IMPROVING FEW-SHOT OBJECT DETECTION WITH OBJECT PART PROPOSALS

Arthur Chevalley, Ciprian Tomoiagă, Marcin Detyniecki

AXA Group Opereations, GETD, Switzerland

Marc Rußwurm, Devis Tuia

EPFL, Switzerland

ABSTRACT

Few-Shot Object Detection (FSOD) allows fast adaptation of an object detection model to new classes of objects using few examples per class. This has many applications, in particular in satellite and aerial observation, as it allows learning from experts who can only annotate a few examples for new classes and helps migrate models across tasks. In this work, we present a technique to improve the performance of FSOD in remote sensing by defining a contrastive loss that utilizes parts of objects. For this, we generate, what we call, Object Parts Proposals (OPPs) on the fly for each novel class, and use them to learn more robust features with an additional contrastive objective. We observe that training with OPPs brings a consistent improvement over the state-of-the-art when evaluating on the DIOR dataset.
The code is available at https://github.com/arthurchevalley/Improving-FSOD-on-RSI-using-Sub-Parts.

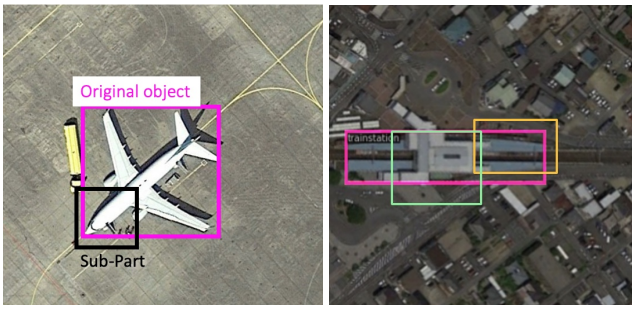

Fig. 1. Main intuition behind our Objects Parts Proposal OPP: a sub-region of an object, such as the cockpit of an airplane (left, black square), provides significant signal regarding the type, localization, and orientation of the entire object. The right image illustrates the new regions proposed by the extraction process: the original bounding box in pink, and the two sub-parts in green and yellow. The sub-parts may extend outside the original bounding box.

1. INTRODUCTION

Remotely-sensed images (RSI) have become an invaluable source of information for various applications, ranging from agriculture and forestry to urban planning and disaster management. Many RSI enabled tasks involve some kind of detection of objects on the ground, hence the development of object detection methods for RSI has seen a tremendous increase, especially since the advent of deep learning methods [1]. One of the key challenges in effectively utilizing RSI for object detection tasks is the need to identify highly specific object classes, often not among those that the base model has been trained on, often difficult to characterize without domain expertise and therefore costly to obtain in large numbers. Consequently, there is a pressing demand for object detection techniques that can operate in a low-data regime, providing accurate results even when few labeled examples are available. This would enable agile models to adapt to new scenarios, which is crucial, for example, in the insurance sector, where the nature of objects to be detected varies greatly.

Moreover, RS images present multiple sources of complexity that distinguish them from natural-view images. These complexities include significant variations in scale, resolution, and object characteristics, and they make it even more difficult for models to generalize effectively when learning from limited examples. Furthermore, the new objects of interest may already be present within the base images, albeit without annotations. This poses an additional challenge for object detection models as they must separate these unannotated objects from the background class, further highlighting the need for robust few-shot object detection methods tailored to the intricacies of RSI.

In industrial settings, the emphasis is often on identifying and detecting novel object classes rather than achieving superior performance on the base classes, which are typically abundant. Consequently, parallel deployment of two models, one specializing in the base classes and the other focusing on the specific industrial objects, is a common approach. This setup allows for improved learning on the novel classes, without hurting the overall detection performance.

In this article we present a technique to improve the performance of few-shot object detection (FSOD) in remotely-sensed images. Our contribution centers around the generation of an additional learning signal, in the form of object parts proposals (OPP). This is exploited through contrastive losses, enabling the model to better capture and discriminate variations within single instances of objects. By explicitly modeling the object structure, our proposed method offers a

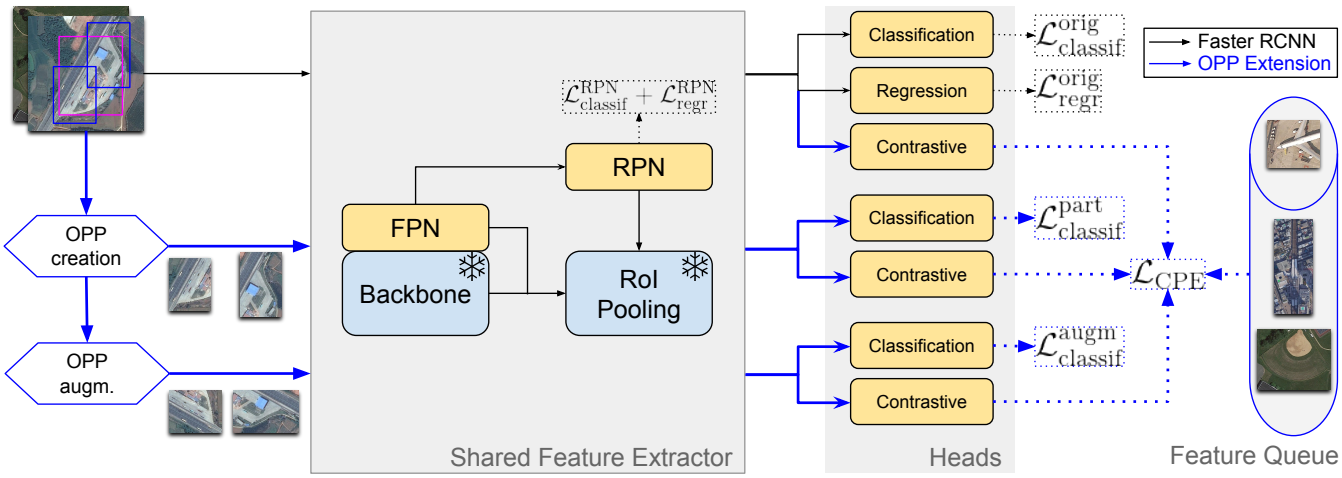

Fig. 2. Architecture of proposed model. The blue path indicates the OPP extension to classical Faster RCNN, comprising parts creation, their augmentations, and the losses given by each. The feature queue plays a major role in contrastive loss effectiveness. In the second stage, only the yellow layers are fine-tuned, while the blue ones are frozen.

viable solution to address the aforementioned limitations.

To evaluate the effectiveness of OPP, we conducted experiments on the DIOR dataset [2]. By comparing our approach against state-of-the-art methods, we observe consistent improvements on the novel classes, demonstrating the efficacy of our proposed technique in addressing the low data regime challenges in remote sensing.

2. RELATED WORK

Exploration of object-part relationships was common in pre-deep learning algorithms. For example, in SIGMA [3], a rule system for understanding aerial images, *Part-Of* relationships are manually specified. However, this avenue received little attention in the deep learning era, as object bounding boxes and class probabilities are predicted directly from the image. This can be done in a single stage, when speed is of concern, or in two stages, as introduced by the seminal Faster RCNN [4]. This model combines a Region Proposal Network (RPN) as a first stage for accurate object localization, with a Convolutional Neural Network for refinement and classification, sharing the weights between the two.

Many approaches for few-shot object detection are based on Faster RCNN. Among the performant ones, the two-stage fine-tuning approach (TFA) [5] is the simplest, training the vanilla model once on base classes and fine-tuning specific layers on novel classes. This works well for natural view images but cannot deal with the challenges of RSI. In this domain, the few-shot object detection model (FSODM) [6] achieves promising results using meta-learning with a YOLO architecture. These were improved by Shared Attention Module [7], who adopt a fine-tuning strategy, enhanced with multi-attention maps that are shared between the two stages.

3. METHOD

Our proposed method, OPP, follows a two-stage training protocol as in previous works in FSOD for remote sensing [6, 7]. First, we train the base Faster RCNN on an abundant dataset of base classes. Then, we fine-tune certain parts of the architecture on a dataset composed of base and novel classes. We implement our model using the MMDetection framework [8].

The OPP model architecture consists of three parts, and is illustrated in Fig. 2. The first and core element is a traditional Faster RCNN [4], made of a feature extractor, a Region Proposal Network (RPN), a Feature Pyramid Network (FPN) [9] and finally two heads: one for classification of boxes using the cross entropy loss $\mathcal{L}_{\text{clasif}}$, and one for regression of their coordinates using smooth L_1 loss $\mathcal{L}_{\text{regr}}$:

$$\mathcal{L}_{\text{classif}}(y_i, \hat{y}_i) = -y_i \log(\hat{y}_i) \tag{1}$$

$$\mathcal{L}_{\text{regr}}(t_i, v_i) = \sum_{d \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_i^d, v^d). \tag{2}$$

Here, y_i and t_i are the ground truth class and coordinates of a box i, respectively, and $\hat{y}_i$, v_i are the predicted counterparts.

3.1. Object Parts Proposals(OPPs)

The second part of the architecture, and our main contribution, is a branch exploiting the Object Parts Proposals. It stems from the observation that a part of an object is often sufficient to locate it and classify it. As illustrated in Fig. 1, the cockpit suffices to classify an airplane and have an idea of the airplane's size and orientation. Therefore, the representation of this region should be similar to that of the entire airplane.

To achieve this, we implement a data loader pipeline that works on the fly to generate two sub-regions for each ground

truth region of an image. They are of size $w_p \times h_p$, such that $w/3 < w_p < w, \quad h/3 < h_p < h$ and are centered randomly inside the original box surrounding the reference object. This means that the sub-regions can include areas outside the original box ; however, we limit this to be small. Furthermore, we augment the parts using transformations such as flipping, 90° rotations, and brightness changes. The generated parts, together with their augmentations, increase the number of representations of each class, which is critical in few-shot learning. In addition, OPPs are more powerful than per-image data augmentation, as they allow modelling the object-part relationship in the feature space.

To learn from this new data source, we employ three losses. First, we propagate the label from the object to the parts, and classify them accordingly ($\mathcal{L}^{\text{part}}_{\text{classif}}$). The loss of the part classifier is the same as the object classifier (Eq. (1)), only applied to representation extracted from the bounding box describing the part. Furthermore, we repeat the same procedure for an augmentation of the part image ($\mathcal{L}^{\text{augm}}_{\text{classif}}$). Finally, we encourage the similarity of object-part features using the Contrastive Proposal Encoding (CPE) loss [10]

$$\mathcal{L}_{\text{CPE}} = \frac{1}{N} \sum_{i=1}^{N} f(u_i) \cdot L_{z_i} \tag{3}$$

where z_i is the RoI feature encoded by the contrastive head and u_i denotes the Intersection-over-Union. The term

$$L_{z_i} = \frac{-1}{N_{y_i} - 1} \sum_{i=1, j \neq i}^{N} \mathbb{I}_{y_i = y_j} \cdot \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{I}_{j \neq k} \exp(\text{sim}(z_i, z_j)/\tau)} \tag{4}$$

calculates a softmax-normalized ($\frac{\exp(\cdot)}{\sum \exp(\cdot)}$) cosine-similarity (sim) between samples in the batch, indexed by j and i. Only samples within batch where bounding boxes describe the same class ($\mathbb{I}_{y_i = y_j}$) are considered. The total number of samples of the same class is denoted by N_{y_i} and the temperature hyper-parameter τ determines the flatness or tolerance of the softmax normalization, as in the original InfoNCE loss [11].

More generally, this equation uses the label information to group instances of the same class, while increasing the distance between parts extracted from objects pertaining to different classes. In the context of the CPE loss, we would like to draw attention to the important function of parameter $\boldsymbol{f}(u_i)$ in Eq. (3), representing the Intersection-over-Union between proposals and ground truth boxes. This reduces the impact of badly predicted Regions of Interest (RoI).

3.2. Feature queue

As for general contrastive self-supervised learning [12], learning strong features contrasting parts within and between objects depends on the abundance and diversity of the negative examples. To ensure such diversity, in the third part of the architecture we build a queue of features, where features representing parts of objects are stored and reused when necessary. This is inspired from MoCo v2 [12], where the usage of a features queue promoted stability in learning and effective contrasting between several classes at once. In other words, this provides the model a broad panel of negative and positive samples, so it can effectively form and distinguish the groups, increasing its robustness to intra-class diversity. Without the queue, learning would be limited to current batch's images, which is reduced in size due to the memory requirements of contrastive objectives.

Contrary to MoCo v2, our algorithm has access to labels, which can be used to strengthen the queue formation. To this end, we enforce that the queue contains at least five examples of each class, effectively overcoming the class imbalance inside the queue. This promotes proper group formation in the contrastive branch.

4. EXPERIMENTS AND RESULTS

We tested OPP on the DIOR dataset, which offers a considerable number of diverse classes, thus allowing us to effectively simulate a FSOD scenario. We adopt the same data split strategy of related works [6, 7], thereby identifying five *novel* classes for evaluation purposes. These are representative of real-life scenarios: i) small size (*windmill*), ii) large size (*train station*), iii) similarity with base dataset (*airplane, tennis court* vs. airport, basketball court), iv) cluttered environments (*train station*).

Table 1 reports our results on the validation set, employing the mean average precision (mAP) as the evaluation metric. It facilitates the comparison with a strong baseline, described below, and the current state-of-the-art, as reported in Shared Attention Module paper [7] and FSDOM [6].

To establish a robust baseline, we train a vanilla Faster RCNN model, fine-tuning all but the convolutional backbone layers (see Fig. 2). We always use a single image per base class during training, both in the baseline as in OPP. Our findings reveal that this imbalanced fine-tuning approach enhances the performance of novel classes while compromising the base ones. This trade-off is acceptable in the common scenario where both models can be run in parallel.

In the first experiment, we exclusively employ Object Parts Proposals (OPP) during the fine-tuning process on the novel classes (**OPP-FT**). This technique yields a substantial overall improvement, leading to a 5% increase in mAP compared to the baseline approach (Table 1 row **a)**), which is powered by large enhancements in three of the five classes. For the remaining two, we can attribute their underperformance to: i) the overlap between *tennis court* novel class and *basketball court* base class, both consisting of similar sizes, coloring, and often neighboring each other, ii) the small size of *windmill* hindering creation of meaningful sub-parts.

In the subsequent experiment, we incorporate OPP during both the base training and the fine-tuning stages (**OPP-Full**).

Table 1. Detection performance (mAP) of OPP variants and current state-of-the-art. OPP performs best on novel class average (row **a)**), outperforming the others by over 5% for 10 and 20 shots scenarios. While **OPP-FT** underperforms on base classes (row **b)**), applying our technique also during base training (**OPP-Full**) recovers most of the performance.

	Baseline			**FSODM**			**Shared Attn. Mod.**			**OPP-FT**			**OPP-Full**		
↓ class; shots →	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
Baseball field	84.3	**88.2**	90.5	27.0	46.0	50.0	73.0	78.0	81.0	**85.0**	85.7	**91.1**	84.6	87.2	90.4
Airplane	31.0	78.5	82.8	9.0	16.0	22.0	53.0	66.0	67.0	**59.6**	**84.5**	**88.1**	50.0	80.9	84.3
Tennis court	51.2	44.7	61.7	57.0	60.0	66.0	49.0	65.0	**70.0**	40.7	55.7	64.4	**58.2**	**65.5**	**69.5**
Train station	1.9	13.0	9.5	**11.0**	14.0	16.0	2.5	3.5	5.8	4.8	**18.0**	**23.9**	4.2	14.2	20.0
Windmill	0.6	7.5	8.0	**19.0**	24.0	29.0	14.0	**26.0**	**30.5**	1.8	16.8	15.2	1.5	14.3	8.4
a) **Mean Novel**	33.78	46.4	50.50	25.0	32.0	36.0	38.30	47.30	50.90	38.39	52.13	**56.50**	**39.69**	**52.4**	54.51
b) Mean base		22.4			54.0			N/A			30.0			40.5	

By doing so, we expect the model to benefit from improved clustering of base class features, consequently facilitating the fine-tuning process. Indeed, this results in a marginal improvement in overall score, while recovering the performance in *tennis court* class. Furthermore, **OPP-Full** exhibits exceptional gains on the base classes, more than doubling the performance compared to the baseline. In this way, the OPP extension proves significant improvements on novel classes, while remaining competitive on the base ones.

5. CONCLUSION

In conclusion, our approach introduces promising advancements to the field of few-shot object detection for remotely sensed images by incorporating Object Parts Proposals (OPP) during fine-tuning and, optionally, base training. This integration yields notable improvements in performance across visually diverse classes, highlighting the efficiency and potential of our method in enhancing FSOD tasks on the DIOR dataset.

References

[1] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.

[2] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.

[3] T. Matsuyama and V. Hwang, *SIGMA*. Springer, 1990.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Neural Inf. Processing Syst.*, vol. 28, 2015.

[5] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," in *ICML*, 2020.

[6] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Trans. on Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[7] X. Huang, B. He, M. Tong, D. Wang, and C. He, "Few-shot object detection on remote sensing images via shared attention module and balanced fine-tuning strategy," *Remote Sens.*, vol. 13, no. 19, 2021.

[8] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[9] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[10] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," in *CVPR*, 2021.

[11] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[12] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.