, addressing the case of invalid region proposals. While the background appearance of natural images is highly variable and often unrelated to object categories, satellite imagery contains a finite number of backgrounds, i.e. the different earth land cover types (water, pavement, urban, forest, etc.). For this reason, we propose to generate K background prototypes using object-free areas in available images. To this end, we extract image representations of all available images and generate a number of crops per image that do not overlap with any labelled instance. Subsequently, we cluster these representations into K clusters using K-Means, and generate a background prototype per cluster by averaging the embeddings in each cluster into a single embedding vector. Overall, our detector uses J object prototypes and K background prototypes $\mathcal{P} = \{p_1, p_2, ..., p_{J+K}\}$, which are initialized offline.

3.2. Classification

During inference, input image I is fed to the detector and a set of region proposals are generated by the RPN. Classification is subsequently performed as follows: we first extract image features f by applying the pre-trained backbone to image I and upsample it to input image resolution. Then, a similarity map is generated by computing the cosine similarity between extracted features and pre-computed prototypes $\mathcal{P}$:

$$sim_{map}(f, \mathcal{P}) = \frac{f \cdot \mathcal{P}}{\|f\| \|\mathcal{P}\|} \in \mathbb{R}^{H \times W \times (J+K)}. \quad (3)$$

Afterward, we extract the sim_{map} similarity values inside each bounding box proposal $b_i \in \mathbb{R}^4$, and compute the average bounding box similarity for every prototype as follows:

$$sim_{avg}(b_i) = \frac{\sum_{(l,h)\in b_i} sim_{map}(f, \mathcal{P})_{l,h}}{X_i Y_i} \in \mathbb{R}^{J+K}, \quad (4)$$

where $sim_{map}(f, \mathcal{P})$ is the cosine similarity map and X_i and Y_i are the width and height of b_i, respectively. Then, b_i is classified as the class of higher $sim_{avg}(b_i)$ according to its prototype. Boxes classified as a background prototype are discarded. Lastly, non-maximum suppression is applied to avoid detecting the same object multiple times.

3.3. Fine-tuning prototypes

As detailed in Section 3.1, prototypes are built by averaging available object representations into a single reference vector. Nevertheless, bounding boxes include not only the object of interest but a portion of the background as well, introducing undesirable information into class prototypes. Hence, averaging their features might not be the most appropriate approach to classify such cases. Inspired by Parisot *et al.* [26], which uses learnable word embeddings to find the best suited class names for unseen categories given a VLM, we derive a fine-tuning approach to improve the discriminative capabilities of our prototypes. Similarly, we propose to fine-tune the pre-computed prototypes $\mathcal{P}$ to learn better representations for each class, given available images.

To this end, we optimize prototypes to classify available ground truth boxes between the set of object and background classes. In addition, we randomly sample image crops that do not intersect with object ground truth bounding boxes, and use them as negative examples that need to be classified as background prototypes. For each negative example, we define its ground truth as the class with the most similar prototype amongst the set of background prototypes. More formally, the classifier of our detector predicts a class $\hat{c}_i$ given a region proposal b_i, an image feature representation f and set of prototypes $\mathcal{P}$:

$$Classifier(b_i, f, \mathcal{P}) = \hat{c}_i. \tag{5}$$

Thus, learning a set of prototypes $\hat{\mathcal{P}}$ that optimizes the cross-entropy loss objective function over the annotated bounding boxes, given their ground truths. The weights of the backbone are kept frozen at all time.

4. Experiments

Implementation details. We use the visual model DINOv2 [25] and the vision-language model CLIP [27] as pre-trained backbones. We evaluate different versions for the latter, including two remote sensing-tailored models. We empirically set $K = 200$ background prototypes and explore different numbers of negative region proposals per image. We apply several spatial augmentations to input images, consisting in random horizontal and vertical flips, random rotations, color jitter, padding and random resized crops of size 602×602. Detailed information on the training setup and hyper-parameters is provided in the supplementary material.

Experimental setup. We select the DOTA [33] dataset as $\mathcal{D}^{train}$, which contains 2,806 large images with 403,318 annotated instances of 16 general classes, such as *plane*, *ship*, *small vehicle* or *storage tank*. We pre-process the dataset to obtain images of size 800×800 with an overlap of 200 pixels. Then, a Faster-RCNN model is trained on the entire train set, and the resulting RPN is extracted to serve as our region proposal network. We consider SIMD [8] and DIOR [18] as target datasets $\mathcal{D}^{target}$. The SIMD dataset comprises 5,000 images of resolution 1024 × 768 with 45,096 annotated objects, which consist of several aircraft types, e.g. *propeller-aircraft*, *fighter-aircraft* and *airliner*, amongst other more general vehicles, e.g. *car*, *van* or *truck*. Refer to the supplementary material for a detailed description of the base and novel classes of each dataset. The DIOR dataset contains 23,463 images and 192,472 annotations over 20 different classes. While some categories are shared with the DOTA dataset, other significantly different classes are found, such as *chimney*, *express-toll-station*, *airport* or *trainstation*. To address the few-shot detection performance of novel classes, we generate test subsets for different numbers of examples per class $N = \{5, 10, 30\}$. Class imbalance is highly common in optical remote sensing, as pointed out in Section 2, and we often observe multiple annotations in a single image. Thus, randomly generating a subset of exactly N instances for all classes can be challenging. For this reason, certain classes in our subsets can contain slightly more or fewer examples than N. All data splits will be publicly released, and detailed information can be found in the supplementary material. Lastly, we report the mAP50 scores on novel classes to measure the few-shot detection performance of all evaluated models.

Results. We compute the mAP50 results for novel classes on 5-shot, 10-shot and 30-shot of our detector, and benchmark our approach to other methods from the literature. Firstly, we select YOLO as a reference for fully supervised approaches. Hence, we pre-train a YOLOv5 network on the entire DOTA dataset and subsequently fine-tune it using the available few-shot data for each case. Moreover, we consider two relevant FSOD methods, namely the feature reweighting approach (FSRW) introduced by Kang *et al.* [12] and DE-ViT [38], which uses a prototypical approach with DINOv2 features as well. We observe that for the DIOR dataset, which contains novel classes that are significantly different from the objects in DOTA, the performance of the proposed detector exhibits a decline with respect to the SIMD results. We explore this in detail in Section 4.1, concluding that the RPN is a limiting block in cases where target objects notoriously differ from the objects in $\mathcal{D}^{train}$. For this reason, we re-use the FSRW model as RPN and re-classify its proposals using our learned prototypes. All results are provided in Table 1. As shown, our approach reports large improvements concerning all evaluated methods for the SIMD dataset. It is worth mentioning that while all other methods have previous knowledge of the class *airplane*, they struggle to learn the differences between plane types with minimal examples. Conversely, learning representative DINOv2 prototypes proved to be very discriminative with only a handful of exemplars. It is important to note that our approach uses few training pa-

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 [12]	DarkNet-19	11.04	10.20	13.70	15.06	23.77	25.79
DE-ViT [38]	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**

Table 1. Results (mAP50) on DIOR and SIMD benchmarks for 5-shot, 10-shot and 30-shot detection. Several representative object detection methods are evaluated, including fully supervised and FSOD approaches. In the last row, we use the FSRW approach as RPN, and we re-classify each bounding box using the learned prototypes.

rameters, as only the learnable prototypes are optimized. As for the DIOR dataset, combining the prototypical approach with FSFR yields the best results. This illustrates both the potential of DINOv2 prototypes to classify region proposals and the limitations of pre-training the RPN on the DOTA dataset. As FSRW re-weights pre-trained RoI features of a one-stage detector, the region proposals improve with respect to the base training, resulting in better-suited region proposals for classes that largely diverge from the categories in $\mathcal{D}^{train}$. Qualitative results for the SIMD and DIOR datasets are illustrated in Figure 4.

4.1. Ablation study

While the best results obtained by the proposed detector are reported in Table 1, we conduct in-depth ablation studies on the key components of our approach. Hence, we expand on (1) the choice of pre-trained backbone, (2) the classification abilities of learned DINOv2 prototypes, and (3) the impact of fine-tuning prototypes as described in Section 3.3.

Visual vs. vision-language features. Despite VLM having become popular for OVD and FSOD, a question arises when selecting a pre-trained backbone: are vision-language features superior to purely visual features? We aim to shed some light on this issue by comparing DINOv2 and CLIP representations. Furthermore, we add to our analysis two CLIP-based VLMs developed for the remote sensing domain: RemoteCLIP [21] and GeoRSCLIP [39]. To this end, we evaluate our detector on the SIMD dataset for different backbones and $N = 10$ examples per class. Results are shown in Table 2, which displays mAP50 scores for both base and novel classes. Average prototypes, as described in Section 3.1, are used on the top part, while fine-tuned prototypes, as shown in Section 3.3, are reported on the bottom. Visual features show a clear advantage over vision-language features in novel classes, even for RemoteCLIP and GeoRSCLIP. They also report strong results on base classes, despite RemoteCLIP and GeoRSCLIP achieving higher results. This can be explained by two factors: On one hand, base classes correspond to those contained in DOTA, consisting of very general and common objects in remote sensing datasets. We argue that remote sensing datasets are highly overfitted by such classes, e.g. *plane*, *small vehicle*, *ship*, *storage tank*, *tennis court*, etc. Thus, both RemoteCLIP and GeoRSCLIP have repeatedly seen the concepts related to base classes. On the other hand, novel classes consist in exceptionally rare object categories, including *propeller-aircraft*, *airliner*, *charted-aircraft*, or *stair-truck*. Hence, such fine-grained vocabulary is not known by CLIP models, which rely on image captions to learn image-text representations. These captions often lack the ability to describe all elements in the image, since a single satellite image can contain numerous instances and concepts. Thus, we argue that VLMs are limited by the granularity of image descriptions, which restricts their capabilities for FSOD on fine-grained, rare categories.

Classification abilities of DINOv2 features. The results reported in Table 1 illustrate impressive detection performance on SIMD and a considerable decline when it comes to the DIOR dataset. We hypothesize this is mainly due to one aspect; several objects in DIOR significantly differ from the types of objects in DOTA, thus the pre-trained RPN provides unsuitable proposals for such categories. More precisely, DIOR contains some classes involving land cover areas and buildings, such as *airport*, *trainstation*, *dam* or *toll-station*. These substantially differ from the concept of *object* in DOTA, which considers building and ground areas as background elements. Therefore, categories containing those elements in DIOR will be ignored as object candidates and consequently never detected. To clarify this hypothesis, we evaluate the classification abilities of the learned prototypes using their ground truth box annotations as region proposals. A strong classification performance and a decrease in the disparity between SIMD and DIOR results would indicate that the RPN is indeed a limiting factor of the approach for the DIOR dataset. Hence, we report in Table 3 the classification F-1 score and accuracy of pre-trained prototypes using bounding boxes for the SIMD and DIOR datasets on 5-shot, 10-shot and 30-shot. Furthermore, to

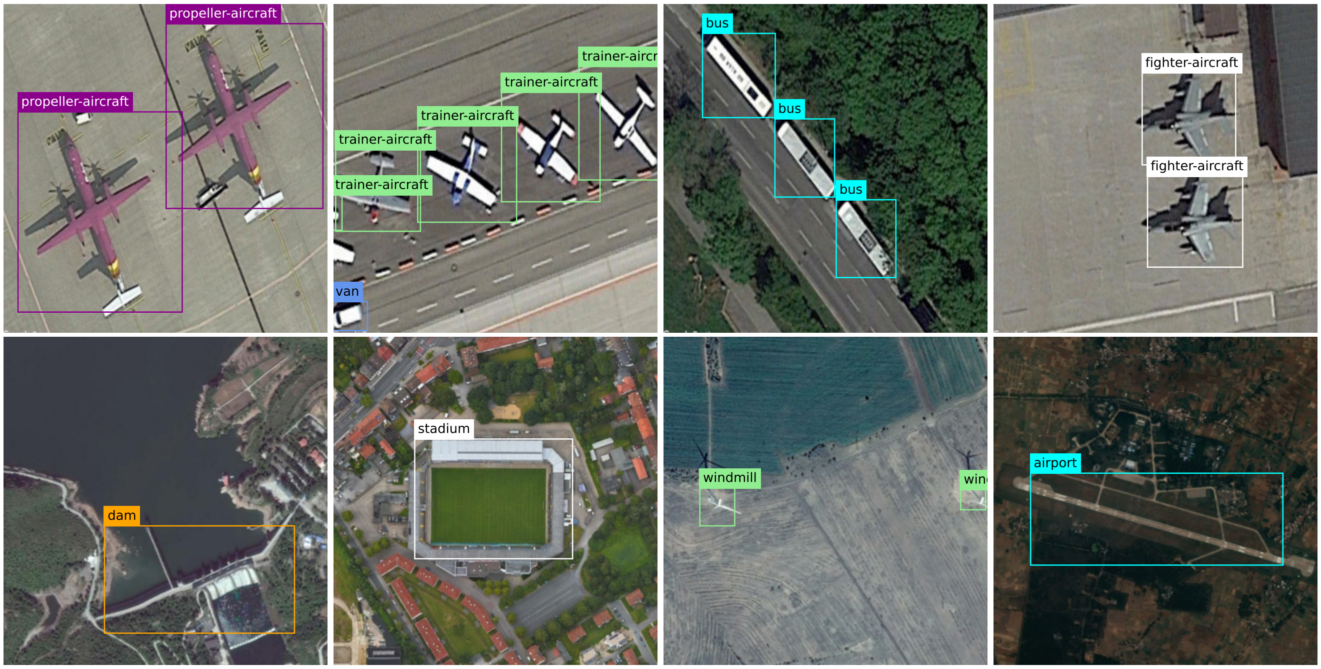

Figure 4. Illustrative qualitative results obtained by the proposed detector. Images on the top row correspond to the SIMD dataset, while images on the bottom belong to the DIOR dataset.

Backbone	Fine-tuned	Architecture	$\mathbf{c}_{novel}$	$\mathbf{c}_{base}$
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

Table 2. Performance (mAP) of different backbones for 10-shot on the SIMD dataset, including a general VLM (CLIP), remote sensing VLMs (GeoRSCLIP and RemoteCLIP) and a purely visual model (DINOv2). Prototypes without fine-tuning are shown on top, while fine-tuned prototypes are reported on the bottom. One negative example per image was used during training. Visual features show higher detection capabilities on novel classes c_{novel}, and they report strong performance in base classes c_{base} as well. DINOv2 largely outperforms RemoteCLIP on novel classes despite having fewer parameters.

minimize background information in object prototypes, we additionally initialize and fine-tune them on the same subsets using segmentation masks instead of bounding boxes. Thus, a segmentation mask is extracted for each object using the Segment Anything Model (SAM)[15] by using its ground truth bounding box as an input prompt. Results indicate an impressive classification ability for prototypes learned via both boxes and segmentation masks. To complement this analysis, we select the best performing FSOD method on DIOR, FSRW [12], and use it as RPN to pair it with the prototypical classification. FSRW re-weights the features of a pre-trained one-stage detector and thus improves the bounding box proposals with respect to the base training. As reported in Table 1, re-classifying FSRW proposals with our approach improves FSRW itself for the DIOR dataset. This ablation clarifies one aspect of our detector: selecting an RPN pre-training dataset with an object definition that better aligns with your target classes will considerably increase the performance. Alternatively, one can re-purpose FSRW as RPN if no suitable dataset is available, as shown. Lastly, the impact of fine-tuning using segmentation masks on classification remains unclear. Given the substantial computational overhead introduced by SAM, we opt to continue using exclusively bounding boxes.

Impact of prototype fine-tuning. We complement our evaluation with an ablation of the proposed prototype fine-tuning. Consequently, we report the mAP50 results for the SIMD dataset with $N = \{5, 10, 30\}$ without model fine-tuning, fine-tuning without negative examples, i.e. only fine-tuning object prototypes, and fine-tuning with 1, 5, and 10 negative examples per image, respectively. Table 4 shows the obtained results. As seen, the best scores in all

	fine-tuning type	SIMD		DIOR	
		F-1 score	Acc	F-1 score	Acc
5-shot	boxes	**57.96**	**62.43**	**59.58**	**73.35**
	masks	50.49	58.17	48.98	62.68
10-shot	boxes	**64.30**	**69.23**	60.60	75.74
	masks	63.42	68.23	**71.88**	**86.67**
30-shot	boxes	**66.88**	**68.95**	**72.23**	**84.36**
	masks	60.97	66.65	72.02	80.13

Table 3. Classification results for fine-tuned prototypes using both boxes and masks (retrieved using SAM). The F-1 score and classification accuracy are provided for the SIMD and DIOR datasets, on 5, 10 and 30-shot, respectively.

fine-tuning	# of negatives	N=5	N=10	N=30
✗	0	0.297	0.303	0.339
✓	0	**0.354**	**0.390**	**0.412**
✓	1	0.336	0.362	0.382
✓	5	0.322	0.363	0.349
✓	10	0.308	0.365	0.333

Table 4. Ablation of the proposed prototype fine-tuning approach. Different numbers of examples per class $N = \{5, 10, 30\}$ are evaluated on mAP50 for prototypes with no fine-tuning, fine-tuning only object prototypes, and fine-tuning with 1, 5 and 10 negative boxes per image, respectively.

cases correspond to prototype fine-tuning with no negative examples. Interestingly, we use one negative example per image in Table 2 experiments as we observed that CLIP backbones yield a poor performance otherwise, assigning nearly all proposals to background prototypes. We attribute this behavior to the fact that captions in satellite images often describe the land cover, thus CLIP is more familiar with the background of objects than the objects themselves. Conveniently, DINOv2 does not require negative examples, enabling a considerable speed-up of the fine-tuning process as only a few prototypes are optimized. We believe this to be due to the way class-reference prototypes are learned. When using negative examples and training both object and background prototypes altogether, we are trying to yield vectors that separately characterize objects and the background. However, region proposals will partially contain the background in addition to the object. Furthermore, DINOv2 representations not only describe local information but also the relationship with nearby patches, i.e. the background. Consequently, learning DINOv2 reference vectors that completely decouple foreground from background information might be ill-posed.

4.2. Limitations and future work.

Region proposals. Our framework can be used for two different applications: On one hand, one could seek fine-grained detection of common, general objects, e.g. vehicles or aircraft, as is the case of the SIMD dataset. This scenario is very suitable to our approach, as the RPN is capable of reliably detecting the general object, while the learned prototypes allow for fine-grained classification of its sub-categories. On the other hand, one could want to detect new, unseen categories that significantly differ from the ones in the pre-training dataset. As illustrated by the DIOR results, the RPN limits the performance of the method, as it fails to provide suitable region proposals for these challenging objects. Even though using FSRW as RPN improves the proposed bounding boxes, future work should focus on a better adaptation of RPN to novel classes using available annotations and prototypes.

Classifying boxes. Features inside region proposals contain object and background information. Hence, using all patch similarities inside the box for classification might negatively impact the results. In our experiments, averaging all similarities yielded better results than taking the maximum similarity or the top k most similar patches. Nevertheless, this could be studied in depth in further analysis.

5. Conclusion

In this article, we thoroughly explore recent ideas in open-vocabulary and few-shot object detection for remote sensing applications. More precisely, we develop a few-shot object detector that builds prototypes of objects and their backgrounds using robust features and fine-tunes them via automatic prompt engineering. Furthermore, we explore the use of visual (DINOv2) and vision-language (CLIP) representations, including two VLMs specifically tailored for the remote sensing domain. We find that visual features are largely superior to vision-language representations, as they do not have the vocabulary and/or knowledge for fine-grained remote sensing object detection. We demonstrate the capabilities of DINOv2 features to represent and classify rare objects in satellite imagery with only a handful of examples. Lastly, we compare our simple approach with other fully supervised and few-shot methods on two challenging datasets, SIMD and DIOR, for 5,10, and 30-shot. Our approach provides large improvements for the SIMD dataset, while a simple change of region proposal network allows us to beat other methods on the DIOR dataset as well.

Acknowledgements. This work was funded by AID-DGA (l'Agence de l'Innovation de Défense a la Direction Générale de l'Armement—Minitère des Armees), and was performed using HPC resources from GENCI-IDRIS (grants 2023-AD011011801R3, 2023-AD011012453R2, 2023-AD011012458R2) and from the "Mésocentre" computing center of CentraleSupélec and ENS Paris-Saclay supported by CNRS and Région Île-de-France (http://mesocentre.universite-paris-saclay.fr). Centre Borelli is also with Université Paris Cité, SSA and INSERM.

References

[1] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117:11–28, 2016. 3

[2] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016. 3

[3] Gong Cheng, Bowei Yan, Peizhen Shi, Ke Li, Xiwen Yao, Lei Guo, and Junwei Han. Prototype-cnn for few-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2022. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[5] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 2

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 1, 2

[8] Muhammad Haroon, Muhammad Shahzad, and Muhammad Moazam Fraz. Multisized object detection using spaceborne optical imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3032–3046, 2020. 3, 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[12] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 2, 5, 6, 7

[13] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multimodal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, 2023. 2, 3

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 1

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 7

[16] Mona Köhler, Markus Eisenbach, and Horst-Michael Gross. Few-shot object detection: a comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[17] Ke Li, Gong Cheng, Shuhui Bu, and Xiong You. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2337–2348, 2017. 3

[18] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. 1, 3, 5

[19] Xiang Li, Jingyu Deng, and Yi Fang. Few-shot object detection on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 3

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[21] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *CoRR*, abs/2306.11029, 2023. 2, 3, 6

[22] Xiaonan Lu, Xian Sun, Wenhui Diao, Yongqiang Mao, Junxi Li, Yidan Zhang, Peijin Wang, and Kun Fu. Few-shot object detection in aerial imagery guided by text-modal knowledge. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–19, 2023. 3

[23] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 1, 2

[24] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Je-

gou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 5

[26] Sarah Parisot, Yongxin Yang, and Steven McDonagh. Learning to name classes for vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23477–23486, 2023. 2, 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[28] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016. 3

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 2

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2, 3

[31] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[32] Stefan Wolf, Jonas Meier, Lars Sommer, and Jürgen Beyerer. Double head predictor based few-shot object detection for aerial imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 721–731, 2021. 3

[33] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 3, 5

[34] Youzi Xiao, Zhiqiang Tian, Jiachen Yu, Yinshu Zhang, Shuai Liu, Shaoyi Du, and Xuguang Lan. A review of object detection based on deep learning. *Multimedia Tools and Applications*, 79:23729–23791, 2020. 2

[35] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 1

[36] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[37] Tianyang Zhang, Xiangrong Zhang, Peng Zhu, Xiuping Jia, Xu Tang, and Licheng Jiao. Generalized few-shot object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:353–364, 2023. 3

[38] Xinyu Zhang, Yuting Wang, and Abdeslam Boularias. Detect every thing with few examples. *arXiv preprint arXiv:2309.12969*, 2023. 2, 3, 5, 6

[39] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2023. 2, 3, 6

[40] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30 (11):3212–3232, 2019. 2

[41] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1, 2

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[43] Igor Ševo and Aleksej Avramovic. Convolutional neural network based automatic object detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, 13:1–5, 2016. 3