Apprentissage de réseaux de neurones de super-résolution pour la détection d'objets de petite taille dans les images de télédétection

Luc Courtrai, Minh-Tan Pham, Jean-Christophe Burnel, Sébastien Lefèvre

Apprentissage de réseaux de neurones de super-résolution pour la détection d'objets de petite taille dans les images de télédétection

Luc Courtrai, Minh-Tan Pham, Jean-Christophe Burnel, Sébastien Lefèvre

Univ. Bretagne Sud - IRISA, UMR 6074, 56000 Vannes, France
{luc.courtrai,minh-tan.pham,jean-christophe.burnel,sebastien.lefevre}@irisa.fr

Résumé

Cet article traite du problème de détection de petits objets dans les images satellites ou aériennes [1] *au moyen d'un réseau de neurones spécialisé dans la super-résolution d'images. Nous montrons comment améliorer l'apprentissage d'un générateur, réseau de neurones basé sur des blocs résiduels (le réseau de super-résolution) en lui ajoutant un réseau adversaire (discriminateur ou critique) puis en l'intégrant dans un cycle avec un réseau de réduction d'images. En outre, en ajoutant un réseau auxiliaire de détection de petits objets dans l'architecture, nous améliorons considérablement l'apprentissage de ce même réseau de super-résolution. Les résultats de super-résolution et de détection ont été obtenus sur une base de véhicules extraites du jeu de données public ISPRS 2D Semantic Labeling Contest composé d'images aériennes acquises sur la ville de Potsdam.*

Mots Clef

Télédétection, détection d'objets, super-résolution, apprentissage profond, réseaux adversaires (GAN), Cycle-GAN, Wasserstein GAN, réseau auxiliaire.

Abstract

This paper presents deals with detection of small objects in satellite or aerial images by relying on a neural network specialized in image super-resolution. We show how to improve the learning of a generator, a neural network based on residual blocks (the super-resolution network) by adding to it an adversarial network (discriminator or critic) and then integrating it in a cycle with an image reduction network. Besides, thanks to adding an auxiliary network for small object detection in the architecture, we also considerably improve the learning of this same super-resolution network. The results of super-resolution and detection have been obtained on a set of vehicles extracted from the public dataset ISPRS 2D Semantic Labeling Contest made of aerial images acquired over the Potsdam city.

1. Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme ASTRID (projet DeepDetect, ANR-17-ASTR-0016).

Keywords

Remote sensing, object detection, super-resolution, deep learning, Generative Adversarial Network (GAN), cycle-GAN, Wasserstein GAN, auxiliary network.

1 Introduction

La détection d'objets de petite taille dans les images de télédétection est connue comme un problème difficile en vision par ordinateur du fait du faible nombre de pixels représentant ces objets dans l'image. Par exemple, sur des images satellites Pléiades (50cm/pixel), les véhicules sont contenus dans une surface d'environ 40 pixels (4x10 pixels). Pour améliorer la détection des petits objets tels que des véhicules sur des images satellites, on peut spécialiser les détecteurs classiques de l'état de l'art comme SSD (Single Shot Multibox Detector) [9], YOLOv3 (You Only Look Once) [10], Faster R-CNN [11] en réduisant les tailles des ancres pour cibler ces faibles tailles d'objets. Dans [4], les auteurs ont exploité et adapté le détecteur YOLOv3 pour la détection des véhicules dans les images Pléiades à 50cm/pixel. Pour ce faire, une base comprenant 88k véhicules a été annotée manuellement pour l'entraînement du réseau. Cette approche présente cependant le défaut d'une annotation manuelle coûteuse et, dans le cas de d'une application à des images issues d'un autre capteur, il serait nécessaire de procéder à une nouvelle phase d'annotation. De plus, en réduisant encore la résolution de ces images (1m/pixel), on arrive aux limites des possibilités de ces détecteurs.

Une approche alternative qui nous intéresse est d'effectuer la super-résolution pour augmenter la résolution spatiale des images (et donc la taille des objets) avant de réaliser la détection. Pour faire face au manque de détails présents dans les images, les dernières techniques de super-résolution (SR) par réseau de neurones comme SI-IR [14], SR-CNN [15], MDSR (Multiscale Deep Super-resolution) et EDSR (Enhanced Deep Residual Super-resolution) [8] visent à augmenter de façon significative la résolution d'une image, et ce bien mieux qu'une simple interpolation bicubique. Le lecteur intéressé par ces réseaux est invité à consulter l'article de synthèse [1]. Un réseau EDSR [8] utilise un empilement des blocs résiduels pour obte-

nir l'image super-résolue, le réseau n'apprenant que sur les images de même type que celles ciblées. Certains travaux actuels de détection de petits objets exploitent ainsi la super-résolution pour augmenter la résolution des images afin que le détecteur puisse chercher des objets de plus grande taille. Dans [3], les auteurs associent un réseau de super-résolution en amont du détecteur SSD pour la détection de véhicules sur des images satellites, et en ne modifiant que légèrement les premières couches de SSD. Ils ont montré qu'un SSD travaillant sur des images super-résolues (d'un facteur de 2 et 4) conduit à une amélioration significative par rapport à l'utilisation des images de basse résolution. De plus, l'article [13] décrit le gain apporté par la super-résolution avec l'EDSR pour différentes résolutions dans les images satellites. Les auteurs ont montré que ces techniques permettent d'améliorer considérablement les résultats de super-résolution pour des images de 30cm avec un facteur de 2 (permettant d'atteindre une résolution de 15cm), mais pas avec un facteur plus élevé (de 4, 6 ou 8 par exemple).

En effet, plus le facteur de super-résolution demandé est important (résolution encore plus faible), plus le nombre de blocs résiduels et les tailles de ces blocs (dimension des couches de convolution) doivent être importants pour espérer reconstruire l'image correctement. Un simple réseau avec un critère d'évaluation et d'optimisation du réseau de type MAE (Mean Absolute Error) ou MSE (Mean Square Error) est ainsi particulièrement difficile à entraîner du fait du nombre important de paramètres.

Dans cet article, nous débutons en section 2 par une présentation de différents modèles de super-résolution et leur évaluation sur des images de télédétection. Nous introduisons ensuite dans les sections 3 et 4 différentes améliorations de l'architecture EDSR, respectivement par intégration de réseaux adversaires et ajout d'un réseau auxiliaire. Pour illustrer nos différentes propositions, nous considérons un cadre expérimental de détection de petits objets en imagerie aérienne. Plus précisément, nous cherchons à extraire des véhicules sur des images dont la résolution a été artificiellement réduite à 1m/pixel. Les objets recherchés ont ainsi une surface inférieure à 10 pixels (2×5 pixels). Nous rapportons plusieurs résultats qualitatifs et quantitatifs de super-résolution et de détection d'objets qui illustrent expérimentalement l'intérêt de nos propositions. Finalement, la section 5 fournit quelques conclusions et pistes de recherche.

2 Réseau de super-résolution par blocs résiduels

Nous décrivons brièvement ici les différentes méthodes d'apprentissage profond pour la super-résolution avant de nous focaliser sur l'architecture EDSR [8]. On trouvera dans l'article [1] un panorama relativement complet des techniques de super-résolution en apprentissage profond.

Un réseau de neurones spécialisé en super-résolution reçoit en entrée une image de basse résolution LR et sa version haute résolution HR comme référence. Le réseau renvoie en sortie une image de résolution améliorée $SR = f(LR)$, en minimisant la distance entre $f(LR)$ et HR. Les architectures les plus "simples" sont des CNN constitués d'un empilement de couches de convolutions suivies d'une ou plusieurs couches de réarrangement de pixels. Une couche de réarrangement permet un changement de dimension d'un ensemble de couches passant de la dimension (B, Cr^2, H, W) à (B, C, Hr, Wr). Par exemple avec $r = 2$, on double la dimension en X et Y de la matrice en sortie (l'image). On peut citer ici SR-CNN [15] et SI-IR [14], ce dernier ne traitant que le canal de luminance des images. Ces premiers réseaux montrent déjà une nette amélioration de l'image résultante, par rapport à une solution classique par interpolation bicubique, tout en permettant une exécution rapide du fait de leur faible complexité (5 couches de convolution pour l'approche SI-IR).

Une amélioration de ces réseaux consiste à remplacer les couches de convolution par des blocs résiduels. Une partie de l'information d'entrée d'une couche est ajoutée telle quelle à la sortie de la couche. On peut mentionner ici l'EDSR [8] qui utilise un ensemble de blocs résiduels, et dont une version simplifiée est illustrée en figure 1. Dans cette figure, la super-résolution est réalisée avec un facteur de 4, en utilisant simplement 4 blocs résiduels. La couche de mise à l'échelle (d'un facteur 2) des couches inférieures est assurée par l'opération de réarrangemment, *pixel shuffle* en Pytorch, indiquée en orange dans la figure.

Comme indiqué précédemment, nous évaluons le comportement et les performances des approches discutées ici à l'aide du jeu de données ISPRS 2D Semantic Labeling Contest [12] [2]. Plus spécifiquement, nous nous intéressons aux images aériennes acquises sur la ville de Potsdam qui sont fournies avec une résolution spatiale de 5cm/pixel. Ce jeu de données a été initialement conçu pour évaluer les méthodes de segmentation sémantique, en considérant 6 classes : surfaces imperméables, bâtiments, végétation basse, arbres, véhicules, et autre/fond). Il peut être cependant utilisé pour des tâches de détection de véhicules [2], en ne retenant que les composantes connexes des pixels étiquetés comme appartenant à la classe véhicules. La base ainsi construite contient près de 10 000 véhicules. Nous considérons ces images bien résolues et de bonne qua-

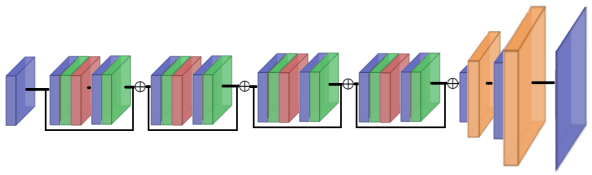


FIGURE 1 – Architecture d'EDSR avec 4 blocs résiduels, avec des couches de convolution (bleu), normalisation (vert), d'activation ReLU (rouge), et de réarrangement (orange).

2. http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

lité (dont nous n'exploitons que les bandes RVB) pour conduire des tests avec différentes résolutions artificiellement fixées à 12.5cm/pixel, 50cm/pixel et 100cm/pixel, et pour entraîner le réseau de super-résolution. La résolution 12.5cm/pixel nous sert de référence (version HR). Nous utilisons une architecture EDSR avec 16 blocs résiduels de taille 64x64. Pour optimiser les résultats, EDSR effectue en plus une standardisation-normalisation des pixels de l'image sur les trois bandes lors de l'inférence, en fonction des valeurs moyennes des pixels des images utilisées pour l'entraînement. Sur ce réseau, le calcul de l'erreur s'effectue par la fonction de coût L1 et l'optimiseur utilisé est ADAM [6].

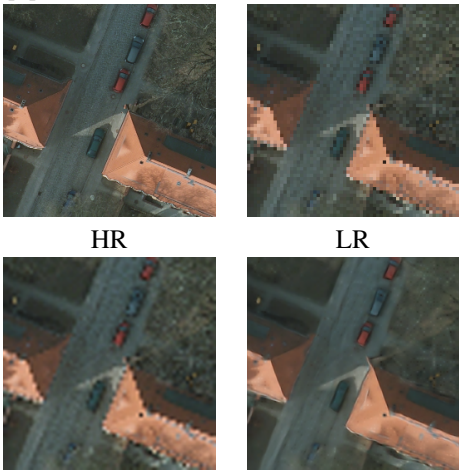


FIGURE 2 – Illustration des résultats fournis par l'EDSR : image haute résolution (HR) à 12,5cm/pixel et sa version artificiellement réduite (LR) à 50cm/pixel (agrandie pour une meilleure lisibilité), résultats fournis à 12,5cm/pixel par l'interpolation bicubique et par la super-résolution EDSR (facteur 4).

La figure 2 montre l'effet d'une super-résolution d'un facteur 4, qui accroît la résolution de 50cm/pixel à 12,5cm/pixel. On peut constater visuellement le gain significatif par rapport à une mise à l'échelle par interpolation bicubique. Pour évaluer quantitativement la performance de la super-résolution dans notre contexte de détection de petits objets, nous utilisons le détecteur YOLOv3 [10]. YOLOv3 est ici entraîné sur les images de haute résolution à 12,5cm/pixel. Yolo retourne par chaque détection une valeur de confiance, que nous comparons à un seuil fixé à $0,25$. La mesure d'IoU (Intersection Over Union) est utilisée pour calculer un indice de surface entre le rectangle englobant l'objet détecté et la vérité terrain. L'objet est considéré comme détecté si son IoU avec la vérité terrain est supérieure au seuil. Les objets à détecter étant petits, nous choisissons un seuil pour cette IoU faible soit $0,25$. Nous mesurons les vrais positifs (TP), les faux positifs (FP) et le F1score, calculé à partir de TP, FP et de l'effectif réel. Nous calculons enfin la mAP (mean average precision) comme indice de précision pour différentes valeurs du rappel. La table 1 montre que la super-résolution par l'EDSR fournit des résultats proches de ceux obtenus sur les images de haute résolution (HR), bien au-delà de ceux obtenus avec une interpolation bicubique.

Méthode	TP	FP	F1-score	mAP
HR	707	32	0,90	92,94
Bicubique	268	13	0,48	71,82
EDSR-4	648	27	0,86	90,73

TABLE 1 – Évaluation quantitative de EDSR-4 (seuil de confiance de 0,25 et un IoU de 0,25) et comparaison avec l'image originale (HR) et l'interpolation bicubique.

En descendant la résolution de l'image à 1m/pixel (résolution visée dans l'étude) commme l'illustre la figure 2, la super-résolution par blocs résiduels d'un facteur 8 ne permet pas de reconstruire une image de qualité. La table 2 confirme ainsi les mauvaises performances de détection avec le détecteur YOLOv3 sur les images interpolées ou super-résolues avec un facteur de 8.

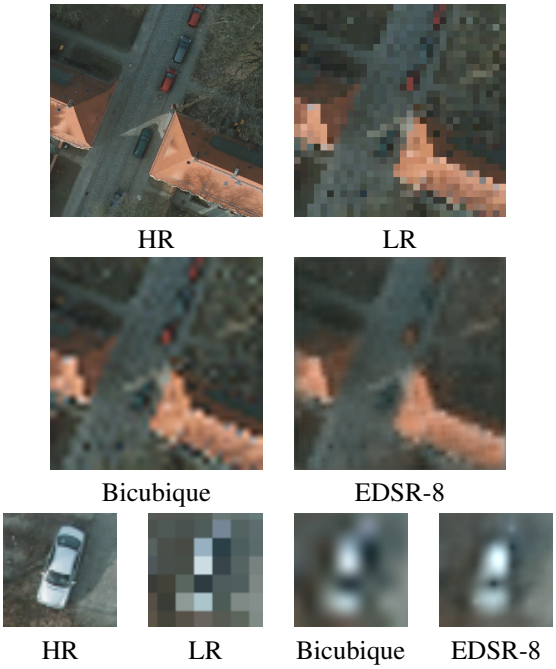


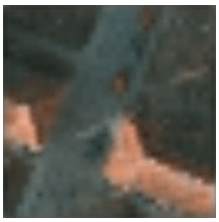
FIGURE 3 – Super-résolution d'un facteur 8 : image haute résolution (HR) à 12,5cm/pixel et sa version artificiellement réduite (LR) à 100cm/pixel (agrandie pour une meilleure lisibilité), résultats fournis à 12,5cm/pixel par l'interpolation bicubique et par la super-résolution EDSR (facteur 8). En bas : zoom sur un véhicule (50×50 pixels) : images HR et LR, résultats d'interpolation bicubique et EDSR-8.

Pour améliorer la qualité des images super-résolues, il faut augmenter le nombre de paramètres du réseau de super-résolution. Pour cela, nous passons le nombre de blocs résiduels de 16 à 32 et la taille de ces blocs de 64×64 à 96×96. Le nombre de paramètres du réseau passe alors de 1 665 307 à 6 399 387. La figure 4 montre la différence des images issues des deux réseaux.

La table 3 montre l'amélioration importante des résultats

Méthode	TP	FP	F1-score	mAP
HR	707	32	0,90	92,94
Bicubique	34	9	0,02	22,04
EDSR-8	14	1	0,03	18,10

TABLE 2 – Évaluation quantitative de EDSR-8 (seuil de confiance de 0,25 et un IoU de 0,25) et comparaison avec l'image originale (HR) et l'interpolation bicubique.



EDSR-8 (16 64×64)

EDSR-8 (32 96×96)

FIGURE 4 – Comparaison entre une image super-résolue par EDSR avec 16 blocs résiduels de taille 64×64 (à gauche) et avec 32 blocs de taille 96×96 (à droite).

du nouveau réseau en détection de véhicules grâce à cette augmentation du nombre de paramètres.

Version	TP	FP	F1score	mAP
HR	707	32	0,89	92,94
EDSR-8 (16 64×64)	14	1	0,03	18,10
EDSR-8 (32 96×96)	116	8	0,25	39,43

TABLE 3 – Effet du nombre et de la taille des blocs résiduels de EDSR.

Le problème reste l'entraînement de ce type de réseau qui comporte un nombre important de paramètres. Un critère d'évaluation du réseau de type MAE (Mean Absolute Error, ou moyenne des valeurs absolues des écarts) ou MSE (Mean Square Error) avec une optimisation de type ADAM reste limité. Plusieurs entraînements sur des jeux de données du réseau issus d'un même ensemble d'images de super-résolution de 32 blocs résiduels de taille 96×96 ont donné des résultats forts différents, en fonction notamment de l'ordre des images lors de l'entraînement (voir table 4) et de l'utilisation du générateur aléatoire à différents niveaux.

Cycle	TP	FP	F1-score
1	116	8	0,25
2	53	12	0,12
3	85	15	0,18
4	105	8	0,22
5	46	10	0,10
moyenne	81	10,6	0,17
écart-type	± 27	$\pm 2{,}65$	$\pm 0{,}06$

TABLE 4 – Variabilité des résultats selon les cycles d'entraînement.

La section suivante décrit des architectures de réseau qui permettent un apprentissage plus orienté pour les résultats visés et plus stable.

3 Amélioration du réseau

Nous explorons ici différentes stratégies pour améliorer les résultats préliminaires obtenus dans la section précédente. Nous nous intéressons ainsi d'abord aux réseaux adversaires avant de nous orienter vers les réseaux cycliques.

3.1 Réseaux adversaires

Les travaux récents utilisent des réseaux adversaires (GAN) dans l'apprentissage du réseau de super-résolution. L'association du réseau de super-résolution (le générateur) avec un discriminateur permet de paramétrer au mieux le générateur. On peut citer ici SR-GAN [7] qui reprend un réseau avec des blocs résiduels de type EDSR pour le générateur. Dans un GAN, le réseau discriminateur (figure 5) doit distinguer les images réelles de celles obtenues par super-résolution. Le générateur doit tromper le discriminateur sachant que celui-ci s'améliore au fur et à mesure des itérations. Les résultats du discriminateur interviennent dans l'entraînement du générateur via le calcul de la fonction de coût qui dépend de l'image obtenue par le générateur G. L'image finale sera proche de la cible via le critère de la fonction de coût MSE ou L1 et plus réaliste via le discriminateur.

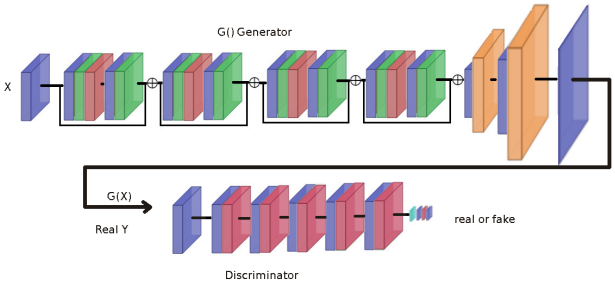


FIGURE 5 – Architecture de SR-WGAN avec en haut le générateur (réseau de super-résolution) et en bas le discriminateur. Les couches sont de type : convolution (bleu), réduction 1×1 (bleu clair), activation ReLU (rouge), normalisation (vert), réarrangement (orange).

Nous évaluerons expérimentalement cette solution en choisissant comme générateur le réseau de super-résolution avec 32 blocs résiduels décrit dans la section précédente, et en y ajoutant un réseau discriminateur (ou critique). Nous utilisons la version Wasserstein GAN (WGAN) [5] avec l'ajout d'une pénalité des gradients (deuxième partie d'équation) :

$$\min_{\theta} \max_{\phi} \sum_{x \sim \widehat{\mathbb{P}}_r} [\mathcal{D}_\phi(x)] - \sum_{z \sim \widehat{\mathbb{P}}_z} [\mathcal{D}_\phi(Gsr_\theta(z))] + \sum_{\hat{x} \sim \widehat{\mathbb{P}}_{\hat{x}}} [(\|\nabla \mathcal{D}_\phi(\hat{x})\|_2 - 1)^2] \,. \quad (1)$$

où x est une entrée HR, z l'entrée LR associée à x, Gsr_θ le réseau de super-résolution, $\mathcal{D}_\phi$ le discriminateur, $\nabla\mathcal{D}_\phi(\hat{x})$ le gradient du discriminateur et $\hat{x}$ un élément aléatoire construit avec $Gsr_\theta(z)$ et x.

3.2 Réseaux cycliques

Une seconde amélioration consiste à intégrer le réseau de super-résolution dans un cycle. Un second réseau a alors pour but de prendre une image de haute résolution et d'en générer une version de basse résolution (figure 6). L'image ainsi obtenue est comparée à l'image de basse résolution initiale. Symétriquement, celle de haute résolution sera comparée au résultat obtenu après passages successifs par les réseaux de basse et de haute résolution. Le système utilise l'ensemble des comparaisons ci-dessous pour le calcul de la fonction de coût (ou *loss*) où x_{HR} est l'image de haute résolution d'entrée, x_{LR} la version de basse résolution, Ghr le générateur de super-résolution et Glr celui générant l'image de basse résolution :

$$\begin{aligned}\mathcal{L}^{L1}(x_{HR}, Ghr(x_{LR})) + \mathcal{L}^{L1}(x_{LR}, Glr(x_{HR}))\\ +\mathcal{L}^{MSE}(x_{HR}, Ghr(Glr(x_{HR})))\\ +\mathcal{L}^{MSE}(x_{LR}, Glr(Ghr(x_{LR})))\end{aligned} \quad (2)$$

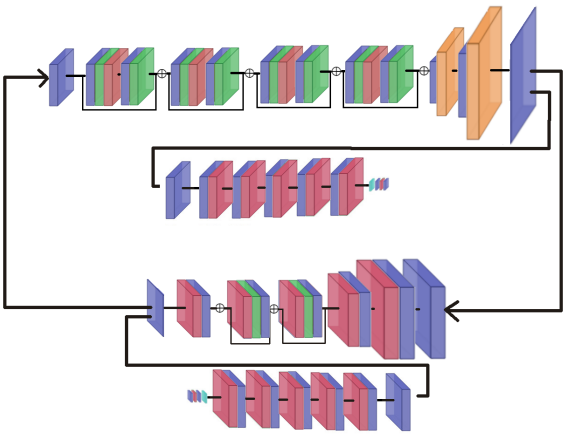


FIGURE 6 – Architecture d'un réseau cyclique : générateur de super-résolution et son discriminateur (en haut), et générateur de basse résolution et son discriminateur (en bas).

On peut remarquer sur la figure 7 des détails plus importants sur les images générées par les réseaux WGAN simple et WGAN dans un cycle par rapport à la version EDSR. De même, les performances (table 5) montrent une augmentation du taux de détection par les réseaux WGAN par rapport à EDSR et plus particulièrement par le réseau en cycle. Dans le cycle, 4 réseaux interviennent dans l'apprentissage des poids du réseau de super-résolution (le nombre de blocs résiduels du réseau de super-résolution est toujours égal à 32).

4 Ajout d'un réseau auxiliaire

Nous proposons maintenant d'ajouter un réseau auxiliaire à l'architecture précédente. Celui-ci détectera lors de l'entraînement les objets dans l'image super-résolue et le calcul de la fonction de coût $\mathcal{L}$ prendra en compte cette détection. Pour cela, nous utilisons YOLOv3 pour la détection des objets lors de l'apprentissage du réseau de super-résolution. Il a été préalablement appris sur des images de même type et à la même résolution que les images qui seront produites par la super-résolution. Lors de l'apprentissage, les images produites par le générateur sont donc passées en entrée de YOLOv3 qui calcule sa fonction de coût $\mathcal{L}$ de prédiction d'objets dans les boites englobantes. L'ajout du réseau auxiliaire est implémenté dans les deux versions WGAN et Cycle WGAN décrites dans la section précédente. La figure 8 illustre l'ajout du réseau YOLOv3 dans l'architecture WGAN du modèle de super-résolution. Les images issues du réseau de super-résolution sont ainsi passées à YOLOv3 qui calcule les prédictions. Lors de la rétro-propagation du gradient, les poids du réseau du super-résolution sont également mis à jour via la fonction de coût sur la détection de YOLOv3 ; les poids de YOLOv3 restent quant à eux figés.

La fonction de coût globale est composée de trois fonctions :

— $\mathcal{L}_{\text{Gsr}}$ du générateur ;

$$|Gsr(x) - x|$$

— $\mathcal{L}_{\text{Critic}}$ du critique ou discriminateur (Wasserstein

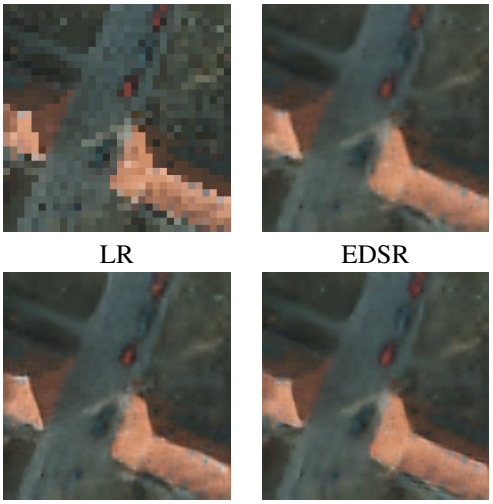


FIGURE 7 – Illustration des images super-résolues par différentes méthodes.

Méthode	IoU = 0,05	IoU = 0,25	IoU = 0,5
HR	95,04	92,93	81,04
Bicubique	26,68	22,08	9,51
EDSR	43,80	39,48	27,63
WGAN	54,96	50,45	36,39
Cycle WGAN	59,38	55,68	40,85

TABLE 5 – Résultats de mAP obtenus avec différents niveaux de IoU.

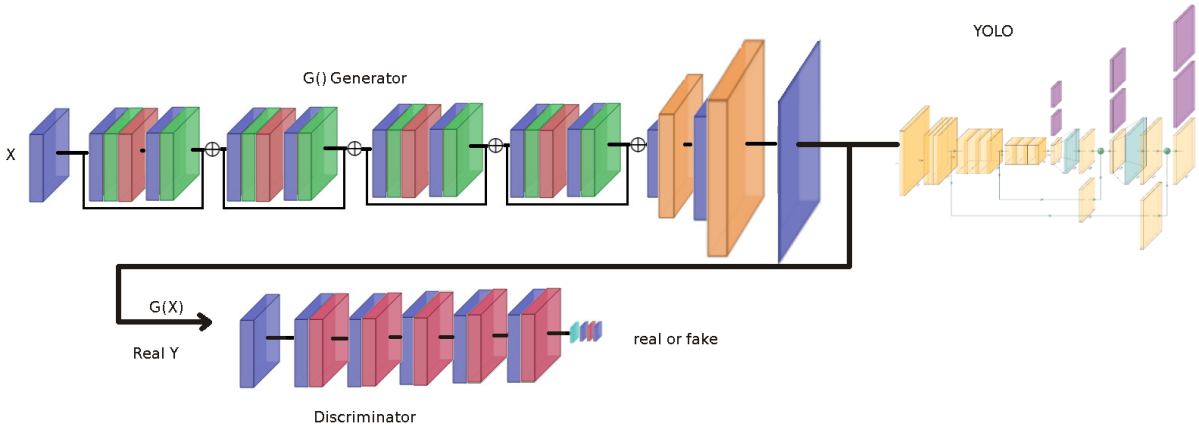


FIGURE 8 – Architecture SR WGAN avec l'ajout du réseau YOLOv3 comme réseau auxiliaire.

GAN) :

$$\min_{\theta}\max_{\phi}\sum_{x\sim\widehat{\mathbb{P}}_r}[\mathcal{D}_\phi(x)]-\sum_{z\sim\widehat{\mathbb{P}}_z}[\mathcal{D}_\phi(Gsr_\theta(z))]+\sum_{\hat{x}\sim\widehat{\mathbb{P}}_{\hat{x}}}[(\|\nabla\mathcal{D}_\phi(\hat{x})\|_2-1)^2]$$

— $\mathcal{L}_{\text{Yolo}}$ du détecteur YOLO, qui cherche à minimiser l'écart entre les boites englobantes (définies par une position x,y, une hauteur h, et une largeur v) des objets détectés par YOLO et les annotations d'entrée :

$$\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{obj}\left(\left(x_i-\hat{x}_i\right)^2+\left(y_i-\hat{y}_i\right)^2+\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right)$$

La fonction de coût globale s'écrit alors simplement comme une somme pondérée :

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{Gsr}}+\alpha\,\mathcal{L}_{\text{Critic}}+\beta\,\mathcal{L}_{\text{Yolo}} \quad (3)$$

et nous prenons $\alpha=10$ et $\beta=10^{-1}$ pour garder un équilibre des directions données par les trois fonctions de coût lors de l'apprentissage.
Pour mieux qualifier la performance du réseau, nous avons ajouté le calcul du *F1-score* sur le lot d'images courant pour suivre le taux de détection au fur et à mesure des itérations. À titre de comparaison, la fonction a aussi été ajoutée aux réseaux décrits dans les sections précédentes. On constate sur la figure 9 l'amélioration rapide des détections grâce à l'ajout de la fonction du coût du détecteur dans le réseau. Vient ensuite une phase de légère décroissance qui s'explique par le rééquilibrage de la fonction de coût identité $Ghr(LR)=HR$ du générateur. Si l'objectif du réseau est la détection des objets, on pourrait arrêter l'entraînement avant cette décroissance. Au début de l'entraînement (moins de 100 étapes), les images générées sont de mauvaise qualité (en noir et blanc et peu réalistes) mais le détecteur y trouve des véhicules (le seuil de détection est volontairement bas pour amorcer l'influence du détecteur dans l'apprentissage).

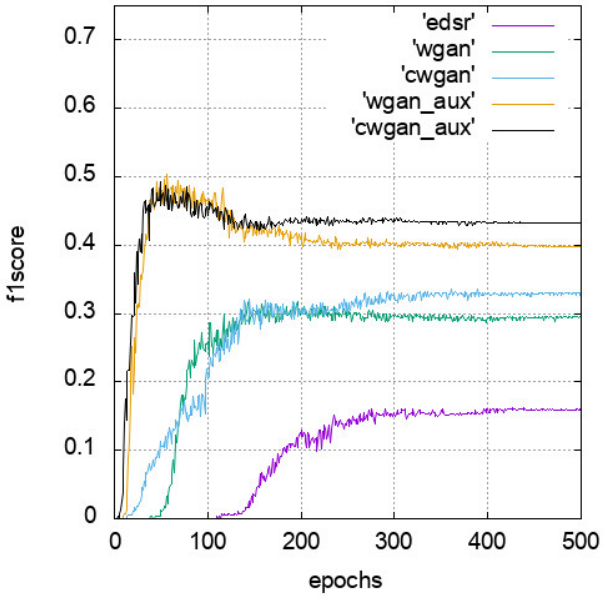


FIGURE 9 – Mesures de F1-score obtenues avec les différentes réseaux lors de l'entraînement (celles notées "aux" possèdent un détecteur comme réseau auxiliaire).

La figure 10 présente des résultats obtenus par les réseaux WGAN et Cycle WGAN, avec ou sans détecteur comme réseau auxiliaire. Les réseaux reconstruisent des détails de véhicule perdus dans la version de basse résolution. L'ajout d'un détecteur auxiliaire focalise l'apprentissage de la super-résolution sur les objets à détecter.
La figure 11 illustre la capacité de reconstruction des différentes approches discutées dans cet article, en considérant une image d'entrée de basse résolution où la surface du véhicule est de 8 pixels (résolution de 1m/pixel). On

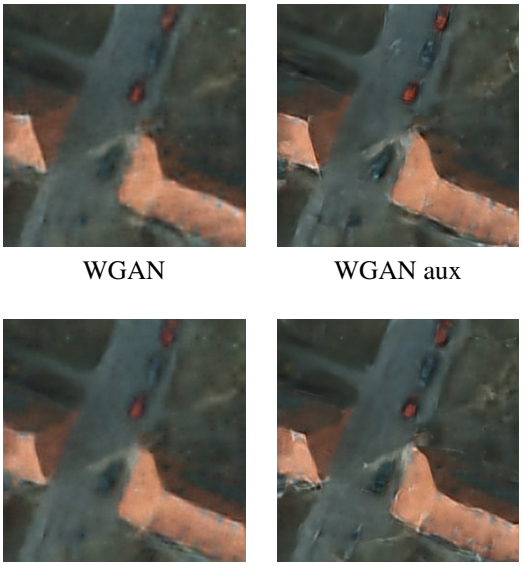


FIGURE 10 – Images super-résolues par WGAN, WGAN auxiliaire, Cycle WGAN, Cycle WGAN auxiliaire.

peut constater que les résultats s'approchent visuellement de l'image haute résolution de référence.

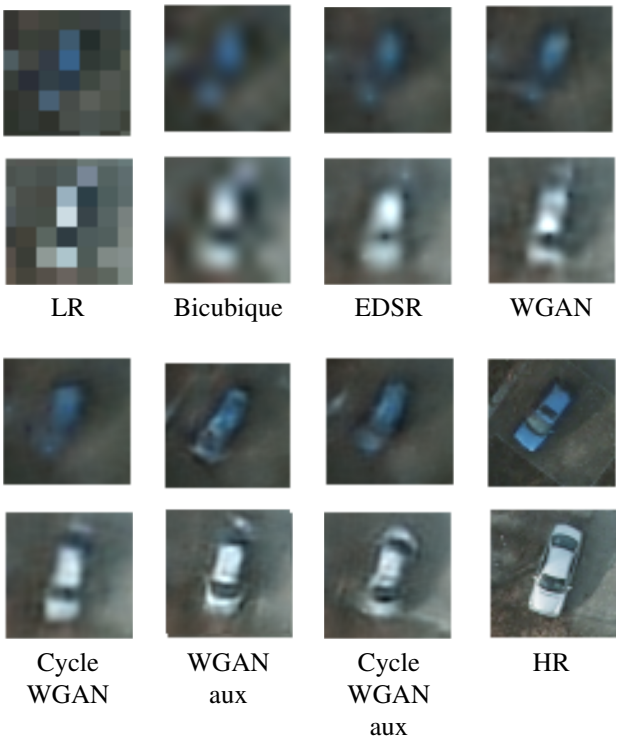


FIGURE 11 – Comparaison visuelle de la reconstruction de véhicules par le réseau de super-résolution EDSR et ses différentes évolutions.

La table 6 montre qu'en ajoutant au WGAN ou Cycle WGAN un réseau auxiliaire de détection, on peut atteindre des mAP supérieurs à 60% en détection sur les images super-résolues. Il faut noter également que la super-résolution apprise avec un détecteur n'engendre pas davantage de faux positifs que les versions sans détecteur.

Méthode	IoU = 0,05	IoU = 0,25	IoU = 0,5
HR	95,04	92,93	81,04
Bicubique	26,68	22,08	9,51
EDSR	43,80	39,48	27,63
WGAN	54,96	50,45	36,39
Cycle WGAN	59,38	55,68	40,85
WGAN aux	63,16	58,66	44,41
Cycle WGAN aux	64,16	60,18	45,17

TABLE 6 – Résultats de mAP obtenus avec différents niveaux de IoU pour l'ensemble des méthodes explorées dans cet article.

On peut finalement voir en figure 12 les résultats obtenus en super-résolution et détection d'objets avec les différentes versions présentées dans cette étude pour un autre extrait du jeu de données. On constate sur cet exemple que tous les véhicules sont détectées sur la version WGAN aux de la super-résolution. La version Cycle WGAN aux n'a détecté que trois véhicules sur les quatre mais avec une confiance plus importante que la version WGAN aux, probablement du fait d'une meilleure définition spatiale des véhicules. Rappelons que la table 6 montrait que la version avec un cycle s'est avérée légèrement meilleure en détection.

5 Conclusion

Dans cet article, nous avons montré que la super-résolution permettait d'améliorer la détection d'objets de petites tailles dans des images aériennes ou satellites, lorsqu'elle est associée à un détecteur tel que YOLOv3. En augmentant la taille et le nombre des blocs résiduels du réseau de super-résolution, et en le couplant à un détecteur comme réseau auxiliaire dans la phase d'entraînement, on améliore grandement la qualité de l'image super-résolue. On accroît ainsi la détection en spécialisant la super-résolution aux objets cibles. L'architecture proposée avec cinq réseaux et les différentes fonctions de coût permettent un apprentissage dirigé, ici sur la détection de petits objets.

Nous envisageons à l'avenir de conduire des travaux visant à affiner dynamiquement les coefficients des différents fonctions de coût au cours de l'entraînement pour converger vers une solution optimale. De plus, l'utilisation d'une version orientée du détecteur devrait permettre de mieux viser les pixels des objets à détecter.

Remerciements

Les auteurs souhaitent remercier l'ISPRS et BSF Swissphoto pour la mise à disposition du jeu d'images aériennes sur la ville de Potsdam.

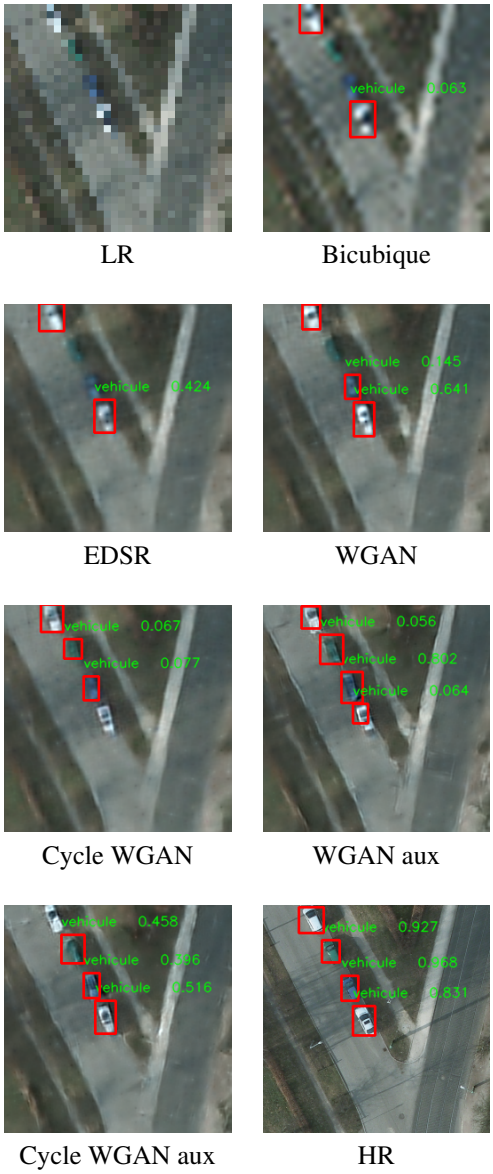


FIGURE 12 – Exemples de résultats de détection.

Références

[1] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution : A survey. *arXiv preprint arXiv :1904.07523*, 2019.

[2] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Segment-before-Detect : Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing*, 9(4) :1–18, 2017.

[3] Syeda Nyma Ferdous, Moktari Mostofa, and Nasser M Nasrabadi. Super resolution-assisted deep aerial vehicle detection. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100617. International Society for Optics and Photonics, 2019.

[4] Alice Froidevaux, Andréa Julier, Agustin Lifschitz, Minh-Tan Pham, Romain Dambreville, Sébastien Lefèvre, and Pierre Lassalle. Vehicle detection and counting from vhr satellite images : efforts and open issues. *arXiv preprint arXiv :1910.10017*, 2019.

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777, 2017.

[6] Diederik P Kingma and Jimmy Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.

[7] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[8] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR-WS*, pages 136–144, 2017.

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd : Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[10] Joseph Redmon and Ali Farhadi. Yolov3 : An incremental improvement. *arXiv preprint arXiv :1804.02767*, 2018.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn : Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[12] Franz Rottensteiner and et al. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals I-3 (2012), Nr. 1*, 1(1) :293–298, 2012.

[13] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *CVPR-WS*, pages 0–0, 2019.

[14] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016.

[15] Liguo Zhou, Zhongyuan Wang, Yimin Luo, and Zixiang Xiong. Separability and compactness network for image recognition and superresolution. *IEEE transactions on neural networks and learning systems*, 30(11) :3275–3286, 2019.