Vehicle Detection in Remote Sensing Images Leveraging on Simultaneous Super-Resolution

Hong Ji, Zhi Gao, Tiancan Mei, and Bharath Ramesh

***Abstract*— Owing to the relatively small size of vehicles in remote sensing images, lacking sufficient detailed appearance to distinguish vehicles from similar objects, the detection performance is still far from satisfactory compared with the detection results on everyday images. Inspired by the positive effects of super-resolution convolutional neural network (SRCNN) for object detection and the stunning success of deep CNN techniques, we apply generative adversarial network frameworks to realize simultaneous SRCNN and vehicle detection in an end-to-end manner, and the detection loss is backpropagated into the SRCNN during training to facilitate detection. In particular, our work is unsupervised and bypasses the requirement of low-/high-resolution image pairs during the training stage, achieving increased generality and applicability. Extensive experiments on representative data sets demonstrate that our method outperforms the state-of-the-art detectors. (The source code will be made available after the review process.)**

***Index Terms*— Faster region-based convolutional neural network (R-CNN), feature fusion, remote sensing images, super-resolution convolutional neural network (SRCNN), vehicle detection.**

I. Introduction

VEHICLE detection from remote sensing images captured by cameras mounted on satellites or airplanes has numerous useful applications in traffic, military, security, and other domains. In spite of the great achievements made, the performance of vehicle detection in remote sensing images, however, is still far from satisfactory compared with the results of the state-of-the-art detectors on benchmark data sets (e.g., ImageNet Large Scale Visual Recognition Competition [21], Microsoft common objects in context [15]). This is generally because the size of the vehicle in remote sensing image is much smaller than that of the everyday image, lacking sufficient detailed appearance to distinguish the vehicle from similar objects. As shown in Fig. 1, there is a significant gap between the detection performance of the faster region-based convolutional neural network (R-CNN) [20] on low-resolution image and its high-resolution counterpart. Inspired by the stunning success of deep CNNs (DCNNs) in both object detection and super-resolution CNN (SRCNN), it is quite natural to consider combining these two tasks in a unified framework to achieve improved detection performance.

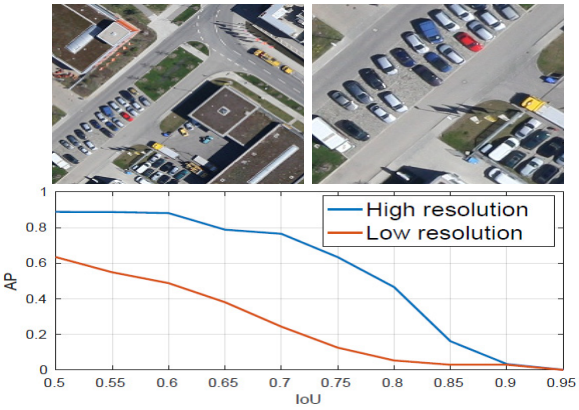

Fig. 1. (Top left) Examples of low-resolution remote sensing image and (Top right) its high-resolution counterpart. (Bottom) Vehicle detection results on low- and high-resolution images using a Faster R-CNN detector.

Traditionally, vehicle detection in the image was addressed by using low-level, handcrafted visual features (e.g., color histogram, texture, and local pattern) and classifiers such as support vector machine and AdaBoost [6], [18]. However, such handcrafted or shallow-learning-based features are usually computationally expensive and their representation power is limited. Thus, their detection performance is less competitive. Recently, inspired by the great success of region-based CNNs, such as R-CNN [8], Fast R-CNN [7], Faster R-CNN [20], and you only look once (YOLO) [19], for object detection on benchmark data sets (of everyday images), many DCNN-based methods have been proposed for vehicle detection in aerial images [2], [13], [23], [25]. Therein, promising techniques including multiscale feature fusion and hard example mining have been investigated for improved performance. The positive effects of SRCNN on object detection have been verified in [22], and methods that realize simultaneous SRCNN and object detection have been proposed [3], [9]. In [9], the Single Shot MultiBox Detector (SSD) [17] was fixed, and the detection loss was backpropagated to a SRCNN for training. In [3], a generative adversarial network (GAN) was proposed. The generator is a SRCNN and the discriminator is a multi-task network for real/fake authentication, classification, and localization. In [9] and [3], low-/high-resolution image

Manuscript received April 10, 2019; revised June 10, 2019; accepted July 17, 2019. Date of publication August 8, 2019; date of current version March 25, 2020. This work was supported in part by the Wuhan Institute Key Project under Grant 1WHS20171003. *(Corresponding authors: Zhi Gao; Tiancan Mei.)*

H. Ji and T. Mei are with the Electronic and Information School, Wuhan University, Wuhan 430072, China (e-mail: 2013301220036@whu.edu.cn; mtcwlb@aliyun.com).

Z. Gao is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Temasek Laboratories, National University of Singapore, Singapore 117411 (e-mail: gaozhinus@gmail.com).

B. Ramesh is with the Singapore Institute for Neurotechnology, National University of Singapore, Singapore 117456 (e-mail: bharathramesh@nus.edu.sg).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2019.2930308

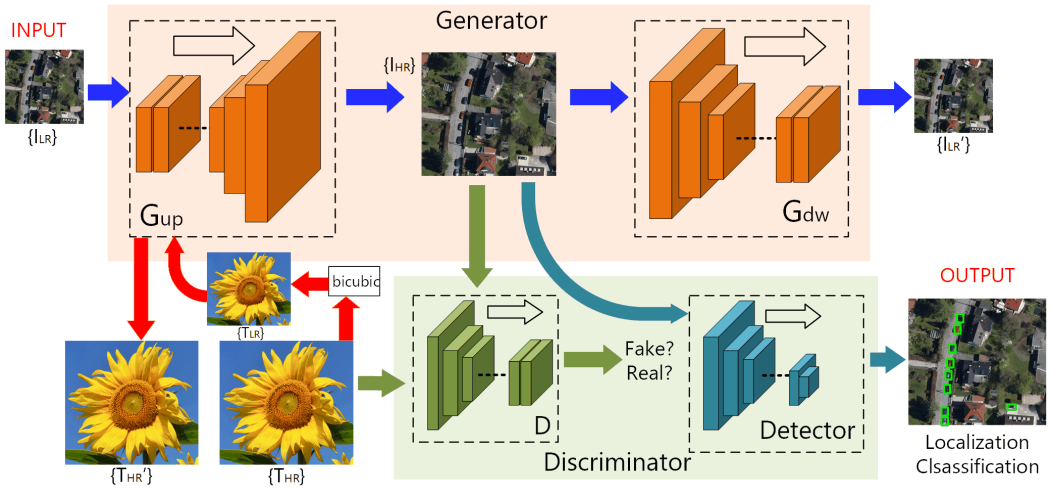

Fig. 2. Illustration of the pipeline of our method. Red region: generator networks. G_{UP} and G_{dw} correspond to the SRCNN and downsampling networks, respectively. Green region: discriminator networks. D discriminates the input image into real or fake. *Detector* completes vehicle detection task and outputs the results. I_{LR} is the input low-resolution image, I_{HR} is the super-resolved high-resolution image from I_{LR}, and I'_{LR} is of low-resolution generated from I_{HR}. T_{HR} is the high-resolution image provided as reference from the high-quality DIV2K data set [1], which is not related to any vehicle detection purpose.

pairs are required. Such a prerequisite is difficult to be met in practical applications.

Based on the above discussions that inspire our work from the beginning, we propose a network for vehicle detection in remote sensing images leveraging on simultaneous unsupervised SRCNN, in which the CycleGAN structure is investigated to super-resolve the original image to facilitate small object detection without the requirement of low-/high-resolution image pairs. Leveraging on the multitask learning strategy, the detector is assigned as a discriminator whose detection loss is backpropagated to the generator. Consequently, our generator provides the most appropriate super-resolved image under the guidance and also serves for the detector. Extensive experiments on representative data sets demonstrate that our method outperforms the state-of-the-art detectors. The source code will be made available (after the review process) to facilitate future research in the community.

II. Our Network

The pipeline of our network is shown in Fig. 2, and the details are elaborated in the following.

A. CycleGAN-Based Super-Resolution Network

SRCNN has benefited from recent advances in DCNN. SRCNN [5] enhanced the spatial resolution of an input image by handcrafted upsampling filters, followed by refinement using CCNs. To further reduce the blurring artifacts, a super-resolution GAN (SRGAN) [12] has been proposed via combining both perceptual similarity measurement and adversarial losses. As the high-resolution counterpart of the low-resolution image is available (see Fig. 3), the pixelwise loss is enforced, and the SRGAN loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{SRGAN}} = {} & \mathbb{E}_{I_{\mathrm{HR}}\sim P_{\mathrm{data}}(I_{\mathrm{HR}})}[\log(D(I_{\mathrm{HR}}))] \\ & + \mathbb{E}_{I_{\mathrm{LR}}\sim P_{\mathrm{data}}(I_{\mathrm{LR}})}[\log(1 - D(G(I_{\mathrm{LR}})))] \\ & + \mathbb{E}_{I_{\mathrm{LR}}\sim P_{\mathrm{data}}(I_{\mathrm{LR}})}[||G(I_{\mathrm{LR}}) - I_{\mathrm{HR}}||_2] \qquad (1)\end{aligned}$$

where the first two terms correspond to the GAN loss. The third term is the pixelwise mean squared error (MSE). I_{LR}, $G(I_{\mathrm{LR}})$, and I_{HR} denote the low-resolution image, the generated super-resolved image, and the real high-resolution image, respectively, which requires sufficient low-/high-resolution image pairs for training. In our CycleGAN-based network, called CycGANSR, we bypass such data preparation via replacing the pixelwise MSE term with our new term.

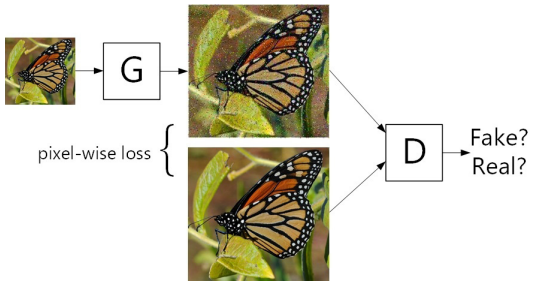

Fig. 3. Schematic of the SRGAN method.

Inspired by the development of unsupervised learning in image-to-image translation [24], we apply the CycleGAN structure for unsupervised SRCNN. The generator is shown in Fig. 4, which consists of two subgenerators G_{UP} and G_{dw}. We define the cycle-consistency MSE loss

$$\mathcal{L}_{\mathrm{cyc}} = \mathbb{E}_{I_{\mathrm{LR}}\sim P_{\mathrm{data}}(I_{\mathrm{LR}})}[||G_{\mathrm{dw}}(G_{\mathrm{UP}}(I_{\mathrm{LR}})) - I_{\mathrm{LR}}||_2] \qquad (2)$$

where I_{LR} and $G_{\mathrm{UP}}(I_{\mathrm{LR}})$ denote the original image and the super-resolved high-resolution image, respectively. $G_{\mathrm{dw}}(G_{\mathrm{UP}}(I_{\mathrm{LR}}))$ represents the downsampled image I'_{LR}, which has the same resolution as I_{LR}. As G_{UP} and G_{dw} are coupled, it is difficult to guarantee the convergence of G_{UP} to a network that we want. Thus, leveraging on the additional high-resolution images (which are from some high-quality image data set, and not related to our remote sensing image), we define an identity loss term to ensure the convergence of G_{UP}, as follows:

$$\mathcal{L}_{\mathrm{Idt}} = \mathbb{E}_{T_{\mathrm{LR}}\sim P_{\mathrm{data}}(T_{\mathrm{LR}})}[||G_{\mathrm{UP}}(T_{\mathrm{LR}}) - T_{\mathrm{HR}}||_2]. \qquad (3)$$

TABLE I
ARCHITECTURE OF UPSAMPLING GENERATOR G_{UP}

layer	conv	residual block $\times 16$	conv	element-wise sum	conv	pixelshuffle	conv	pixelshuffle	conv
kernel size	3	3	3	-	3	-	3	-	3
kernel num	64	64	64	-	256	-	256	-	64
stride	1	1	1	-	1	$\frac{1}{2}$	1	$\frac{1}{2}$	1

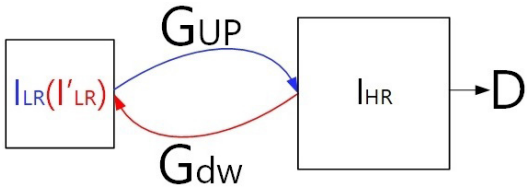

Fig. 4. Pipeline of the generator of our CycleGAN network.

TABLE II
ARCHITECTURE OF DOWNSAMPLING GENERATOR

layer	conv	conv$\times 2$	residual block $\times 6$	conv$\times 2$	conv
kernel size	7	4	3	3	7
kernel num	64	64	64	64	3
stride	1	2	1	1	1

TABLE III
ARCHITECTURE OF DISCRIMINATOR

layer	conv	conv	BN	conv	BN	conv	BN	conv
kernel size	4	4	-	4	-	4	-	4
kernel num	64	128	-	256	-	512	-	1
stride	2	2	-	2	-	1	-	1

As shown in Fig. 2, T_{LR} is the low-resolution image obtained via bicubic downsampling on the high-quality reference image T_{HR}. Here, only G_{UP} is included. Our method ensures the similarity between the input image and the super-resolved image without using any prepared low-/high-resolution pairs. The network architectures of our two generators are shown in Tables I and II, respectively.

Our discriminator D employs the ResNet-50 as the backbone network (described in Table III), which plays the role of distinguishing the real high-resolution images from the generated super-resolved images. For this specific task, sigmoid loss function is applied in the last fully connected layer. Now, we can formulate our cycle-GAN loss, which consists of GAN loss, cycle-consistency MSE loss, and identity loss

$$\mathcal{L}_{\mathrm{cycGAN}} = \mathcal{L}_{\mathrm{GAN}} + \lambda_1 \mathcal{L}_{\mathrm{cyc}} + \lambda_2 \mathcal{L}_{\mathrm{Idt}}. \tag{4}$$

B. Deep Detection Network With Multiscale Feature Fusion

In our recent work [10], we presented a multiscale feature fusion strategy for vehicle detection in remote sensing images based on Faster R-CNN. Here, we summarized the key idea and operations of our feature fusion technique. As shown in Fig. 5, we combine the feature from pyramid level 5 (C5), 4 (C4), 3 (C3), and 2 (C2) to generate the finest feature map for the subsequent region-of-interest (RoI) pooling and detection. We exclude the fifth layer (C5) for proposal detection, since the vehicle is so small and that it cannot be retained in this level. In each intermediate fused level P4, P3, and P2, we apply a CNN, to detect the regions that could possibly

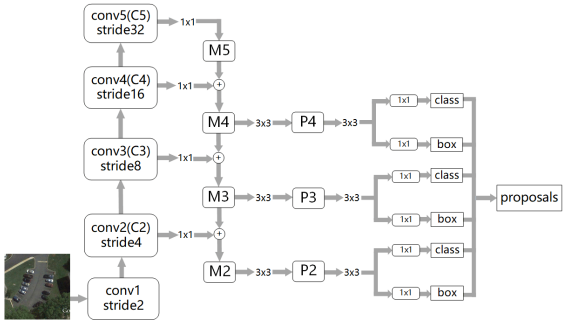

Fig. 5. Bottom-up feature extraction and top-down multiscale feature fusion for region proposal generation. M5, M4, M3, and M2 indicate the middle levels of intermediate operation. P4, P3, and P2 indicate the pyramid levels. 1×1 and 3×3 represent the convolution layer with kernel sizes 1 and 3, respectively.

contain target. For each proposal, we perform RoI pooling and then feed the pooled area to two sibling fully connected layers for classification and localization. We train the same objective function of multi-task loss as (here, λ is 1)

$$\begin{aligned}
\mathcal{L}_{\mathrm{Det}} &= \mathcal{L}_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{reg}} \\
\mathcal{L}_{\mathrm{cls}} &= \mathbb{E}_{I_{\mathrm{LR}} \sim P_{\mathrm{data}}(I_{\mathrm{LR}})}[-\log(\mathrm{Det}_{\mathrm{cls}}(G_{\mathrm{UP}}(I_{\mathrm{LR}})))] \\
\mathcal{L}_{\mathrm{reg}} &= \mathbb{E}_{I_{\mathrm{LR}} \sim P_{\mathrm{data}}(I_{\mathrm{LR}})} \\
&\quad \times [\mathrm{smooth}_{L1}(\mathrm{Det}_{\mathrm{loc}}(G_{\mathrm{UP}}(I_{\mathrm{LR}})), \mathbf{t}_*)] \\
\mathrm{smooth}_{L1}(\boldsymbol{x}) &= \begin{cases} 0.5\boldsymbol{x}^2, & \text{if } |\boldsymbol{x}| < 1 \\ |\boldsymbol{x}| - 0.5, & \text{otherwise.} \end{cases}
\end{aligned} \tag{5}$$

We now combine the SRCNN and the detection networks to formulate the overall loss function, as follows:

$$\mathcal{L}_{\mathrm{overall}} = \mathcal{L}_{\mathrm{GAN}} + \lambda_1 \mathcal{L}_{\mathrm{cyc}} + \lambda_2 \mathcal{L}_{\mathrm{Idt}} + \lambda_3 \mathcal{L}_{\mathrm{Det}}. \tag{6}$$

C. Implementation Detail

The upsampling generator is initialized by the pretrained model released from [14]. The downsampling generator and the discriminator are trained from scratch. All the generators and the discriminators are trained using an Adam optimizer [11]. Their initial learning rates are set to 0.0001 and reduced by a factor of 10 after 40 000 iterations. The batch size is 2 and the networks are totally trained for 80 000 iterations. When training generators, the parameters of the discriminator are fixed and objective function is shown as (7), just without the classification loss (third term) and localization loss (fourth term). Here, we also take λ_1 and λ_2 to control the contribution of GAN loss, cycle-consistency loss, and identity loss. Through experiments, we figure out that these terms are of the same importance for the whole model optimization. Therefore, λ_1 and λ_2 are both set to 1. For training discriminator, we fix the generators and the objective function is shown as (8), but without detector loss (second and third

TABLE IV
KEY INFORMATION OF OUR DATA SETS

Datasets	HR size	#TRI	#TEI	#TRV	#TEV
DLR Munich	600*600	924	369	10770	3304
UCAS-AOD	600*600	3660	908	24873	5700

*TRI indicates training image. TEI indicates testing image. TRV indicates training vehicle. TEV indicates testing vehicle

TABLE V
RESULTS OF MUNICH DLR DATA SET

Method*	AP	AP@0.5	AP@0.75	mRecall	Time (s)
R-FCN	0.321	0.613	0.303	0.396	0.018
SSD	0.249	0.521	0.212	0.258	0.093
YOLOv3	0.262	0.574	0.186	0.273	0.086
FRCNN	0.342	0.691	0.292	0.362	0.027
FRCNN+Bicubic	0.487	0.795	0.571	0.554	0.078
FRCNN+EDSR	0.450	0.784	0.530	0.538	0.095
FRCNN+CycGANSR	0.541	0.801	0.658	0.628	0.100
Ours	**0.599**	**0.889**	**0.684**	**0.648**	0.101

*FRCNN indicates the Faster R-CNN++ method, which is the improved version of Faster R-CNN with FPN.

terms). The detector is initialized with ResNet-50 trained on ImageNet and trained with a stochastic gradient descent optimizer. Initialized learning rate is 0.0025 and it reduced to 0.00025 after 40 000 iterations. It is totally trained for 60 000 iterations. We adopt two images for training per mini-batch

$$\begin{aligned}\arg\min_{G^*}\ & \frac{1}{N}\sum_i \left\|D\left(G_{\mathrm{UP}}\left(I^i_{\mathrm{LR}}\right)\right)-1\right\|_2 \\ &+\frac{1}{N}\sum_i \lambda_1\left\|G_{\mathrm{dw}}\left(G_{\mathrm{UP}}\left(I^i_{\mathrm{LR}}\right)\right)-I^i_{\mathrm{LR}}\right\|_2 \\ &+\frac{1}{N}\sum_i \lambda_2\left\|G_{\mathrm{dw}}\left(G_{\mathrm{UP}}\left(T^i_{\mathrm{LR}}\right)\right)-T^i_{\mathrm{HR}}\right\|_2 \\ &+\frac{1}{N}\sum_i -\lambda_3\log\left(\mathrm{Det}_{\mathrm{cls}}\left(G_{\mathrm{UP}}\left(I^i_{\mathrm{LR}}\right)\right)\right) \\ &+\frac{1}{N}\sum_i \lambda_3[u^i\geq 1]\left(\mathrm{Det}_{\mathrm{reg}}\left(G_{\mathrm{UP}}\left(I^i_{\mathrm{LR}}\right),\mathbf{t}^{\mathbf{i}}_*\right)\right. \qquad (7)\end{aligned}$$

$$\begin{aligned}\arg\min_{D^*}\ & \frac{1}{N}\sum_i \left(\left\|D(G_{\mathrm{UP}}\left(I^i_{\mathrm{LR}}\right))\right\|_2+\left\|D\left(T^i_{\mathrm{HR}}\right)-1\right\|_2\right) \\ &+\frac{1}{N}\sum_i -\omega\log\left(\mathrm{Det}\left(G_{\mathrm{UP}}\left(I^i_{\mathrm{LR}}\right)\right)\right) \\ &+\frac{1}{N}\sum_i \omega[u^i\geq 1]\left(\mathrm{Det}_{\mathrm{reg}}\left(G_{\mathrm{UP}}\left(I^i_{\mathrm{LR}}\right),\mathbf{t}^{\mathbf{i}}_*\right)\right. \qquad (8)\end{aligned}$$

After training the CycGANSR network and the detection network, we train them jointly. The training procedure is same as the CycGANSR and its objective functions are similar to those in (7) and (8). λ_3 and ω are set to 0.01 and 0.1, respectively.

III. EXPERIMENTS AND ANALYSIS

To evaluate the performance, we conduct extensive experiments and comparison against the state-of-the-art approaches.

TABLE VI
RESULTS OF UCAS-AOD DATA SET

Method*	AP	AP@0.5	AP@0.75	mRecall	Time (s)
R-FCN	0.316	0.605	0.297	0.391	0.022
SSD	0.264	0.566	0.188	0.286	0.089
YOLOv3	0.281	0.593	0.196	0.311	0.083
FRCNN	0.337	0.682	0.288	0.362	0.029
FRCNN+Bicubic	0.481	0.805	0.526	0.569	0.082
FRCNN+EDSR	0.486	0.804	0.526	0.559	0.097
FRCNN+CycGANSR	0.516	0.804	0.611	0.594	0.099
Ours	**0.572**	**0.885**	**0.637**	**0.653**	0.103

*FRCNN indicates the Faster R-CNN++ method, which is the improved version of Faster R-CNN with FPN.

A. Preparation

1) Data: DLR Munich data set [16] is taken over the area of Munich, using DLR 3K camera system. It contains 20 images (of resolution 5616×3744 pixels), with approximately 13-cm ground sampling distance (GSD). UCAS-AOD data set [26] includes 510 satellite images with resolution 659×1280, but without GSD information. All the images are randomly divided into 410 training and 100 testing images. Based on observation, the sizes of vehicles in this data set are usually larger than that of the Munich data set, but the quality is much poorer. The key information of these data sets related to our task is summarized in Table IV.

2) Methods for Comparison: We perform a comparison with the state-of-the-art detectors, Faster R-CNN++, YOLOv3, region-based fully convolutional network (R-FCN) [4], SSD, and their combinations with different SRCNN modules including the basic bicubic upsampling, enhanced deep SRCNN (EDSR) [14], and our CycleGANSR component. Note that when we mention our CycleGANSR component in the experiment part, it indicates our CycleGAN-based SRCNN where the detection loss is not backpropagated for training.

B. Experimental Results

We report the vehicle detection results visually and numerically. Tables V and VI demonstrate the detection results on the Munich and UCAS-AOD data sets, respectively, with the metrics of average precision (AP), AP at different intersections of union (IoU) levels, and mean of recall (mRecall). First, we test the detection performance of R-FCN, SSD, YOLOv3, and Faster R-CNN++ on the input low-resolution images (without any SRCNN operation), and the results are poor. In particular, Faster R-CNN++ outperforms those one-step detectors SSD, YOLOv3, and R-FCN. Therefore, we combine Faster R-CNN++ with different SRCNN modules and compare with our method. Clearly, when the Faster R-CNN++ detector is combined with our CycGANSR component, the detection results are better than other combinations, indicating the benefits of our SRCNN component. Moreover, our method where the SRCNN and the detection network are jointly trained achieves the best results for all metrics and obtains about 5% better AP than the second best result. When the level of IoU is low, the superiority of our method is more obvious. In addition, our method obtains better result on the Munich data set than that of the UCAS-AOD data set, except the metric of mRecall.

Fig. 6. Examples of detection results of our method on the Munich (first row) and UCAS-AOD (second row) data sets. Green boxes: correctly detected vehicles. Red boxes: false alarms. Blue boxes: missing alarms.

This is for the reason that the image quality of the Munich data set is better, resulting in less false alarm, although missing may still happen. Some examples of the detection results are shown in Fig. 6. In our work, all codes are implemented using Caffe2 and run on an NVIDIA GeForce GTX1080Ti with 12 GB on-board memory. We report the average processing time for an input of size 200×200.

IV. Conclusion

In this letter, we realize simultaneous SRCNN and vehicle detection in remote sensing images in an end-to-end manner. In particular, our method bypasses the requirement of low-/high-resolution image pairs via applying the CycleGAN strategy, achieving increased generality and applicability. We are going to increase the robustness and generalization ability of our method.

References

[1] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.

[2] N. Ammour, H. Alhichri, Y. Bazi, B. Benjdira, N. Alajlan, and M. Zuair, "Deep learning approach for car detection in UAV imagery," *Remote Sens.*, vol. 9, no. 4, p. 312, 2017.

[3] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 210–226.

[4] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2016, pp. 379–387.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[6] M. ElMikaty and T. Stathaki, "Detection of cars in high-resolution aerial images of complex urban environments," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5913–5924, Oct. 2017.

[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[9] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," Mar. 2018, *arXiv:1803.11316*. [Online]. Available: https://arxiv.org/abs/1803.11316

[10] H. Ji, Z. Gao, T. Mei, and Y. Li, "Improved faster R-CNN with multiscale feature fusion and homography augmentation for vehicle detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, to be published.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[12] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.

[13] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R^3-net: A deep network for multi-oriented vehicle detection in aerial images and videos," Aug. 2018, *arXiv:1808.05560*. [Online]. Available: https://arxiv.org/abs/1808.05560

[14] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 1132–1140.

[15] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[16] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.

[17] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[18] T. Moranduzzo and F. Melgani, "Detecting cars in UAV images with a catalog-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6356–6367, Oct. 2014.

[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[21] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[22] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," Dec. 2018, *arXiv:1812.04098*. [Online]. Available: https://arxiv.org/abs/1812.04098

[23] H. Tayara, K. G. Soo, and K. T. Chong, "Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network," *IEEE Access*, vol. 6, pp. 2220–2230, 2018.

[24] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, p. 814.

[25] J. Zhong, T. Lei, and G. Yao, "Robust vehicle detection in aerial images based on cascaded convolutional neural networks," *Sensors*, vol. 17, no. 12, p. 2720, 2017.

[26] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3735–3739.