Amélioration de la détection d'objets few-shot à travers une analyse de performances sur des images aériennes et naturelles

Pierre LE JEUNE[1, 2], Anissa MOKRAOUI[1]

[1]Laboratoire de Traitement et Transport de l'Information, Université Sorbonne Paris Nord
99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France

[2]COSE, 5 bis Route de Saint-Leu, 95360 Montmagny, France
pierre.lejeune@edu.univ-paris13.fr, anissa.mokraoui@univ-paris13.fr

Résumé – L'entraînement de modèles de détection d'objets requiert l'accès à de grandes bases de données annotées. Étant donné le coût élevé des annotations, en particulier pour la détection, celles-ci ne sont souvent pas disponibles pour des applications réelles. La détection d'objets *few-shot* (FSOD) tente de combler cette lacune en entraînant des modèles à partir de peu de données annotées. Les méthodes existantes se concentrent sur des images naturelles, et notamment les jeux de données MS COCO et Pascal VOC. Lorsque l'on applique naïvement ces méthodes sur des images aériennes, les performances sont moins bonnes que sur des images naturelles. Une analyse détaillée des performances et des jeux de données utilisés est réalisée afin de comprendre l'origine de ces différences. Des adaptations compatibles avec la plupart des méthodes FSOD sont finalement proposées afin d'améliorer les performances sur les images aériennes, permettant ainsi de gagner en moyenne 5% de performance.

Abstract – Object detection models require a large amount of annotated data during training, making their deployment for real-world tasks difficult. Few-Shot Object Detection (FSOD) aims to solve this shortcoming by training object detection models from limited annotated data. However, existing methods mostly focus on natural images such as MS COCO and Pascal VOC datasets. In this paper, we study FSOD on aerial images. At first glance, performance seems to decrease compared to natural images with similar datasets. We perform an in-depth analysis to understand the performance discrepancies between natural and aerial images. In the light of this analysis, we propose several improvements to boost the detection quality on aerial images. These modifications increase the mAP by approximately 5% on average.

1 Introduction

L'apprentissage *few-shot* vise à adapter une tâche spécifique sur de nouvelles classes à partir de peu de données annotées. Cet apprentissage est, en général réalisé en deux phases successives : un entraînement de base avec des classes pour lesquelles on dispose de suffisamment de données puis un affinage sur les classes dites nouvelles. La détection d'objets *few-shot* (FSOD) s'inscrit dans ce paradigme et tente de détecter des objets appartenant à des classes pour lesquelles peu d'annotations sont disponibles. Cela est d'autant plus intéressant pour la détection d'objets, car les modèles existants demandent des quantités conséquentes de données pour atteindre des performances satisfaisantes. De plus, les annotations sont très coûteuses pour la détection, en particulier dans le cas des images aériennes dans lesquelles se trouvent de nombreux objets de petite taille. Une première approche dans cette direction est proposée par [1]. Depuis, un certain nombre de méthodes ont été mises au point, s'en inspirant de près ou de loin. Cependant, la plupart de ces méthodes se concentrent sur des images naturelles provenant des jeux de données Pascal VOC [2] et MSCOCO [3]. Rares sont les contributions qui s'intéressent à des images aériennes. Cette étude s'intéresse en particulier à deux d'entre elles. La première étend la méthode de pondération des caractéristiques (*feature reweighting* – FRW) proposée par [4] à plusieurs échelles. La seconde propose un mécanisme d'attention adaptatif [5] qui extrait les informations contenues dans les images de *support* (c.a.d. les quelques exemples disponibles pour chaque nouvelle classe). Ces informations sont ensuite utilisées afin de conditionner la détection dans les images de *query*. Ces deux méthodes sont au cœur de cette étude et seront respectivement abrégées FRW et SAA. Elles sont testées sur trois jeux de données distincts : DOTA [6] et DIOR [7], constitués d'images aériennes et Pascal VOC comme référence sur les images naturelles. La Figure 1 présente les performances des deux méthodes sur chaque jeu en comparant aux performances d'un modèle de détection classique, c.a.d. sans *few-shot* et entraîné avec suffisamment de données pour chaque classe (baseline), en l'occurrence FCOS [8]. Comme on ne peut directement comparer les performances sur des jeux de données différents, on compare les écarts avec la *baseline*. On observe une baisse mesurée des performances sur les classes de base, mais une grande diminution sur les classes nouvelles. Cependant, cette diminution est beaucoup plus grande pour les images aériennes. Afin de comprendre cette différence de comportement, une analyse détaillée des jeux de données et des

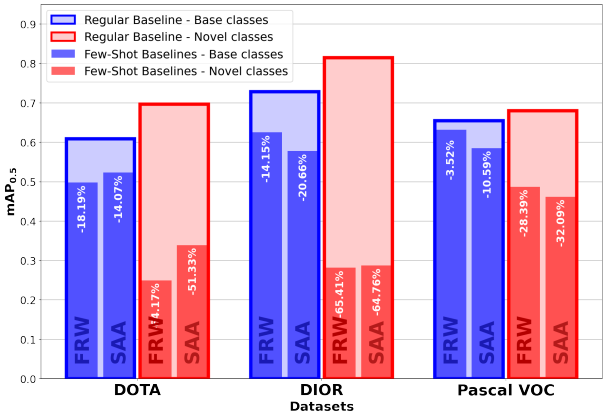

FIGURE 1: Comparaison des performances de la *baseline* classique et de deux méthodes *few-shot* : FRW [4] et SAA [5] sur trois datasets, DOTA, DIOR et Pascal VOC.

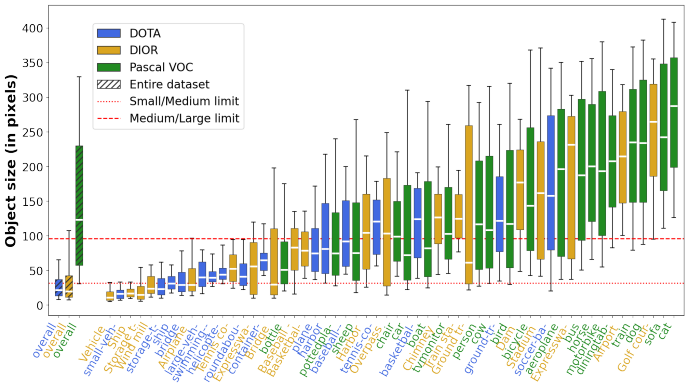

FIGURE 2: Boxplot représentant la taille des objets dans les datasets DOTA, DIOR et Pascal VOC. Sur la gauche, les boîtes représentent la distribution des tailles dans les datasets entiers. Sur la droite, les distributions sont séparées par classes.

performances est réalisée, individuellement sur chaque classe. DIOR et DOTA contiennent des objets sensiblement plus petits que Pascal VOC. Il est généralement plus difficile de détecter des objets de petite taille, mais ici, ce n'est pas le cas. Les performances de la *baseline* sont proches sur les trois jeux. En revanche, les écarts de performances des méthodes *few-shot* avec la *baseline* diffèrent beaucoup entre les images naturelles et aériennes. Il est probablement plus difficile d'extraire des informations concernant la classe d'un objet lorsque celui-ci est petit. Ainsi, le conditionnement des méthodes FSOD pour la détection de petites classes est de moins bonne qualité. Finalement, des améliorations sur la méthode d'extraction d'information des images de support sont proposées afin de résoudre ces problématiques.

2 Analyse de performances sur des images naturelles et aériennes

Cette étude se concentre sur les performances des méthodes FSOD sur trois jeux de données distincts : DOTA, DIOR et Pascal VOC. Le but étant de comprendre pourquoi elles performent mieux sur des images naturelles. D'abord, une analyse statistique est réalisée sur la taille des objets dans chaque jeu. Ensuite, ces statistiques sont mises en relation avec les performances sur chaque classe.

2.1 Taille des objets

DOTA, DIOR et Pascal VOC sont trois datasets à première vue similaires. Ils contiennent un nombre de classes relativement proche (16, 20 et 20 respectivement) et ont un nombre d'objets du même ordre de grandeur. En revanche, lorsque l'on s'intéresse à la taille de ces objets, des différences notables sont visibles, comme le montre la Figure 2. DOTA et DIOR ont globalement des objets beaucoup plus petits que Pascal VOC. De plus, il existe une plus grande variété de tailles dans les jeux d'images aériennes. DOTA et DIOR possèdent des classes petites, dont les objets mesurent moins de 32 pixels de large et des objets grands (largeur supérieure à 96 pixels). Pascal VOC au contraire n'a que des classes de tailles moyennes et larges. Cette classification, petits, moyens et grands objets a été introduite dans MS COCO et sera utilisée dans la suite de cette étude. Cette grande variance dans la taille des objets dans DOTA et DIOR nécessite une plus grande robustesse vis-à-vis de la taille des objets.

2.2 Comparaison des performances sur DOTA, DIOR et Pascal VOC

Deux méthodes FSOD ont été choisies pour faire cette comparaison : FRW [4] et SAA [5]. Afin d'éviter des différences dans les détails d'implémentation, ces deux méthodes ont été réimplémentées au sein du *framework* proposé par [9]. Ce *framework* permet d'implémenter différentes méthodes FSOD en utilisant une architecture de base identique. Cela permet de mieux comparer différentes méthodes. De plus, le *framework* permet également d'entraîner des détecteurs sans *few-shot* en conservant le plus de paramètres fixés. L'entraînement des méthodes FSOD est fait en suivant une stratégie épisodique comme proposée par [1] : à chaque épisode un ensemble de support est sélectionné contenant un sous-ensemble des classes disponibles. Les images de *support* sont ensuite utilisées pour réaliser la détection dans l'ensemble de *query*, sur lequel le modèle est optimisé.

La *baseline* classique atteint des performances relativement bonnes et similaires sur les trois datasets. À partir de cela, il est légitime de penser que les méthodes FSOD atteindront des performances moins élevées, mais similaires. Ce n'est en réalité pas le cas (voir Figure 1). Les performances sur les jeux d'images aériennes sont bien moins bonnes que sur Pascal VOC, relativement à la *baseline*. Cela est d'autant plus vrai pour les classes nouvelles, l'objectif final de la FSOD. L'analyse des datasets a montré une grande différence dans la taille des objets. De plus, pour DOTA et DIOR, des différences importantes entre les classes existent. Il convient donc d'étudier

les performances sur chaque classe individuellement. Cela est représenté sur la Figure 3. À gauche, la *mean Average Precision* (mAP) est représentée par classe en fonction de la taille moyenne des objets des classes. Les performances de la *baseline* et de FRW sont visibles, mais pour FRW, une distinction est faite entre classes de base (en bleu) et classes nouvelles (en rouge). Il semble que les performances de détection augmentent avec la taille moyenne au sein des classes, à la fois pour la *baseline* et FRW. Ensuite, les performances de FRW sur les classes nouvelles sont toujours moins bonnes que la *baseline*. Cela n'est pas surprenant, car peu d'exemples ont été vus par le modèle pour ces classes. Pour les classes de base, on observe des performances inférieures à la baseline pour les petites classes, mais supérieures pour les grandes classes. Afin de mieux visualiser ce phénomène, la partie droite de la Figure 3 représente l'écart entre FRW et la *baseline* en fonction de la taille moyenne des classes. Ainsi, on observe une tendance claire : utiliser des exemples de supports est bénéfique pour les grandes classes, mais détériore les performances pour les petites classes. Cela est également visible pour les classes nouvelles. Pour Pascal VOC en revanche, cette tendance est moins marquée, car les classes sont considérablement plus grandes.

Pour résumer, les petits objets sont plus difficiles à détecter de manière générale, mais dans le cas du *few-shot*, il est également plus difficile d'extraire des informations sur leur classe, ce qui rend plus difficile le conditionnement du modèle à partir des exemples. Il convient donc d'améliorer la méthode d'extraction des exemples de support en se concentrant sur les petits objets. Dans cette direction, nous proposons des adaptations dans la section 3.

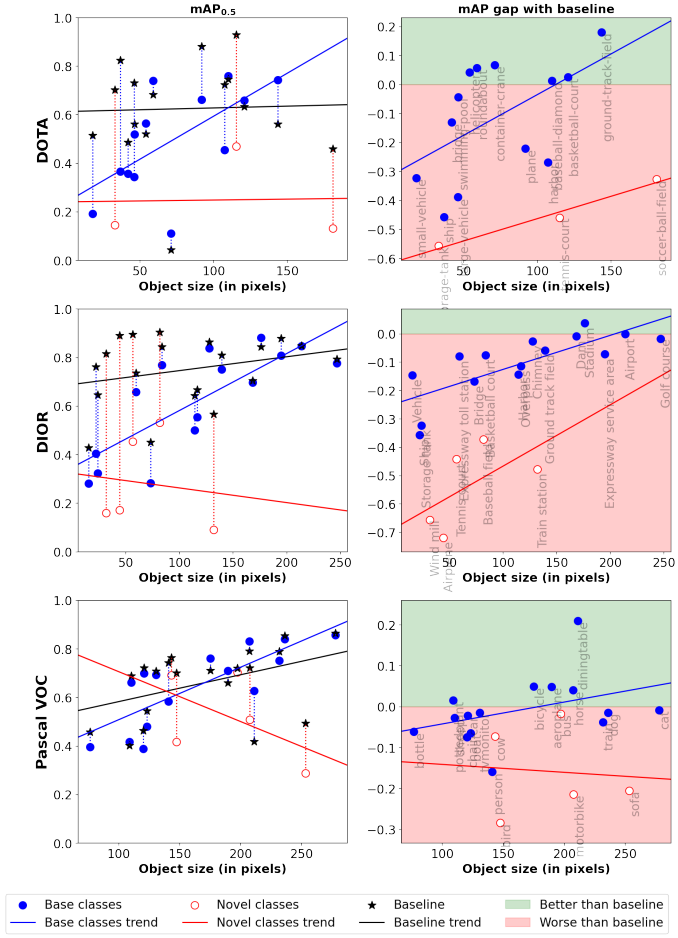

FIGURE 3: Comparaison des performances entre FRW (points bleus et rouges) et de la *baseline* classique (étoiles noires) sur trois datasets : DOTA, DIOR et Pascal VOC. (**gauche**) mAP par classes pour FRW et la *baseline*. (**droite**) écart de performance par classe entre FRW et la *baseline*.

3 Extraction sémantique des supports

La section 2.2 met en avant un défaut dans la stratégie d'extraction d'information des images de support. Il est difficile d'extraire des informations pertinentes à partir de petits objets. Dans les deux baselines *few-shot*, l'objet d'intérêt dans une image de support est découpé, puis comblé avec des zéros jusqu'à obtenir une image de 128×128 pixels (méthode appelée **default**). Les objets dépassant cette taille sont simplement redimensionnés. Lorsque l'on traite des petits objets, il n'est probablement pas optimal d'utiliser des zéros pour combler l'espace dans l'image. Cela dilue les caractéristiques de l'objet et rend ainsi la détection plus difficile. Une alternative serait de découper une région de 128×128 pixels autour de l'objet (**no-padding**). Mais là encore, les caractéristiques des petits objets sont dominées par le fond de l'image. Plusieurs autres méthodes sont donc proposées et testées.

Reflection : à la place du *zero-padding*, on peut répéter la partie découpée dans toutes les directions. Ainsi, les caractéristiques de l'objet ne sont plus dominées par le fond de l'image. **Same-size** : afin d'éviter les inconvénients des méthodes basées sur le *padding*, on redimensionne l'objet en 128×128 pixels, peu importe sa taille, mais en préservant le rapport hauteur/largeur. **Mixed** : il s'agit d'une méthode hybride utilisant *default*, pour les petits objets et *same-size* pour les moyens et grands objets. Ces stratégies sont comparées dans le Tableau 1. On peut noter que la stratégie qui fonctionne le mieux pour les classes nouvelles est *same-size*. Cependant, ce gain de performance est surtout visible pour les classes moyennes et grandes. Dans le cas des petits objets, le redimensionnement peut être néfaste, car les caractéristiques des objets

	Base classes				Novel classes			
	Mean	Small	Medium	Large	**Mean**	Small	Medium	Large
Default	0.237	**0.099**	0.261	0.254	0.132	0.034	0.132	0.178
No padding	0.243	0.074	0.281	0.240	0.136	0.034	0.115	0.245
Same size	0.238	0.085	0.271	0.241	**0.153**	0.030	**0.168**	**0.300**
Reflection	**0.247**	0.086	**0.282**	0.253	0.128	**0.048**	0.139	0.246
Mixed	**0.247**	0.079	0.281	0.247	0.142	0.030	0.124	0.285

TABLE 1: Comparaison de performance des différentes stratégies d'extraction proposées. Les performances sont mesurées comme dans [3] et rapportées séparément selon la taille des objets : petits (S), moyens (M) et grands (L).

	DOTA				DIOR					Pascal VOC			
	FRW		SAA		FRW			SAA		FRW		SAA	
	Baseline	Ours	Baseline	Ours	Baseline	Ours	[4]	Baseline	Ours	Baseline	Ours	Baseline	Ours
Base classes	**0.495**	0.485	**0.523**	0.467	**0.625**	0.615	0.540	0.578	**0.618**	**0.647**	0.610	**0.585**	0.531
Novel classes	0.283	**0.371**	0.339	**0.351**	0.282	**0.356**	0.320	0.287	**0.334**	0.522	**0.549**	0.462	**0.488**

TABLE 2: $mAP_{0.5}$ avec 10 shots sur les trois datasets DOTA, DIOR et Pascal VOC. Pour chaque dataset et chaque méthode, les performances obtenues avec les améliorations proposées sont comparées à la *baseline*.

redimensionnés sont très différentes de celles de l'objet initial. Afin de résoudre cela, une stratégie mixte a été proposée, mais celle-ci ne performe pas mieux que *resize*. L'utilisation de deux méthodes différentes en fonction de la taille des objets peut conduire à des différences importantes dans les caractéristiques des objets lorsque leur taille est proche de la limite. Cela peut induire le modèle en erreur pendant l'entraînement. Bien que *same-size* soit relativement simple, c'est la meilleure stratégie pour les nouvelles classes.

En complément de cette meilleure stratégie d'extraction d'information, plusieurs augmentations de données ont été utilisées pour obtenir une meilleure robustesse dans les détections. Bien sûr, de nombreuses augmentations existent déjà pour les images, mais peu sont utilisées dans le cadre de la détection d'objets. En effet, certaines augmentations risquent de masquer certains objets, les rendant inexploitables pendant l'entraînement. Deux méthodes d'augmentation ont notamment été adaptées : **random cut-out** et **random crop-resize**. Dans les deux cas, l'application de ces méthodes peut conduire à des images sans objets visibles. Les domaines d'application de ces méthodes sont donc restreints afin d'éviter cela. Le *cut-out* est appliqué sur les objets plutôt que sur l'image entière et le *crop-resize* choisit aléatoirement une zone qui contient au minimum un objet. En plus de cela, des effets miroirs et des altérations de couleurs sont appliqués. L'utilisation de ces augmentations conduit également à des détections de meilleure qualité, en particulier pour les classes nouvelles.

4 Résultats et conclusion

Afin de valider les résultats de la section 3, des expériences sont réalisées de manière systématique sur DOTA, DIOR et Pascal VOC avec les deux méthodes *few-shot* FRW et SAA. Les résultats de ces expériences sont disponibles dans le Tableau 2. Les améliorations apportées sont bénéfiques pour chacune des méthodes et sur chaque dataset. La méthode proposée surpasse même l'état de l'art [4] sur le dataset DIOR. Les améliorations sont plus importantes sur DOTA et DIOR que sur Pascal VOC, ce qui confirme l'efficacité de la méthode proposée sur les images aériennes et les petits objets.

Pour conclure, l'analyse conduite sur les trois datasets a mis en lumière la difficulté d'appliquer les méthodes de détection *few-shot* sur des images aériennes, mais surtout que cette difficulté est liée à une plus grande proportion de petits objets. Les améliorations proposées atteignent des performances compétitives avec l'état de l'art. Il s'agit là d'un premier pas pour combler les différences de performance entre images aériennes et naturelles dans le cadre de la détection *few-shot*.

Remerciements

Les auteurs remercient l'entreprise COSE pour leur collaboration étroite et le financement de ce projet.

Références

[1] 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] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2) :303–338, 2010.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco : Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Xiang Li, Jingyu Deng, and Yi Fang. Few-shot object detection on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–14, 2021.

[5] Zixuan Xiao, Jiahao Qi, Wei Xue, and P. Zhong. Few-shot object detection with self-adaptive attention network for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14 :4854–4865, 2021.

[6] 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 CVPR*, pages 3974–3983, 2018.

[7] 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.

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos : Fully convolutional one-stage object detection. In *Procee-*

dings of the IEEE/CVF International Conference on Computer Vision, pages 9627–9636, 2019.

[9] Pierre Le Jeune and Anissa Mokraoui. A unified framework for attention-based few-shot object detection. *arXiv preprint arXiv :2201.02052*, 2022.