SMALL OBJECT DETECTION FROM REMOTE SENSING IMAGES WITH THE HELP OF OBJECT-FOCUSED SUPER-RESOLUTION USING WASSERSTEIN GANS

Luc Courtrai, Minh-Tan Pham, Chloé Friguet, Sébastien Lefèvre

Univ. Bretagne Sud - IRISA, 56000 Vannes, France
{luc.courtrai,minh-tan.pham,chloe.friguet,sebastien.lefevre}@irisa.fr

ABSTRACT

In this paper, we investigate and improve the use of a super-resolution approach to benefit the detection of small objects from aerial and satellite remote sensing images. The main idea is to focus the super-resolution on target objects within the training phase. Such a technique requires a reduced number of network layers depending on the desired scale factor and the reduced size of the target objects. The learning of our super-resolution network is performed using deep residual blocks integrated in a Wasserstein Generative adversarial network. Then, detection task is performed by exploiting two state-of-the-art detectors including Faster-RCNN and YOLOv3. Experiments were conducted on small vehicle detection from both aerial and satellite images from the VEDAI and xView data sets. Results showed that object-focused super-resolution improves the detection performance and facilitates the transfer learning from one data set to another.

***Index Terms*—** Small object detection, deep learning, super-resolution, Wasserstein GANs, remote sensing imagery

1. INTRODUCTION

The detection of small objects (less than 10×10 pixels), such as vehicles or animals from aerial or satellite remote sensing images, requires specialization and adaptation of existing state-of-the-art detectors in computer vision such as Faster-RCNN [1] (Faster Region-based Convolutional Neural Network), SSD (Single Shot multibox Detector) or YOLOv3 (You Only Look Once version 3) [2]. The latest super-resolution (SR) techniques based on deep neural networks greatly improve the spatial resolution of an image compared to a simple bi-cubic interpolation (see a recent review in [3]). Therefore, coupling such a super-resolution framework with an object detector should allow us to first increase the size of objects with good quality within an image, and then improve the detection performance.

In the literature, some recent studies have been proposed to pursue this direction. In [4], the authors evaluated the effects of SR techniques using the VDSR (Very Deep Super-Resolution) and the RFSR (Random Forest Super-Resolution) frameworks on object detection across multiple resolutions. They proposed to quantify the gain in detection performance with regard to the spatial resolution of satellite images and proved that these techniques provide great improvement in case of super-resolving 30-cm images by a factor of 2 (meaning to bring to 15-cm resolution), but less beneficial at higher factors (from 4, 6 or 8 for example). No modification regarding the architectures of VDSR and RFSR, as well as SSD and YOLT (You Only Look Twice) detector has been proposed within this study. Next, the authors in [5] proposed to associate an SR network based on Generative Adversarial Network (SR-GAN) [6] to the SSD detector to improve the performance of vehicle detection from aerial images. As their objective, they showed that an SSD trained and predicted on super-resolved images (by a factor of 2 and 4) could achieve great improvement compared to the case using low-resolution (LR) images. Their experiments were conducted on the VEDAI dataset [7] by only considering 1 class of vehicle instead of the 9 classes from the original data, from which the global influence on multiple classes has been ignored.

In this paper, the aim is to improve the exploitation of super-resolution to assist the detection of small objects of multiple classes from both aerial and satellite remote sensing images. We propose to train the super-resolution focused on the objects to be detected (target objects) using an existing CNN-based SR method. To do this, we modify an EDSR (Enhanced Deep Residual SR) [8] based on residual blocks learned in a Wasserstein GAN. We prove that training such a network focused on image patches including target objects rather than on the entire images could provide better detection performance on the VEDAI dataset [7]. Moreover, we show that our technique could benefit the transfer learning from one high-resolution (HR) dataset to another lower-resolution (LR) one. We provide promising results using the SR learned on the aerial ISPRS Potsdam dataset [9] to assist the detection of vehicles from the satellite xView dataset [10]. In the remainder of the paper, Section 2 describes the object-focused SR technique and provides the improvement on vehicle detection from the VEDAI data. In Section 3, the SR based on modified EDSR learned in a Wasserstein GAN framework is studied in order to improve the performance of SR to assist detection task. Section 4 shows the advantage of the proposed tech-

This work is funded by ANR/DGA through the DEEPDETECT project (ANR-17-ASTR-0016).

nique within a multi-resolution transfer learning context and Section 5 finally draws a conclusion of our work.

2. OBJECT-FOCUSED SUPER-RESOLUTION

The main motivation in our approach is to create a specific super-resolution network for the detection of small objects (here, vehicles in aerial or satellite images). Instead of learning the super-resolution on the entire images from a data set, we will focus the learning only on the target objects. To do this, we split the images into a set of patches each containing at least one object. For illustration, we perform the SR-IR [11] on the VEDAI database [7]. The resolution of VEDAI images is 12.5cm/pixel and the database contains approximately 3500 vehicles divided into 9 classes. To focus the learning of SR on target vehicles, we create image patches of size 64×64 pixels. Due to the small size of these patches, only a 5-layer network SR-IR is sufficient here for learning the SR by a factor of 4. Figure 1 shows the result of the object-focused SR (64×64 patch) compared to the same SR learned on the full 1024×1024 VEDAI images.

Fig. 1. Qualitative performance of object-focused SR. Left to right: LR, SR on the full image, object-focused SR, HR.

For quantitative evaluation, we now compare the effect of super-resolution on a detection task by running SR-IR trained on the full VEDAI images or on image patches containing the vehicles. For the detection task, we exploit YOLOv3 [2] which is one of the best one-stage detectors in the literature. Figure 2 shows the performance gain of the detection on super-resolved images (*YOLOv3Sr4Full* and *YOLOv3Sr4* green points) over the detection on low-resolution images without SR (*YOLOv3* green point). The YOLOv3 red point represents the reference which is the detection performance of YOLOv3 on VEDAI images with high spatial resolution of 12.5cm/pixel. In our experiment, the SR was achieved with a factor 4. The low-resolution images had a resolution of 50cm/pixel, which are the results of a bi-cubic downsampling from the original images. For YOLOv3 detection, the class confidence threshold is fixed to 0.25 and the IoU (Intersection over Union) threshold is set to 0.1.

In this figure, the two green points *YOLOv3Sr4Full* and *YOLOv3Sr4* show the result of detection on super-resolved images trained on whole images and on sub-images focused on vehicles, respectively. We observe that with the object-focused SR, YOLOv3 achieved a better performance in both precision and recall than when the SR is performed on whole images. Although these two green points do not approach the reference red point because of the simple technique for super-resolution (SR-IR) in use, this experiment shows us a good practice when using SR to assist object detection task: performing super-resolution focused on objects to be detected.

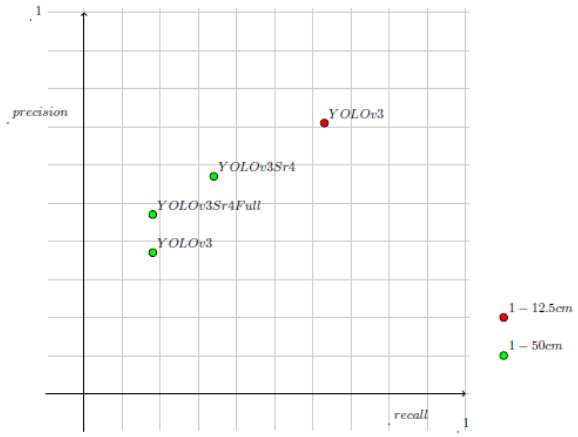

Fig. 2. Detection performance on the VEDAI data using the YOLOv3 detector associated with SR by a factor of 4. VEDAI images with different resolutions: red for 12.5cm/pixel and green for 50cm/pixel. Here, YOLOv3 is combined with SR trained on full images (*YOLOv3Sr4Full*) and SR focused on target objects (*YOLOv3Sr4*).

3. SUPER-RESOLUTION BASED ON RESIDUAL BLOCKS AND WASSERSTEIN GAN

In order to improve the performance of super-resolution, a usual approach is to increase the number of network layers. As proposed in the EDSR (Enhanced Deep Residual Super-resolution) [8], we exploit residual block layers but we distribute here the uploading layers (pixel shuffle) by a factor of 2 on the network. The advantage of such a technique is to reduce the number of layers in the network according to the SR scale factor. For example, for sets of 2 residual blocks, we use 2 blocks for a factor of 2, 4 blocks for a factor of 4 and 6 blocks for a factor of 8. Figure 3 shows the network architecture: in yellow the uploading layers (pixel shuffle) interposed between sets of residual blocks, here with a SR by a factor of 4 and sets of 2 residual blocks. Since the object sizes are small, there is no need to use a large subset of residual blocks (4 residual blocks for an object-focused SR here against 16 in the standard SR over full images). Table 1 shows the network size (number of parameters) according to the SR factor.

factor	2	4	8
parameters	480k	780k	1070k

Table 1. Network size according to the SR factor.

To learn the SR network, an adversarial network based on the Wasserstein distance (WGAN) as proposed in [12] is

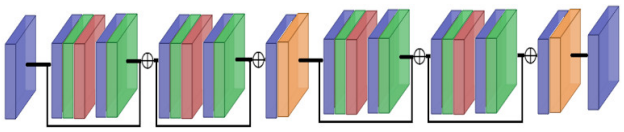

Fig. 3. Modification of the EDSR architecture (blue: convolution layer, green: batch normalisation layer, red: ReLU layer and orange: pixel shuffle layer)

exploited. For more details about the functionality of WGAN compared to a standard GAN, readers are referred to [12]. This Wasserstein GAN version adds of a gradient penalty

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

where x is an HR version, z is the LR version associated with x, Gsr_θ the super-resolution part, $\mathcal{D}_\phi$ the discriminator part, $\nabla\mathcal{D}_\phi(\hat{x})$ the gradient of the discriminator and $\hat{x}$ a random element constructed with $Gsr_\theta(z)$ and x.

Fig. 4. Super-resolution results obtained by the WGAN compared to SR-IR methods. From left to right: LR image, SR-IR result, SR-WGAN result and HR image.

Figure 4 provides some examples of SR results using the SR-IR and SR-WGAN methods on vehicles from the VEDAI database. We can observe that the quality of the objects is clearly improved. The SR-WGAN is learned on images from the ISPRS Potsdam dataset [9] with images in TIFF format and high spatial resolution of 5cm/pixel. The quality of the images obtained by SR-WGAN focused on the target objects is sufficient for the latter detection task. No specific training is required since the quality of the super-resolved image is quite close to the high-resolution version. We show in Table 2 the computational time according to the SR factor.

SR factor	2	4	8
Time (ms)	4.35	4.75	5.17

Table 2. Computational time on GPU NVidia RTX 2080ti

We now experiment the super-resolution yielded by SR-WGAN coupled with the Faster-RCNN detector [1]. Our motivation is to show the impact of object-focused super-resolution on any detector used, as YOLOv3 has been experimented in the previous section. For a comparative study, the super-resolution was performed by SR-IR and then by SR-WGAN. The results obtained for the VEDAI database are shown in Figure 5. It is now observed that the precision/recall curves from the super-resolved images approach the reference curve (purple), especially with the SR-WGAN method. The purple HR curve is the reference curve in high resolution i.e. 12.5cm/pixel, the LR (green) curve is the detector on 50cm/pixel images after bi-cubic interpolation. The blue curve is the result of Faster-RCNN on super-resolved images with SR-IR by a factor of 4, the detector was learned on HR images. The black curve is the detection on super-resolved images with SR-IR, the detector was also learned on super-resolved images. The red curve is the result of Faster-RCNN on super-resolved images with SR-WGAN by a factor of 4, the detector was learned on HR images. The figure shows that in the case of SR-WGAN a specific training is not necessary. These results also show that object-focused SR is relevant regardless of the detector used, at least with YOLOv3 and Faster-RCNN as shown in our experiments.

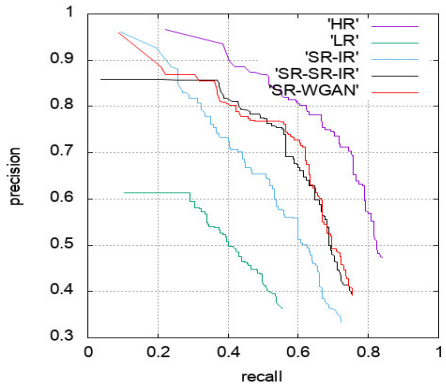

Fig. 5. Detection performance using object-focused super-resolution by SR-IR and SR-WGAN jointly with the Faster-RCNN detector (VEDAI dataset).

4. TRANSFER LEARNING

In this section, we show that object-focused super-resolution could be useful within a transfer learning context. We exploit here the SR-WGAN learned from the high-resolution ISPRS Potsdam images [9] to detect vehicles in the xView satellite images [10] with lower resolution. Figure 6 shows the results of SR-WGAN super-resolution on a sample xView image. The super-resolution adds details to the vehicles, which were learned from the high-resolution Potsdam images. We can see that since this SR is focused on the objects, it does not necessarily improve the background of the image.

We now seek to evaluate the performance of the super-resolution influencing the detection task, not the detector itself. Figure 7 shows the recall/precision curves obtained from the xView data set. The super-resolution provides a gain in

Fig. 6. Example of object reconstruction with super-resolution (factor of 4) of an xView satellite image by SR-WGAN learned from HR aerial images of the Potsdam data.

detection compared to the baseline standard images. From

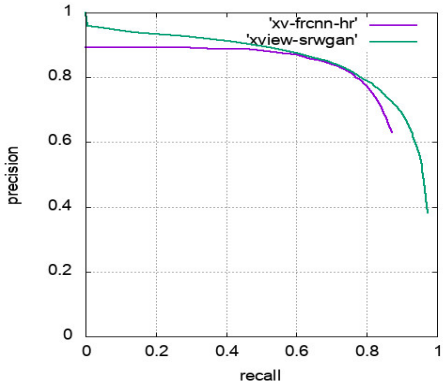

Fig. 7. Detection performance using Faster-RCNN detector on both standard and super-resolved images using SR-WGAN (xView dataset).

Figure 8, we can observe and compare the detection performance (with a confidence threshold of 0.5) on the same part of an xView image by using the original and super-resolved images: the false positives detected in red as well as the cars not detected in blue on the right original image but well detected on the super-resolved image. The original image is here zoomed in for better viewing.

Fig. 8. Detection results from super-resolved image by SR-WGAN (left) and from original image (right).

5. CONCLUSIONS

We have shown that super-resolution improves the detection of small objects on aerial or satellite images using state-of-the-art detectors such as YOLOv3 and Faster-RCNN. The improvement of this detection is more significant with super-resolution focusing on the objects to be detected. Such a technique requires only a reduced number of network layers and depends on the desired scale factor as well as the target object sizes. The proposed network architecture, derived from EDSR learned in a WGAN model, is here well adapted to this scale factor. Our future work should couple more strongly the super-resolution with detectors by ensuring a joint learning of the two neural networks (super-resolution and detector).

6. REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[2] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement." arXiv:1804.02767, 2018.

[3] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey." arXiv:1904.07523, 2019.

[4] J. Shermeyer and A. Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *CVPR-WS*, 2019.

[5] S. N. Ferdous, M. Mostofa, and N. M. Nasrabadi, "Super resolution-assisted deep aerial vehicle detection ," in *SPIE Conf. Artif. Intell. Mach. Learning Multi-Domain Operations Appl.*, vol. 11006, pp. 432 – 443, 2019.

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, pp. 4681–4690, 2017.

[7] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Visual Comm. Image Represent.*, vol. 34, pp. 187–203, 2016.

[8] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPR-WS*, pp. 136–144, 2017.

[9] F. Rottensteiner and et al., "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Annals I-3 (2012), Nr. 1*, vol. 1, no. 1, pp. 293–298, 2012.

[10] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xview: Objects in context in overhead imagery." arXiv:1802.07856, 2018.

[11] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, pp. 1874–1883, 2016.

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *NIPS*, pp. 5767–5777, 2017.