IMPROVING FEW-SHOT OBJECT DETECTION WITH OBJECT PART PROPOSALS

Arthur Chevalley, Ciprian Tomoiagă, Marcin Detyniecki

AI Research
GETD, AXA

Marc Rußwurm, Devis Tuia

ECEO laboratory
EPFL, Switzerland

Few-Shot Object Detection (FSOD) allows fast adaptation of an object detection model to new classes of objects using few examples per class [1]. This has several applications across industry and academia: it allows learning from experts who can only annotate a few examples for new classes and helps migrate models across tasks. As the base model remains available for use, it is critical to have high performance in the novel classes. In this work, we present a technique to improve the performance of FSOD in remote sensing by learning about parts of objects with contrastive losses. Object-Part Proposals (OPP) follow the line of two-stage fine-tuning adaptation [2] and improve it with a custom branch dedicated to parts of objects. We follow the protocol of previous works in FSOD for remote sensing [3, 4, 5] and evaluate OPP on the DIOR dataset [6]. We observe a consistent improvement over the state-of-the-art on the novel classes. Our code will be made available after publication.

1. METHOD

Our main contribution consists of a part proposal module that extends the traditional Faster-RCNN + FPN architecture [7, 8] and forces the detector to learn more generic representations for the novel classes, which lead to more detections (Fig. 1). The module adds two smaller bounding boxes for each novel object, $b^{\text{obj}} \rightarrow b^{\text{part}}_{1,2}$, which the network should classify the same as their parent bounding boxes using a cross-entropy loss $\mathcal{L}_{\text{cls}}(b^{\text{part}}, b^{\text{obj}})$. The part-bounding boxes partially overlap the ground-truth box, thus capturing at the same time a part of the object and part of the surrounding context.

In addition, we follow the intuition that in remote sensing an object part is similar to the whole. Therefore, we encourage the representations of the parts to be similar to that of the whole object by using a Contrastive Proposal Encoding loss [9], $\mathcal{L}_{\text{CPE}}(\boldsymbol{f}(b^{\text{part}}), \boldsymbol{f}(b^{\text{obj}}))$. This acts on the features $\boldsymbol{f}(\cdot)$ of parts and objects; it increases their similarity when they are of the same class, and spaces them apart when they differ. Furthermore, we generalize these representations by applying data augmentation specific to the part proposal branch, and we push the augmented parts $b^{\text{aug}}_{1,2}$ to be similar to the non-augmented ones, as well as their parent objects. A features queue, which is continually updated as the parts are processed by the model, is also added in order to provide enough negative examples and produce a more generic feature space.

Our final training objective is:

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}_{\text{original}} + \frac{1}{2}(\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{CPE}}), \quad \text{where} \\
\mathcal{L}_{\text{original}} &= \mathcal{L}_{\text{cls}}(b^{\text{obj}}, b^{\text{GT}}) + \mathcal{L}_{\text{regression}}(b^{\text{obj}}, b^{\text{GT}}), \\
\mathcal{L}_{\text{cls}} &= \mathcal{L}_{\text{cls}}(b^{\text{part}}, b^{\text{obj}}) + \mathcal{L}_{\text{cls}}(b^{\text{aug}}, b^{\text{obj}}), \\
\mathcal{L}_{\text{CPE}} &= \mathcal{L}_{\text{CPE}}(\boldsymbol{f}(b^{\text{part}}), \boldsymbol{f}(b^{\text{obj}})) + \mathcal{L}_{\text{CPE}}(\boldsymbol{f}(b^{\text{aug}}), \boldsymbol{f}(b^{\text{obj}})) + \mathcal{L}_{\text{CPE}}(\boldsymbol{f}(b^{\text{part}}), \boldsymbol{f}(b^{\text{aug}})).
\end{aligned}$$

2. EXPERIMENTS & RESULTS

Setup We follow the experimental setup of the current state-of-the-art, the Shared Attention Module (SAM) [5], using the same split of DIOR training dataset: five novel objects (baseball field, airplane, tennis court, train station, windmill), and the remaining 15 objects as base training set. For each novel class, we pick the k images which have the fewest object instances in order to minimize the amount of information given to the model. In the first phase, we train the original Faster-RCNN on the large base set. Then, we freeze all layers except the head, the Feature Pyramid Network (FPN), and the Region Proposal

Correspondance to Ciprian Tomoiagă

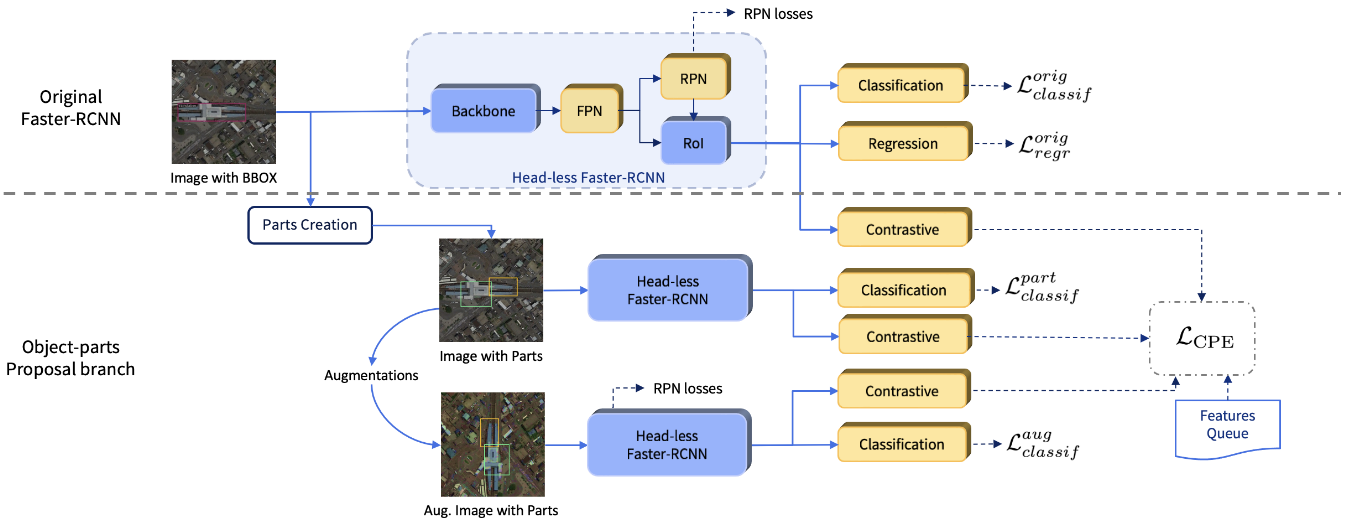

Fig. 1. Architecture of the model. Yellow layers are fine-tuned, while blue ones are frozen.

Class	**Ext. FT TFA**			**FSODM**			**Shared Attention Module**			**OPP (ours)**		
	5 Shots	10 shots	20 shots	5 Shots	10 shots	20 shots	5 Shots	10 shots	20 shots	5 Shots	10 shots	20 shots
Baseballfield	79.4	88.1	90.1	27.0	46.0	50.0	73.0	78.0	81.0	**86.6**	**89.8**	**91.3**
Airplane	13.8	53.1	73.0	9.0	16.0	22.0	53.0	66.0	67.0	**54.8**	**77.3**	**83.4**
Tenniscourt	49.7	63.1	62.3	57.0	60.0	66.0	49.0	65.0	**70.0**	**57.6**	**70.2**	62.9
Trainstation	0.7	2.2	6.7	11.0	14.0	16.0	2.5	3.5	5.8	**6.8**	**16.2**	**25.0**
Windmill	0.3	1.7	2.6	**19.0**	24.0	29.0	14.0	**26.0**	**30.5**	0.6	9.9	10.8
Mean	28.78	41.64	46.94	25.0	32.0	36.0	38.30	47.30	50.90	**41.28**	**52.68**	**54.68**

Table 1. Performances per class in terms of novel mAP.

Network (RPN). We add the OPP and fine-tune on the novel classes. Finally, we evaluate on the full validation set. Our code and experiments employ the open-source MMDetection framework [10], using ResNet50 as backbone for the original model.

We track performance using mean average precision (mAP) and compare our approach to: i) TFA [2], ii) FSODM [3], iii) Shared Attention Module (SAM) [5]. As the latter two use the exact same benchmark as us, we report their numbers as given in the original papers, noting that for SAM the mAP values are estimated from their plots. However, TFA was not used in the context of RSI, therefore we train and fine-tune it following the protocol described above; we reference it as *Ext. FT TFA*.

Observing the results in Table 1, we observe that our proposed object-part mining outperforms all previous approaches in all k-shot settings by 3-5%. Looking in more detail at performance per category, we note that our approach is significantly better on categories of airplane, tennis court and train station, performs on par with TFA for baseball field, and can be improved for detection of windmills. We suppose the latter is due to the chosen size of part bounding boxes, as they are larger than ground truth of windmill objects. In the full paper we will provide more insights and comparisons based on other splits of the DIOR dataset, i.e. with different sets of novel classes, as well as ablation studies for different components of our contrastive losses.

In summary, we believe that this work contributes to the state-of-the-art in Few-Shot Object Detection methods where novel classes are learned efficiently with few annotations. This accommodates the open-world nature of objects present in remote sensing imagery where training methods to a fixed set of classes a priori is often too restrictive for many real-world use-cases.

3. REFERENCES

[1] Sixu Liu, Yanan You, Haozheng Su, Gang Meng, Wei Yang, and Fang Liu, "Few-shot object detection in remote sensing image interpretation: Opportunities and challenges," *Remote Sensing*, vol. 14, no. 18, 2022.

[2] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu, "Frustratingly simple few-shot object detection," in *International Conference on Machine Learning (ICML)*, July 2020.

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

[4] Zhitao Zhao, Ping Tang, Lijun Zhao, and Zheng Zhang, "Few-shot object detection of remote sensing images via two-stage fine-tuning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[5] Xu Huang, Bokun He, Ming Tong, Dingwen Wang, and Chu He, "Few-shot object detection on remote sensing images via shared attention module and balanced fine-tuning strategy," *Remote Sensing*, vol. 13, no. 19, 2021.

[6] 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*, vol. 159, pp. 296–307, 2020.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, Curran Associates, Inc.

[8] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 7352–7362.

[10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.