Détection des petits objets dans les images aériennes et satellitaires : le suréchantillonnage et la distillation suffisent

Abdelbadie BELMOUHCINE[1] Alexandre BRAVO[1] Minh-Tan PHAM[1]
[1]IRISA, Université Bretagne Sud, Vannes, France

Résumé – En télédétection, en raison de la distance entre la caméra et les objets, beaucoup de ces derniers apparaissent très petits et deviennent indistinguables après le sous-échantillonnage, ce qui rend la détection plus difficile. Cependant, éviter le sous-échantillonnage augmente considérablement le coût computationnel du réseau lors de l'inférence. Pour résoudre ce problème, au lieu de s'appuyer sur des images de haute résolution ou sur la super-résolution qui est coûteuse en termes de calcul, nous proposons une approche d'entraînement à double réseau, où l'un effectue un suréchantillonnage tôt tandis que l'autre le fait tard. En utilisant la distillation des connaissances, nous transférons les informations du détecteur à suréchantillonnage précoce vers le détecteur à suréchantillonnage tardif, améliorant ainsi la représentation des caractéristiques tout en maintenant l'efficacité computationnelle lors de l'inférence. Nos expériences sur cinq jeux de données d'objets petits montrent que le suréchantillonnage tardif et la distillation des connaissances améliorent la détection des petits objets.

Abstract – In remote sensing, due to the distance between the camera and the objects, many appear very small and become indistinguishable after downsampling, making detection more challenging. However, avoiding downsampling significantly increases the network's computational cost at inference. To address this issue, instead of relying on computationally expensive super-resolution or high-resolution images, we propose a dual-network training approach where one network upsamples early while the other upsamples late. By leveraging knowledge distillation, we transfer information from the early-upsampling detector to the late-upsampling one, improving feature representation while maintaining computational efficiency at inference. Our experiments on five small-object datasets demonstrate that late upsampling and knowledge distillation enhance small-object detection.

1 Introduction

La détection d'objets en télédétection est cruciale pour des applications comme la surveillance militaire, le suivi des navires et les opérations de sauvetage. Cependant, elle présente des défis, notamment la variabilité d'échelle des objets et leur forte densité. Les méthodes classiques, qui réduisent la résolution, échouent souvent à bien détecter ces objets, car le sous-échantillonnage les fusionne ou les mélange. Cependant, éviter ce sous-échantillonnage augmente le coût de calcul.

La super-résolution a été utilisée pour améliorer la détection [2, 7, 13], mais est-elle réellement nécessaire ? Le simple suréchantillonnage des entrées améliore déjà la détection.

Le principal défi dans la détection des petits objets réside dans le sous-échantillonnage, qui fusionne plusieurs objets en une seule représentation. Bien que le suréchantillonnage améliore la détection, il augmente aussi la complexité de calcul. Une solution serait d'augmenter la résolution à la sortie du réseau, mais le détecteur ne pourrait pas apprendre des représentations pertinentes pour la détection des petits objets. Ainsi, nous proposons une approche d'entraînement à double réseau : durant l'entraînement, nous utilisons deux réseaux YOLO – l'un appliquant le suréchantillonnage en début de réseau, l'autre en fin de réseau. Une distillation des connaissances [3] est ensuite employée pour transférer l'information du premier réseau (suréchantillonnage en entrée) vers le second (suréchantillonnage en sortie), lui permettant d'apprendre une stratégie de suréchantillonnage produisant des sorties similaires à celles du premier réseau. Lors de l'inférence, seul le réseau appliquant le suréchantillonnage en sortie est conservé, réduisant ainsi le coût de calcul tout en préservant les performances de détection.

Nous avons validé notre approche sur cinq jeux de données contenant de petits objets et couvrant divers domaines d'application. Notre stratégie prouve son efficacité pour la détection des petits objets, sans ajouter un coût de calcul excessif en inférence.

L'article est structuré comme suit : la Section 2 passe en revue les méthodes existantes de détection des petits objets, la Section 3 détaille notre approche, la Section 4 présente les expériences et les résultats, et la Section 5 conclut l'article.

2 État de l'art

La détection des petits objets a suscité l'intérêt de plusieurs chercheurs, proposant diverses stratégies combinant architectures optimisées [15, 1, 5] et/ou super-résolution [2, 7, 13].

Liu et al. [5] ont proposé une approche reposant sur trois éléments : une version modifiée de FCOS, une branche de reconstruction des objets (ORB) et un module MRFAFEM. Leur méthode utilise FCOS avec des cartes de caractéristiques haute résolution (P2) et améliore la détection grâce à une tâche de segmentation via ORB, qui vise à limiter la perte d'information pour les petits objets. Ils extraient les régions contenant ces objets et appliquent un apprentissage multitâche combinant segmentation et détection. De plus, ils renforcent dynamiquement les caractéristiques issues de différents champs réceptifs. Toutefois, cette approche présente deux inconvénients : elle requiert des annotations spécifiques pour la tâche de segmentation et double le nombre de GFLOPs en utilisant FCOS basé sur des caractéristiques du niveau P2, augmentant ainsi la

complexité de calcul malgré les bonnes performances.

Une approche largement adoptée consiste à coupler la super-résolution avec la détection d'objets. Généralement, la super-résolution est appliquée en amont afin de reconstruire des images haute résolution avant la détection. Rabbi et al.[9] ont proposé EESRGAN, un réseau génératif antagoniste (GAN) optimisé pour l'amélioration des contours, mais cette méthode reste coûteuse en calcul, car la détection intervient après la super-résolution. Liu et al.[6] ont introduit ESRTMDet, qui exploite un réseau de super-résolution pour générer des caractéristiques de haute résolution partagées avec le détecteur. Toutefois, il est difficile d'entraîner simultanément ces deux tâches de manière cohérente. En général, la plupart des approches de super-résolution cherchent à améliorer la qualité des images en augmentant la complexité du modèle. Plusieurs travaux ont adopté cette stratégie afin de faciliter la détection des petits objets. Cependant, cette augmentation de résolution accroît non seulement le coût de calcul du détecteur, mais ajoute aussi le coût de la super-résolution.

D'autres stratégies tentent d'exploiter la super-résolution plus efficacement. Zhang et al. [13] ont proposé un réseau auxiliaire de super-résolution qui génère une image haute résolution à partir des caractéristiques extraites par le backbone du détecteur. Cette méthode s'est révélée efficace, mais elle nécessite des images de haute résolution pour entraîner la branche de super-résolution. Gao et al. [2] ont conçu un réseau pyramidal de super-résolution (SSRFPN) intégré à un modèle étudiant, où la super-résolution (SSR) remplace l'interpolation bilinéaire traditionnelle. Ils ont également introduit une distillation masquée (CSD), où un réseau enseignant est entraîné sur des images haute résolution, puis ses paramètres sont figés. La sortie de SSRFPN du réseau étudiant est ensuite contrainte à s'aligner avec les sorties du FPN de l'enseignant via une fonction de perte L2, avec des couches d'adaptation ajustant les caractéristiques de l'étudiant. Ce mécanisme CSD permet aux modules SSR d'apprendre à mieux représenter les régions contenant de petits objets, améliorant ainsi la détection sans accroître excessivement la complexité.

3 Méthode proposée

Comme illustré dans la figure 1, l'entrée traverse d'abord un premier réseau de détection, où elle est suréchantillonnée avant d'être transmise à un second réseau. Ces deux réseaux sont entraînés simultanément, collaborant ainsi tout au long du processus. Contrairement à l'approche classique enseignant-étudiant, où l'enseignant est pré-entraîné avant l'étudiant, notre méthode repose sur un apprentissage conjoint. Nous avons observé que le suréchantillonnage améliore la détection des petits objets, mais augmente également la consommation mémoire et le nombre d'opérations effectuées par second (GFLOPs). Grâce à la distillation des connaissances [3], nous pouvons maintenir un coût de calcul optimisé en inférence.

La principale différence entre les deux réseaux est que l'un applique le suréchantillonnage dans ses couches finales, tandis que l'autre l'effectue dès l'entrée. Seul le premier réseau (effectuant un suréchantillonage à la fin) est conservé pour l'inférence, garantissant ainsi que les cartes de caractéristiques finales et les sorties des deux réseaux restent équivalentes.

Nous avons employé des techniques de distillation de connaissances et d'alignement de caractéristiques, comme illustré dans la figure 1. En complément des fonctions classiques de perte de la détection, nous avons introduit cinq fonctions de perte supplémentaires. Nous avons calculé la distance L1 entre les pertes de localisation des deux réseaux, notée $\mathcal{L}_{loc}$, ainsi que la distance L1 entre leurs pertes de classification, notée $\mathcal{L}_{cls}$. Soient $\mathcal{L}_{loc}^{lr}$ et $\mathcal{L}_{cls}^{lr}$ les pertes de détection du premier réseau (effectuant le suréchantillonnage en couches finales), et $\mathcal{L}_{loc}^{up}$ et $\mathcal{L}_{cls}^{up}$ celles du second réseau (appliquant le suréchantillonnage en entrée). Les pertes $\mathcal{L}_{loc}$ et $\mathcal{L}_{cls}$ sont définies par :

$$\begin{cases} \mathcal{L}_{loc} = |\mathcal{L}_{loc}^{lr} - \mathcal{L}_{loc}^{up}| \\ \mathcal{L}_{cls} = |\mathcal{L}_{cls}^{lr} - \mathcal{L}_{cls}^{up}| \end{cases} \tag{1}$$

Ensuite, nous calculons les pertes entre les sorties des deux réseaux. Soient (bb^{lr}, p^{lr}) les sorties du premier réseau et (bb^{up}, p^{up}) celles du second. Ici, bb^{lr} et bb^{up} représentent les coordonnées des boîtes englobantes, tandis que p^{lr} et p^{up} correspondent aux probabilités de classification. Nous calculons d'abord $\mathcal{L}_{CIOU}$, correspondant à la fonction de perte CIOU (Intersection sur Union Compléte) [16] entre bb^{lr} et bb^{up}, et $\mathcal{L}_{KL}$, qui est une divergence symétrique de Kullback-Leibler entre p^{lr} et p^{up}. Ces fonctions de perte sont définies par :

$$\begin{cases} \mathcal{L}_{CIOU} = 1 - CIOU(bb^{up}, bb^{lr}) \\ \mathcal{L}_{KL} = \frac{\sum_{k=1}^{C}\left(p_k^{up}\log\left(\frac{p_k^{up}}{p_k^{lr}}\right)+p_k^{lr}\log\left(\frac{p_k^{lr}}{p_k^{up}}\right)\right)}{2} \end{cases}, \tag{2}$$

où C est le nombre de classes et p_k^{lr}, p_k^{up} sont les probabilités pour la classe k. Nous avons choisi une divergence KL symétrisée plutôt qu'une KL standard ou une divergence de Jensen-Shannon, car les deux réseaux sont entraînés simultanément et aucun des deux reseaux n'est figé.

Enfin, puisque les cartes de caractéristiques après suréchantillonnage tardif dans le premier réseau ont les mêmes dimensions que celles du second, nous avons imposé un alignement des caractéristiques en calculant la similarité cosinus entre les caractéristiques au niveau des trois échelles de sortie. Cette perte est notée $\mathcal{L}_{align}$ et est définie comme suit :

$$\mathcal{L}_{align} = \sum_{s=1}^{3}\left(1 - \frac{\sum_{i=1}^{H^s}\sum_{j=1}^{W^s}\cos(F_{1ij}^s, F_{2ij}^s)}{W^s \times H^s}\right) \tag{3}$$

où $s = 1, 2, 3$ correspond aux échelles (/4, /8, /16), W^s, H^s sont respectivement la largeur et la hauteur des cartes de caractéristiques F_1^s et F_2^s (de niveau s) des deux réseaux. La similarité cosinus est calculée sur la dimension des canaux, pour chaque pixel de coordonnées (i, j).

4 Expériences

Dans cette section, nous présentons les jeux de données utilisés, les métriques d'évaluation, ainsi que les résultats expérimentaux accompagnés d'une discussion.

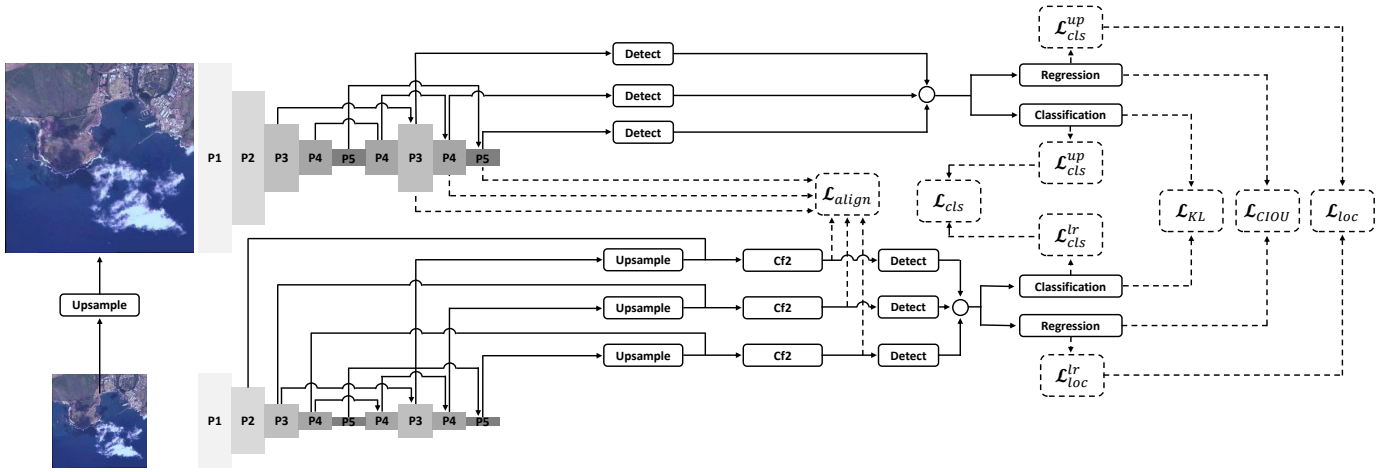

FIGURE 1 : Le processus d'entraînement proposé. Seul le réseau du bas est gardé en inférence.

4.1 Protocole expérimental

4.1.1 Jeux de données

Nous avons utilisé cinq jeux de données dominés par des objets petits et très petits, que nous décrivons ci-après. La figure 2 montre la proportion des objets en fonction de leur taille pour chaque jeu de données. Comme [11], nous considérons les objets ayant une surface inférieure à 8^2 pixels comme très minuscules, ceux ayant une surface entre 8^2 et 16^2 pixels comme minuscules, ceux ayant une surface entre 16^2 et 32^2 pixels comme petits, ceux ayant une surface entre 32^2 et 64^2 pixels comme moyens, et le reste comme grands.

VEDAI : Le jeu de données VEDAI [10] comprend 1 246 images aériennes. Seule la version 512×512 est utilisée dans cet article. Nous avons utilisé 1 089 images pour l'entraînement et 121 pour le test, en suivant un protocole de validation croisée à 10 folds. Le jeu de données contient 3 706 objets, dont 96,6% ont une surface inférieure à 32^2 pixels.

VTUAV-det : VTUAV-det [15], dérivé de VTUAV [14], est un jeu de données dédié à la détection de personnes à partir de drones. Il contient des paires d'images RGB et TIR. En suivant la répartition originale, nous avons utilisé 11 392 images pour l'entraînement et 5 378 pour le test, avec un total de 121 124 annotations de personnes, dont 92,32% ont une surface inférieurse à 32×32 pixels. Pour ce jeu de données, les images ont été redimensionnées en 512×512 tout en respectant le ratio d'aspect.

AI-TODv2 : Le jeu de données AI-TODv2 [12] comprend 28 036 images aériennes, de taille 800×800, annotées avec 752 746 instances d'objets, dont 86,44% mesurent moins de 16×16 pixels. Il est divisé en 11 214 images pour l'entraînement, 2 804 pour la validation et 14 018 pour le test.

LEVIR-Ship : Les images de LEVIR-Ship [1] sont des crops de taille 512×512, qui proviennent des satellites GaoFen-1 et GaoFen-6. Le jeu de données est réparti en 2 320 images pour l'entraînement, 788 pour la validation et 788 pour le test.

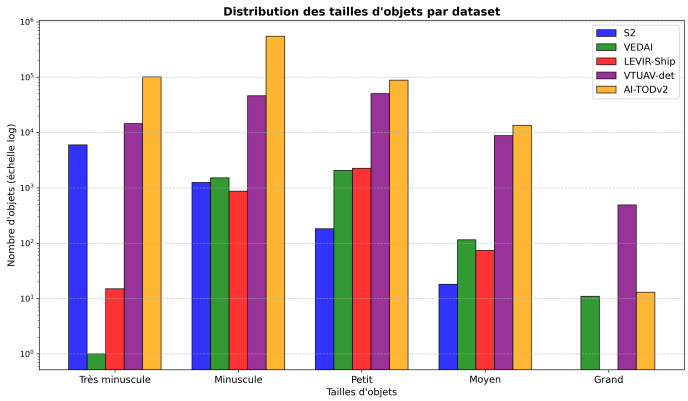

FIGURE 2 : Répartition des objets par tailles dans les cinq jeux de données.

Il contient 3 219 navires, dont 97,62% mesurent moins de 32×32 pixels.

Navires de plaisance sur Sentinel-2 Ce jeu de données [8] est composé de crops de taille 320×320 obtenus à partir de tuiles Sentinel-2 de cinq zones maritimes distinctes le long des côtes finlandaises, en mer Baltique septentrionale. Il contient 7 421 navires pour l'entraînement et la validation, dont 80,39% ont une surface inférieure à 8×8 pixels. Nous avons utilisé environ 12 193 images pour l'entraînement et environ 3 141 pour la validation, en suivant un protocole de validation croisée à cinq folds.

4.1.2 Métrique d'évaluation

Nous avons utilisé le mAP@0,5 (mean Average Precision) pour évaluer l'efficacité de notre approche. Étant donné que la précision (P) correspond au pourcentage d'objets correctement prédits parmi toutes les prédictions, et que le rappel (R) représente le pourcentage d'objets correctement prédits parmi toutes les annotations de vérité terrain, l'Average Precision (AP) est l'aire sous la courbe Précision-Rappel. AP@0,5 représente l'AP en utilisant un seuil d'Intersection sur Union (IoU) de 0,5 pour associer les prédictions aux vérités terrain. Le mAP@0,5 est la moyenne de l'AP@0,5 sur toutes les classes.

4.2 Résultats et discussion

Nous avons réalisé une étude d'ablation afin d'évaluer l'impact du suréchantillonnage sur la détection des objets. La table 1 montre que le suréchantillonnage en sortie du réseau n'offre pas des performances aussi élevées que lorsqu'il est appliqué en entrée, par contre ce dernier a un coût plus élevé. Cependant, en explorant différentes stratégies de distillation, nous observons que la combinaison de toutes les pertes définies dans la section3 conduit aux meilleurs résultats. Parmi celles-ci, les pertes appliquées aux sorties ($\mathcal{L}_{KL} + \mathcal{L}_{CIOU}$) apportent la contribution la plus significative.

Méthode	$\mathcal{L}_{loc} + \mathcal{L}_{cls}$	$\mathcal{L}_{KL} + \mathcal{L}_{CIOU}$	$\mathcal{L}_{align}$	mAP@0,5 ↑	GFLOPs ↓
Suréchantillonage au début	-	-	-	**82,9**	211,38
Suréchantillonage à la fin	✗	✗	✗	80,1	**120,00**
	✓	✗	✗	80,5	
	✗	✓	✗	83,0	
	✗	✗	✓	80,1	
	✓	✓	✗	83,0	
	✓	✗	✓	80,5	
	✗	✓	✓	83,0	
	✓	✓	✓	**83,5**	

TABLE 1 : Étude d'ablation sur VEDAI

À partir de la table 2, nous constatons que notre stratégie améliore la détection sur cinq jeux de données contenant de petits objets, démontrant ainsi son efficacité.

Jeu de données	Méthode	mAP@0,5 ↑
VEDAI	SuperYOLO [13]	75,1
	YOLOv5+CWM+SEM+SRPB [7]	82,6
	YOLOv8	79,6
	$\text{YOLOv8}_{\text{DistUp}}$	**83,5**
VTUAV-det	QFDet [15] (RGB + TIR)	75,5
	YOLOv8 (RGB)	40,1
	YOLOv8 (TIR)	75,4
	$\text{YOLOv8}_{\text{DistUp}}$ (RGB)	44,2
	$\text{YOLOv8}_{\text{DistUp}}$ (TIR)	**78,6**
AI-TODv2	ORFENet [5]	55,4
	ESG_TODNET [4]	55,1
	YOLOv8	63,0
	$\text{YOLOv8}_{\text{DistUp}}$	**65,3**
LEVIR-Ship	ORFENet [5]	**83,3**
	YOLOv8	79,5
	$\text{YOLOv8}_{\text{DistUp}}$	**83,3**
Sentinel-2	YOLOv8	61,9
	$\text{YOLOv8}_{\text{DistUp}}$	**80,3**

TABLE 2 : Comparaison des résultats sur les cinq jeux de données

5 Conclusion

Dans cet article, nous avons montré qu'un simple suréchantillonnage peut suffire pour la détection des petits objets. En combinant cette approche avec la distillation de connaissances, nous avons réussi à améliorer la détection des petits objets sans entraîner une augmentation significative du coût de calcul. Cette méthode a été validée sur cinq jeux de données contenant des petits objets et a démontré son efficacité.

À l'avenir, nous prévoyons d'étudier plus en profondeur les techniques de distillation afin d'améliorer encore la représentation des petits objets.

Références

[1] J. CHEN *et al.* : A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset. *IEEE TGRS*, 60:1–14, 2022.

[2] Y. GAO *et al.* : Feature super-resolution fusion with cross-scale distillation for small object detection in optical remote sensing images. *IEEE GRSL*, 2024.

[3] G. HINTON *et al.* : Distilling the knowledge in a neural network. *arXiv preprint arXiv :1503.02531*, 2015.

[4] D. LIU *et al.* : A tiny object detection method based on explicit semantic guidance for remote sensing images. *IEEE GRSL*, 2024.

[5] D. LIU *et al.* : Tiny object detection in remote sensing images based on object reconstruction and multiple receptive field adaptive feature enhancement. *IEEE TGRS*, 2024.

[6] F. LIU *et al.* : Esrtmdet : An end-to-end super-resolution enhanced real-time rotated object detector for degraded aerial images. *IEEE JSTARS*, 2023.

[7] J. LIU *et al.* : Small-Object Detection in Remote Sensing Images with Super Resolution Perception. *IEEE JSTARS*, 2024.

[8] J. MÄYRÄ *et al.* : Dataset for marine vessel detection from Sentinel 2 images in the Finnish coast , mars 2025.

[9] J. RABBI *et al.* : Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. *Remote Sensing*, 12(9):1432, 2020.

[10] S. RAZAKARIVONY *et al.* : Vehicle detection in aerial imagery : A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34: 187–203, 2016.

[11] J. WANG *et al.* : Tiny object detection in aerial images. *In ICPR*, pages 3791–3798. IEEE, 2021.

[12] C. XU *et al.* : Detecting tiny objects in aerial images : A normalized Wasserstein distance and a new benchmark. *ISPRS Journal of P&RS*, 190:79–93, 2022.

[13] J. ZHANG *et al.* : SuperYOLO : Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE TGRS*, 61:1–15, 2023.

[14] P. ZHANG *et al.* : Visible-thermal UAV tracking : A large-scale benchmark and new baseline. *In CVPR*, pages 8886–8895, 2022.

[15] Y. ZHANG *et al.* : Drone-based RGBT tiny person detection. *ISPRS Journal of P&RS*, 204:61–76, 2023.

[16] Z. ZHENG *et al.* : Distance-IoU loss : Faster and better learning for bounding box regression. *In AAAI conference on AI*, volume 34, pages 12993–13000, 2020.