remote sensing

MDPI

Article

Small Object Detection in Remote Sensing Images with Residual Feature Aggregation-Based Super-Resolution and Object Detector Network

Syed Muhammad Arsalan Bashir [1,2,*] and Yi Wang [1]

[1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China; wangyi79@nwpu.edu.cn
[2] Pakistan Space & Upper Atmosphere Research Commission, Karachi, Sindh 59201, Pakistan
* Correspondence: smarsalan@mail.nwpu.edu.cn or smab1176@yahoo.com

Abstract: This paper deals with detecting small objects in remote sensing images from satellites or any aerial vehicle by utilizing the concept of image super-resolution for image resolution enhancement using a deep-learning-based detection method. This paper provides a rationale for image super-resolution for small objects by improving the current super-resolution (SR) framework by incorporating a cyclic generative adversarial network (GAN) and residual feature aggregation (RFA) to improve detection performance. The novelty of the method is threefold: first, a framework is proposed, independent of the final object detector used in research, i.e., YOLOv3 could be replaced with Faster R-CNN or any object detector to perform object detection; second, a residual feature aggregation network was used in the generator, which significantly improved the detection performance as the RFA network detected complex features; and third, the whole network was transformed into a cyclic GAN. The image super-resolution cyclic GAN with RFA and YOLO as the detection network is termed as SRCGAN-RFA-YOLO, which is compared with the detection accuracies of other methods. Rigorous experiments on both satellite images and aerial images (ISPRS Potsdam, VAID, and Draper Satellite Image Chronology datasets) were performed, and the results showed that the detection performance increased by using super-resolution methods for spatial resolution enhancement; for an IoU of 0.10, AP of 0.7867 was achieved for a scale factor of 16.

Keywords: object detection in satellite images; image classification; vehicle detection; remote sensing; deep learning; generative adversarial networks; residual feature aggregation (RFA)

Citation: Bashir, S.M.A.; Wang, Y. Small Object Detection in Remote Sensing Images with Residual Feature Aggregation-Based Super-Resolution and Object Detector Network. *Remote Sens.* **2021**, *13*, 1854. https://doi.org/10.3390/rs13091854

Academic Editors: Pedro Melo-Pinto and Mercedes E. Paoletti

Received: 25 March 2021
Accepted: 7 May 2021
Published: 10 May 2021

1. Introduction

1.1. Objective and Research Problem

The detection of objects of interest in remote sensing images has been challenging, especially in small objects, because of the limitation of ground sampling distance (GSD). For small objects, such as vehicles, a few pixels represent the whole object; thus, identification and detection become challenging. A very high resolution (VHR) satellite image with a GSD of 25 cm will have a 96 pixel (16 × 6) areal grid for a vehicle with a dimension of 4 × 1.5 m^2. Recently developed deep-learning-based models that tackle the problem of small object detection include Faster RCNN [1], one-stage detector [2], semantic-context-aware network [3], and end-to-end MSCNN (multiscale convolutional neural network) [4]. Alganci et al. [5,6] also compare the object detection networks for further insight into object detection in satellite images; this review is referred to the reader. Researchers have made efforts to adapt the current object detectors to solve object detection in VHR satellite images. In [7], a tree-based CNN was used to detect objects in VHR satellite images; the output layer was fully connected while a single contained a convolution and a pooling layer. In [4], the EssNet was proposed to ensure that the small objects in VHR images are given enough attention using multiscale object detection that was not possible in the fusion–feature

pyramid detection [2] because the deep feature levels dominated the learning process and small objects were lost in the learning process. In [8], multi-spectral features were used to improve object detection accuracy; however, the authors used this method for large object detection, including gold course and a power plant. A deconvolution layer-based Deconv R-CNN was used in [9] after the final convolutional layer, resulting in better localization of the small-sized object features.

1.2. Background and Related Literature

The studies are classified as end-to-end learning for object detection or using a prior network to enhance the image quality before object detection. Thus, for detection of small objects in satellite/aerial images, multiscale learning with feature attention is being used along with the state-of-the-art object detection methods, for instance, Fast/Faster RCNN [10] (region-based convolutional neural network), Feature Pyramid Nets [11], You Only Look Once (YOLOv3 [12], YOLOv4 [13], YOLOv5) Single Shot Detectors (SSD) [14], and RetinaNet [15].

Compared to Faster R-CNN [10], the Deconv R-CNN reported an increase of 13% in mean average precision (mAP) to detect objects. Intersection-over-union (IoU) represents the overall between the actual and detected object boundaries, and the detection performance is dependent on the selection of IoU value; for small objects, choosing a smaller IoU usually results in higher detection accuracy while at the same time, it causes poor localization due to an increased number of false positives [16]. An adaptable IoU-based Deformable R-CNN was proposed in [16], which increased the detection performance by 4.8% compared to baseline R-CNN. In [17], an optimal dense YOLOv2 based method called DOLO was developed to detect small targets from images generated by unmanned aerial vehicles (UAVs); the small targets which YOLOv2 and SSD poorly detected were detected by DOLO (three blocks), and the authors reported the mean average precision (mAP) of 0.762. In [17], Faster R-CNN and YOLOv3 reported a higher mAP of 0.817 and 0.867, respectively. Another method for detecting small objects in UAV was developed by Liu et al. [18], termed UAV-YOLO. In [18], the original darknet in YOLOv3 was modified to incorporate two Resblocks; furthermore, the authors increased the convolution operations within the initial layers to enable the network to collect more spatial information. The method was optimized to perform better on UAV images while it performed similarly to YOLOv3 and SSD on datasets, such as VOC and COCO. In [19], the authors developed a Tiramisu and YOLO-based method for counting vehicles for the images with a ground sampling resolution of 50 cm. A semi-automated annotation method was used to segment and annotate the 87 k vehicles in the dataset, creating a bottleneck in training, as the annotation would take more time if the dataset was changed to images from a different satellite sensor. While in [20], the authors introduced lower sub-sampling factors of four and two in the YOLOv3 model to give attention to small-sized objects, which resulted in improved detection of smaller objects due to smaller detection grids. In [21], Wang et al. proposed SSS-YOLO with dilated Resblocks, which ensures a large receptive field, thereby capturing the ships' information, using a path augmentation fusion network (PAFN). The proposed strategy worked well on SAR datasets, SAR-Ship-Dataset and Gaofen-3, by performing better than YOLOv3 and Faster R-CNN in mAP and inference time.

As mentioned earlier, an alternate way is to perform pre-processing on the dataset by improving the resolution of the images using image SR methods, thereby increasing the level of details in the images before performing object detection as performed by Courtrai et al. [22]. Various techniques help in generating a high-resolution (HR) image from a low-resolution (LR) image where methods, such as single-image super-resolution [23], perform the task using a single image as input to the network. In [24], authors have reviewed deep-learning-based image SR methods, including convolutional neural network (CNN)-based methods, such as SR-CNN [25], deeply-recursive convolutional network (DRCN) [26], fast SRCNN (FSRCNN) [27], and very deep super-resolution (VDSR) [28]. Residual learning-based SR methods include the Laplacian pyramid SR

network (LapSRN) [29], enhanced deep SR network (EDSR) [30], and EnhanceNet [31]. At the same time, some of the current state-of-the-art methods include recurrent back-projection network (RBPN) [32], second-order attention network (SAN) [33], and the wavelet-based residual attention network (WRAN) [34]. In [24], further single image SR methods are explored.

In [35], the authors used progressively supervised learning to detect salient objects in remote sensing images (RSI), using a pseudo-label generation method to train the network in a weakly supervised way while in [36], the authors used a generative adversarial network (GAN) for super-resolution and further used a second stage for detection using SSD with minor modifications. The results of [36] depict that image SR improved the detection and classification performance compared to the LR images. Furthermore, in [22], Courtrai et al. used EDSR as a base network to develop a YOLOv3-based cyclic GAN, which effectively worked for a scale factor of eight in LR images with a GSD of 1 m. In [37], Rabbi et al. combined the enhanced SRGAN (ESRGAN) [38] and edge-enhance GAN (EEGAN) [39] to develop an end-to-end method edge-enhanced SRGAN (EESRGAN). The output of the EESRGAN was fed to the detection network, which was comprised of a faster, region-based convolutional network (FRCNN) and an SSD. The method in [37] improved the detection and classification performance for VHR satellite images with a resolution of 15 cm and 30 cm, using a scaling factor of two and four.

1.3. Proposed Method

For a very low-resolution (LR) image, there is a need for a higher scaling factor, which in the case for residual learning-based methods, which require more residual blocks to capture adequate details from the training dataset and generate a high-resolution image. Therefore, training a single network for detection tasks using conventional optimization based on mean absolute error (MAE) or mean square error (MSE) would be a complex task. In this research, the goal is to utilize super-resolution as a base model to improve VHR images from satellites/aerial vehicles and then perform detection of small objects, i.e., vehicle detection.

We propose a super-resolution-based cyclic GAN for small object detection in remoter sensing and aerial images with residual feature aggregation to enhance the image quality, first using a weighted loss function of the generator, discriminator, and YOLO detector. The proposed method is an improved image super-resolution network with a detector to perform the final task of small object detection. The proposed method incorporates three improvements: RFA based residual blocks for improved super-resolution under very high scaling factors, a framework for object detection independent of the object detector network, and a cyclic approach for improved training detection performance.

1.4. Organization of Research

The paper is organized as follows. Section 2 introduces the actual problem by discussing the EDSR architecture and its performance on LR images and how the scaling factor affects the detection network's performance. Furthermore, the Residual Feature Aggregation is introduced in this section within the EDSR architecture which improved the generated HR image. In Section 3, we propose improving the base EDSR network to increase the overall performance, using a GAN-based cyclic approach while using the loss function of YOLOv3 to train the proposed network. In Section 4, we discuss and share the results of regressive experiments conducted to measure the proposed cyclic super-resolution generative adversarial network's performance with the residual feature aggregation network and YOLO (SRCGAN-RFA-YOLO) on the ISPRS Potsdam data [40]. The research problem is to detect and classify vehicles in remote sensing images whose ground resolution has been artificially reduced by a factor of 16, i.e., from a GSD of 5 cm/pixel to 80 cm/pixel. For a vehicle of size 4×1.5 m^2, this corresponds to 5×2 pixels having a total of 10 pixels. In Section 4.1, we report the performance with the methods discussed in this paper detection to validate our method's performance. In Section 4.2, we compare the

proposed method with other state-of-the-art object detectors. In Section 4.3, we show that this method can be generalized to other datasets using the concept of transfer learning for vehicle detection without any need for training, using two datasets, i.e., Draper Satellite Image Chronology [41] and VAID [42] datasets. Section 5 discusses the results while the conclusions and future research plans are shared in Section 6.

2. Residual Learning for the Image Super-Resolution

This section introduces some basic image super-resolution methods preceding the state-of-the-art method EDSR [30], which is the fundamental model in this research. In recent studies [22,37,43,44], the EDSR model has been used primarily in remote sensing where the objects are of various scales and detection is difficult because of the images' multiscale nature. The selection of the super-resolution method can be changed, as the proposed framework is generic. The image SR focuses on the recovery of HR images from LR image input, as, in principle, the recovered HR image I_{RHR} can be represented as the output of the super-resolution function φ as shown in Equation (1):

$$I_{RHR} = \varphi(I_{LR}, \partial) \quad (1)$$

where φ is the image SR function responsible for converting the LR image to HR image, I_{LR} is the input LR image, whereas ∂ depicts the parameters of the image SR function. Image SR parameters are usually features extracted from the network training on various LR–HR image pairs. One of the fundamental architectures for image SR is the convolutional neural networks (CNNs), which are stacked in layers to perform feature extraction followed by a pixel rearrangement layer that also performs upsampling. Using the extracted feature maps, the pixel rearrangement layer performs rearrangement of the pixels within the image and transforms the input feature maps of dimensions {Bt, Ch^2, H, W} to {Bt, Ch, H $\times$ f, W $\times$ f}. Here, Bt, Ch, H, W represent the batch number, channel number, height, and width of a feature map, while f is the scaling or the upsampling factor. For doubling the feature dimensions in the x and y-axis, a scaling factor of two is required, which will double the feature maps' dimension in the pixel rearrangement layer as performed by Shi et al. [45] and SR-CNN [46]. Compared to simple interpolation methods, such as bicubic interpolation or the patch-based approach [47], these methods performed significantly better in image quality and inference time due to a relatively shallow network architecture, i.e., five convolutional layers, as seen in [45].

Instead of using the conventional CNNs, we can opt for residual learning, using residual blocks where each residual block contains convolution, normalization, and activation layers. Furthermore, the input information is added to the feature maps extracted by the residual maps, as in the EDSR method [30], shown in Figure 1. Figure 1 shows a standard EDSR model for a scaling factor of four where a total of four residual blocks are used, and each residual block has a convolution layer (in green), the normalization layer (in yellow), ReLU activation layer (in blue) and pixel rearrangement layer (in brown). The pixel rearrangement layer in Figure 1 is responsible for a scaling factor of two. Thus, two layers create an overall scaling factor of four.

In Figure 1, the generic residual blocks fuse the residual features with the identity features before propagation to the next module. Thus, the later blocks can only see fused features, ignoring the basic residual features, leading to performance degradation in models using these blocks. A recent approach to address this issue is residual feature aggregation (RFA) [48] that concatenates the local features within the residual features, resulting in improved representation of the features, which improves the final HR image. Figure 2 shows a four-block RFA-based residual network, and the local features of each block are aggregated in the final output, using a 1 $\times$ 1 convolution layer. This section will show how RFA improves performance metrics, such as peak signal-to-noise ratio (PSNR) and structural similarity index metric (SSIM), for image super-resolution. Furthermore, the dataset used in this study is the ISPRS 2D Semantic Labeling Contest dataset of the city of

Potsdam [40]; the whole dataset contains VHR images (6000 × 6000 pixels per patch) of the city divided into 38 patches of the same size with a GSD of cm/pixel.

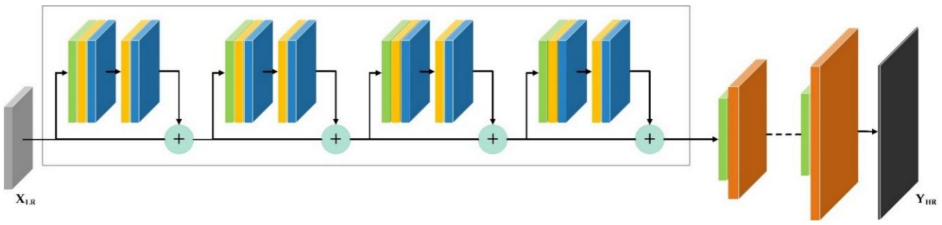


Figure 1. The pipeline of EDSR super-resolution architecture with four residual blocks: the **green** color represents the convolution layers. The **yellow** color represents the normalization layer and the **blue** color represents the ReLU activation layer while the **brown** color represents the pixel rearrangement layer. The pipeline's input is shown in **grey** and labeled as X_{LR} (which is a low-resolution image) while the output, represented in black, is the high-resolution images Y_{HR}.

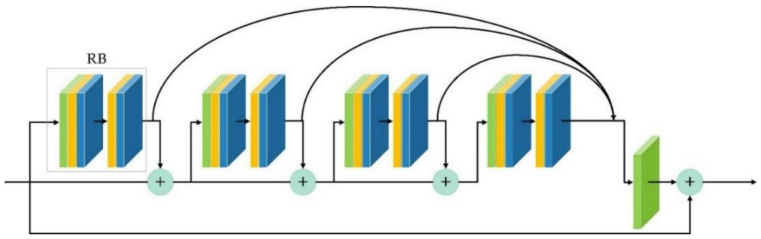


Figure 2. A super-resolution pipeline using Residual Feature Aggregation (RFA) blocks. Color coding is the same as Figure 1. The outputs of all RB (shown in the grey box) are aggregated at the output using a 1 × 1 convolution layer.

The ISPRS Potsdam dataset was used to evaluate and rank segmentation methods based on their capability of classifying the objects into six classes, namely, buildings, vehicles, surface, vegetation, trees, and background/miscellaneous objects. However, this dataset has been used in studies such as [22,49] to perform vehicle detection, using the classification information of vehicle class within the dataset. Overall, there are 2244 vehicles, including 1990 cars, 33 trucks, 181 vans, and 40 pickups; we shall be using the vehicle class, especially the pixel information of cars, for vehicle detection. The images used were in the RGB band, and the ground sampling distances of images used in this study are 5 cm/pixel (original resolution of the dataset), 40 cm/pixel (downscaled by a factor of 8), and 80 cm/pixel (downscaled by a factor of 16). The original patches are reference HR images, and LR–HR image pairs were generated using two degradation functions: bicubic and blur degradation as defined by [50], using an EDSR architecture with 16 residual blocks fixed block size of 64 × 64 and for both conventional residual blocks and RFA Blocks. A normalization layer is used in EDSR to normalize the pixels in three color bands during the training phase, and an L1 loss function was used to optimize the SR network along with Adam optimizer; for a total of N LR–HR image pairs, the loss function (with parameters p) is given in Equation (2).

$$L(p) = \frac{1}{N}\sum_{i=1}^{N}\left\|\varphi\left(I_{LR}^{i}\right) - I_{HR}^{i}\right\|_{1} \quad (2)$$

Figure 3 depicts super-resolution with a scaling factor of eight, generating an HR image with a GSD of 5 cm/pixel from an LR image of GSD 40cm/pixel, using bicubic interpolation, EDSR, and EDSR with RFA (EDSR-RFA). Significant improvement in image quality assessment (IQA) metrics, i.e., PSNR and SSIM, are observed compared to bicubic interpolation, and RFA further improves the quality of the image compared to the EDSR

method. To further assess the performance, we used the YOLOv3 [12] detector for detecting vehicles using the SR images and compared the results with the detection performance on the HR version of the images. For evaluation, we used Intersection over Union (IoU) criterion, which measures the detected box's overlapping on the SR images with the ground truth box. The threshold value for object detection was set to an IoU of 0.25 because objects in LR images occupy few pixels, and a low value of IoU would improve detection performance. Thus, a low IoU would be preferable in this scenario. A confidence score in the range of 0 to 1 was also used for detection. A box with an IoU and confidence score value greater than 0.25 was considered a detected object. For object detection evaluation, we have used the metrics of true positive (TP), false positive (FP), F1 score, and AP (average precision), using two degradation models, i.e., blur degradation (BD) and bicubic degradation (BI), as seen in Table 1. EDSR achieved better results than bicubic interpolation-based detection, while EDSR–RFA achieved the best results with AP of 0.889 and 0.844 for BI and BD degradations, respectively. The image quality metrics of the reconstructed images are summarized in Table 2 for scale factors of 8 and 16.

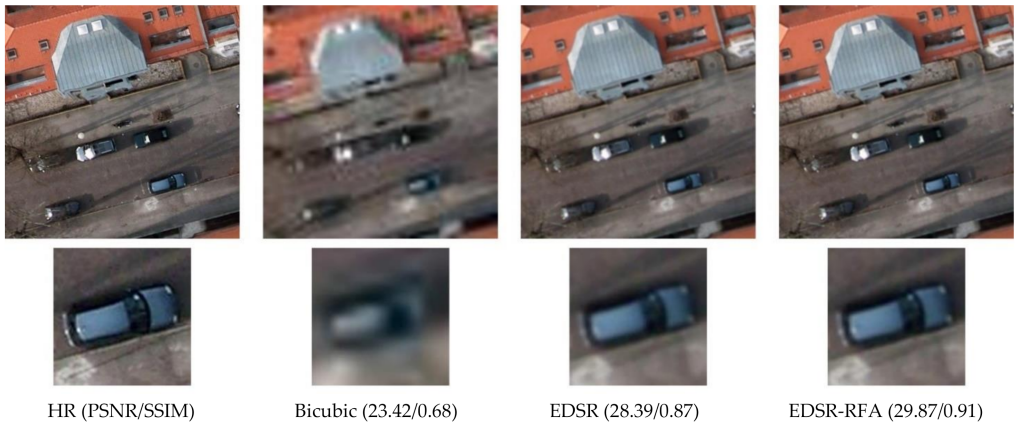


Figure 3. An illustration of the results provided by EDSR and EDSR-RFA: high-resolution (HR) image at the top left has a ground resolution of 5 cm/pixel while the LR version was generated using a scale factor of 8, which corresponds to a ground resolution of 40 cm/pixel, image super-resolution at 5 cm/pixel using bicubic interpolation, EDSR with four residual blocks and EDSR with residual feature aggregation (EDSR-RFA). Top: full image 512 × 512 pixels, Bottom: the zoomed image of 100 × 100 pixels showing a black car.

Table 1. Vehicle detection using YOLOv3 (Confidence and IoU threshold = 0.25): comparing results for HR images, a scaled image with a scale factor 8× using Bicubic interpolation, EDSR, and EDSR-RFA.

Degradation Model	Method	TP	FP	AP	F1 Score
	HR	1390	55	0.962	0.963
Bicubic Degradation	Bicubic	976	393	0.713	0.683
	EDSR	1176	177	0.869	0.864
	EDSR-RFA	1261	145	0.897	0.889
Blur Degradation	Bicubic	934	422	0.689	0.656
	EDSR	1044	240	0.813	0.798
	EDSR-RFA	1213	211	0.852	0.844

Table 2. Bicubic degradation result of super-resolution for downscaling factors of 8 and 16, indicating PSNR and SSIM of the generated HR images from bicubic interpolation, EDSR, and EDSR-RFA.

Scale Factor	Method	PSNR (dB)	SSIM
8×	Bicubic	23.42	0.6814
	EDSR	28.39	0.8732
	EDSR-RFA	29.87	0.9081
16×	Bicubic	15.74	0.4531
	EDSR	18.74	0.6240
	EDSR-RFA	19.47	0.6604

It is worth noticing that for a scaling factor of 8, the detection performance is adequate, but as we further increase the scaling factor to 16 (which corresponds to a GSD of 80 cm/pixel), the detection performance degraded drastically. Both EDSR and EDSR-RFA performed poorly on LR images generated using a factor of 16, as shown in Figure 4, which shows that the reconstructed images are of poor quality.

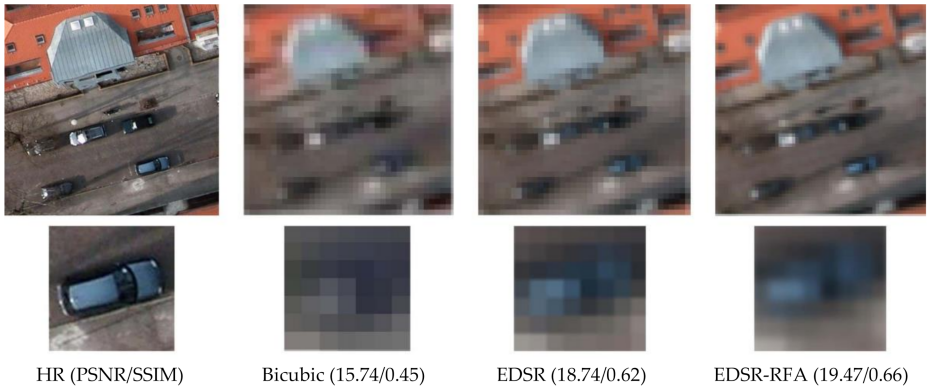


Figure 4. Comparing SR images generated by EDSR and EDSR-RFA: HR image at the top left has a ground resolution of 5 cm/pixel while the LR version was generated using a scale factor of 16, which corresponds to a ground resolution of 80 cm/pixel, image super-resolution at 5 cm/pixel using bicubic interpolation, EDSR, and EDSR-RFA. Top: full image 512 × 512 pixels, Bottom: the zoomed image of 100 × 100 pixels showing a black car.

For a scale factor of 16, the best image quality metrics (PSNR:19.47dB and SSIM: 0.66), as shown in Figure 4, were reported by EDSR-RFA, leading to a detection AP of 0.308 and 0.203 for bicubic and blur degradations, respectively (See Table 3). Although employing image super-resolution enhanced the LR image's resolution, it was still not enough for the detector to perform object detection. Furthermore, as mentioned in previous studies, adding more residual blocks to the network improves the quality of images, so we added 16 more blocks to the network pipeline, and the total number of residual blocks was changed to 32. The block size was also changed from 64 × 64 to 128 × 128; these changes resulted in an increase in the detection AP at the cost of higher training time and a significant increase in network parameters. The training time increased by 445%, and the network parameters increased from 1,145,671 to 3,742,897 (327%) while the image quality (see Figure 5) and detection AP increased (see Table 4). The detection AP increased by ~33% for both bicubic and blur degradations, but the overall detection AP was still low as compared to detection AP for the true HR images (i.e., 0.962).

The EDSR-RFA with 32 residual blocks still comes short in terms of object detection, and due to the limitation of optimization using Adam optimizer, the overall training remains unstable, thereby providing different results for the same distribution of dataset for a block size of 128 × 128. Along with this issue, there is a significant increase in network parameters and training time; thus, adding more residual blocks are not an optimal choice.

In the next section, we have introduced some modifications in the base EDR-RFA method to enhance the detection accuracy with a more stable image super-resolution.

Table 3. Vehicle detection using YOLOv3 SPP (both confidence threshold and IoU threshold were set to 0.25): comparison of results for HR image, the scaled image with a scale factor of 16×, using bicubic interpolation, EDSR, and EDSR-RFA.

Degradation Model.	Method	TP	FP	AP	F1 Score
	HR	1390	55	0.962	0.963
Bicubic Degradation	Bicubic	24	97	0.198	0.028
	EDSR	27	86	0.239	0.034
	EDSR-RFA	28	63	0.308	0.036
Blur Degradation	Bicubic	16	123	0.115	0.018
	EDSR	23	104	0.181	0.028
	EDSR-RFA	25	98	0.203	0.031

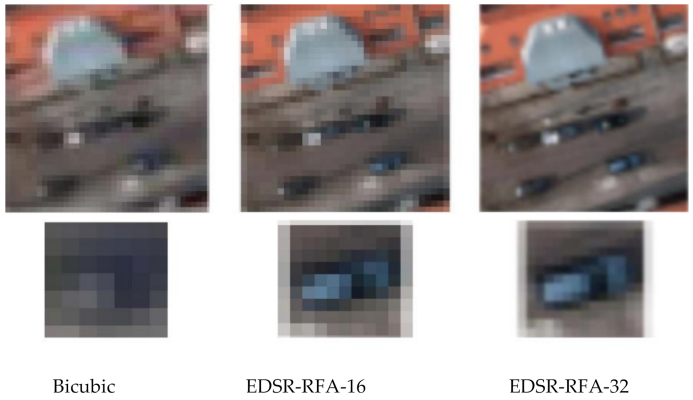


Figure 5. An illustration of the results for changing the number of blocks and block size for scale factor 16.

Table 4. The impact of the total residual block and block size on vehicle detection of EDSR-RFA with a scale factor of 16. For RFA-16, the block size is 64 × 64, while the block size for RFA-32 is 128 × 128.

Degradation	Method	TP	FP	AP	F1 Score
	HR	1390	55	0.962	0.963
Bicubic	EDSR-RFA-16	27	86	0.239	0.034
Degradation	EDSR-RFA-32	173	130	0.571	0.206
Blur	EDSR-RFA-16	23	104	0.181	0.028
Degradation	EDSR-RFA-32	154	148	0.510	0.175

3. Methodology

3.1. Basic Network Architecture for Image SR

Like other SR networks, we use the generic three-part network, namely, the initial layer for extraction of shallow features, the middle part consisting of residual blocks for further extraction of features while the *end part* performs image reconstruction. The shallow features generated by the initial layer (F_0) is given by the following:

$$F_0 = S(I_{LR}) \tag{3}$$

where S is the feature extraction function of the initial layer while F_0 is the extracted feature; this is transferred to the residual blocks for deep feature learning based on residual feature

aggregation using the residual blocks shown in Figure 2. The trunk part is comprised of 48 RFA residual blocks, and for K residual blocks, the feature functions are related as shown below:

$$F_k = M_k(F_{k-1}) = M_k(M_{k-1}(\ldots(M_0(F_0))\ldots)) \quad (4)$$

where M_k represents the k^{th} residual module, which has the input F_{k-1} and output F_k, the reconstruction part is responsible for upscaling the extracted feature F_k using

$$I_{SR} = \Re(F_k + F_0) = \varphi(I_{LR}, \partial) \quad (5)$$

where I_{SR} is the super-resolved image while $\Re$ is the reconstruction function which is equal to $\varphi(I_{LR}, \partial)$ as shown in Equation (1), this is the overall SR function of the network.

The proposed network was optimized using the L1 loss function as shown in Equation (1).

3.2. Network Improvements

In this section, further improvements in the primary network proposed in Section 3.1 are proposed to enhance the overall detection accuracy for small objects in aerial/satellite images. The overall network is improved by making cyclic GAN, using the improved Wasserstein generative adversarial network [50], and the generator of the network is defined in Section 3.1. Finally, a YOLO-v3 detector is added to facilitate object detection and train the network, using the YOLO detector's loss function.

3.2.1. Using Generative Adversarial Network Pipeline

In computer vision and pattern recognition, generative adversarial networks (GANs) have been used for their generic nature, from simple tasks of binary classification to learning image super-resolution, using a generator along with the discriminator that rates the output of the generator. For super-resolution, the first GAN-based proposed was SR-GAN [51], using the EDSR-based residual blocks in the generator network; a similar approach was used in [39], Jiang et al., which proposed Edge-enhanced super-resolution GAN for remote sensing applications. We modify the generator network in the SR-GAN and incorporate residual feature aggregation (See Figure 6) to enhance the generator's performance. The objective of the generator network is to generate SR images, which are then evaluated by the discriminator D to rate the output of the generator as real (HR image) or fake. Feedback from the discriminator is used by the generator to fine-tune the network parameters, using the L1 loss function, to generate more realistic images while at the same time, the discriminator network also improves itself in each iteration.

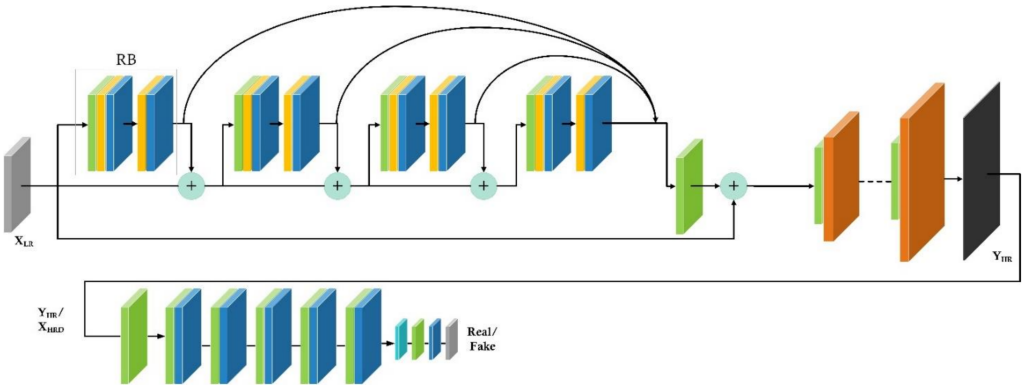


Figure 6. SR-GAN with an EDSR-RFA generator at the top with the discriminator at the bottom. The layers include convolution (RFA block) (**green**), normalization (RFA block) (**yellow**), ReLU activation (RFA Block) (**blue**), a 1 × 1 reduction layer (**light blue**), and a pixel rearrangement layer (**brown**). Output Y_{HR} of the generator is the input X_{HRD} to the discriminator.

The first improvement is incorporated, using a generator network *Gen*, as shown in Figure 6, using RFA-based 32 residual blocks of the size 128 × 128 while the discriminator

network D is the same as in SR-GAN for evaluating the generated HR images. We also utilized the improvements proposed by Arjovsky et al. [50] in the form of Wasserstein GANs, which use gradient penalty in the loss function of the discriminator that is given by the following:

$$\mathcal{L}(SR-CGAN) = \sum_{I_{HR}\sim\mathbb{P}_t}[D(I_{HR})] - \sum_{I_{SR}\sim\mathbb{P}_g}[D(I_{SR})] + \lambda \sum_{I_r\sim\mathbb{P}_r}\left[(\|\Delta D(I_r)\|_2 - 1)^2\right], \quad (6)$$

where I_{HR}, $I_{SR,}$ and I_r respectively are the HR image, SR image, and a random image which is uniformly sampled from I_{HR} and I_{SR} while P_t, $P_{g,}$ and P_r are probability density distributions of true HR images, generated images, and randomly sampled images respectively while λ is the gradient penalty coefficient, which is assigned a high value of 10 as recommended in [52].

3.2.2. Cyclic Network

The next improvement involves using a cyclic approach in GANs as proposed by Zhu et al. [53] by adding another GAN to generate LR images from the output of the first generator as shown in Figure 7.

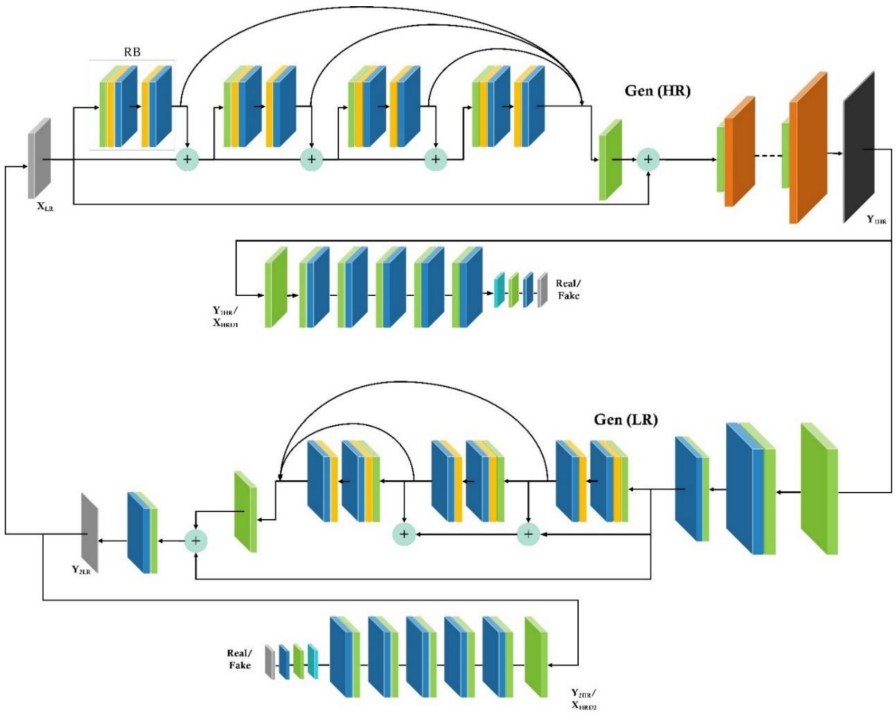


Figure 7. The cyclic approach: GAN-EDSR-RFA. Gen (HR) and its discriminator is shown on top, and Gen (LR) for cyclic feedback is shown in the lower half of the figure.

The final LR image as generated by the second generator is compared with the original LR image. Using the loss function, it adjusts the network parameters as follows:

$$\begin{aligned}\mathcal{L}(SRCGAN) = \mathcal{L}_{L1}(Gen_{HR}(I_{LR}), I_{HR}) + \mathcal{L}_{MSE}(Gen_{HR}(Gen_{LR}(I_{HR})), I_{HR}) + \\ \mathcal{L}_{L1}(Gen_{LR}(I_{HR}), I_{LR}) + \mathcal{L}_{MSE}(Gen_{LR}(Gen_{HR}(I_{LR})), I_{LR})\end{aligned} \quad (7)$$

where I_{HR} and I_{LR} are HR, and LR images; Gen_{HR} and Gen_{LR} are the corresponding HR and LR generators as shown in Figure 7.

As shown in Equation (7), the loss function of cyclic GAN shows our significant improvement, which ensures that the two GAN networks operate to minimize the overall loss function by evaluating one another's output. The final reconstructed LR image (I_{RLR}) generated by Gen_{LR} is compared with the actual LR image (I_{LR}), generated using the

true HR image, using bicubic and blur degradations. The key idea is to ensure the two low-resolution images (i.e., I_{RLR} and I_{LR}) are identical, thereby ensuring that the generated HR image is also close to the true HR image. Further results and analysis of this choice are shown in Section 4.

3.2.3. Detection Network

We further add the state-of-the-art object detection network YOLOv3 to perform the final detection task on the SR images generated by the cyclic GAN. The primary objective is to ensure that the object of interest in the generated HR image is like the true HR images, thus using YOLOv3 to ensure this in the overall model. The YOLOv3 detection network is trained on the HR images. Thereby, its network weights are fixed while integrating it in the cyclic GAN model; the resulting model is shown in Figure 8. In the training phase, the generated HR image is fed into the YOLO network to detect objects while the coordinates of the prediction boxes are used to compute the loss function of the YOLO network, i.e., $\mathcal{L}(Y)$. The total loss function is calculated, and using backpropagation, the weights of generator and discriminator networks are updated, using the gradient of total loss.

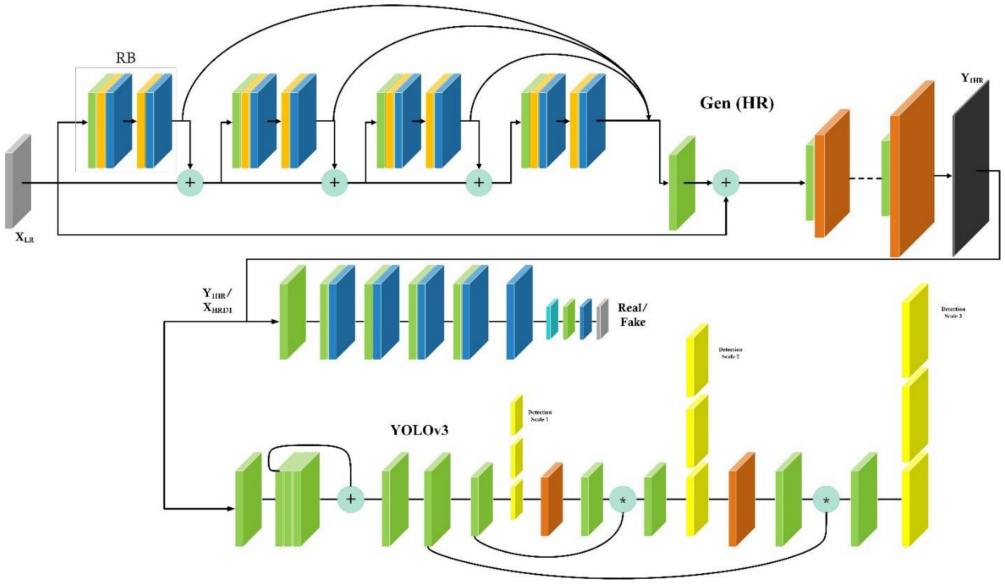


Figure 8. The network architecture for super-resolution SRCGAN with RFA and YOLOv3 detector (SRCGAN-RFA-YOLO).

As seen in Figure 8, there are three components in the proposed method; thus, three different loss functions are used, as shown in Equations (8)–(10):

$$\mathcal{L}(G) = \mathcal{L}(pram) = \frac{1}{N}\sum_{i=1}^{N}\left\| G\left(I_{LR}^{i}\right) - I_{HR}^{i}\right\|_1, \tag{8}$$

where $\mathcal{L}(G)$ is the generator loss, *pram* are parameters, N is the total number of samples, $G\left(I_{LR}^{i}\right)$ is the $i-th$ generated HR image, and I_{HR}^{i} is the ground truth HR image.

A 48 ResBlock-based generative network $G(I_{LR}, \gamma)$ with a block size of 64 × 64 and kernel of 3 × 3 and a discriminator, which is based on the improved Wasserstein GAN [8], with a high value of gradient penalty coefficient in the discriminator loss function is shown in Equation (9):

$$\mathcal{L}(D) = \sum_{I_{HR}\sim\mathbb{P}_t}[D(I_{HR})] - \sum_{I_{SR}\sim\mathbb{P}_g}[D(I_{SR})] + \lambda\sum_{I_r\sim\mathbb{P}_r}\left[\left(\|\Delta D(I_r)\|_2 - 1\right)^2\right], \tag{9}$$

where I_{HR}, $I_{SR,}$ and I_r, respectively, are the HR image, SR image, and a random image which is uniformly sampled from I_{HR} and I_{SR} while P_t, $P_{g,}$ and P_r are probability density distributions of true HR images, generated images, and randomly sampled images, respectively, while λ is the gradient penalty coefficient, which is assigned a high value of 10 as recommended in [52].

The detection network (YOLO) loss function is given by the following:

$$\mathcal{L}(Y) = \sum_{i=0}^{DG^2} \sum_{j=0}^{A^2} \left[(x - x')^2 + (y - y')^2 + (h - h')^2 + (w - w')^2 \right], \tag{10}$$

where DG and A represent the detection grid and the number of anchor boxes associated with the YOLO detector while (x, y, h, w) and (x', y', h', w') depict the coordinates of the ground truth and predicted boundary boxes; (x, y) and (x', y') show the coordinate of the center point while (h, w) and (h', w') represent the dimensions of boundary boxes. The overall loss of the network shown in Figure 8 is as follows:

$$\mathcal{L}(T) = \mu_1 \mathcal{L}(G) + \mu_2 \mathcal{L}(D) + \mu_3 \mathcal{L}(Y)\,, \tag{11}$$

where μ_1, μ_2, and μ_3 are the weights associated with the generator, discriminator, and detector networks. During the training process, weights were assigned values of μ_1 = 0.95, μ_2 = 10, and μ_3 = 0.1 to ensure the range of the errors is similar for the three loss functions, which guaranteed balanced learning in all three aspects. Using super-resolution to achieve HR images, using the generator, realistic-looking images are ensured by the discriminator, and object detection and localization are performed by the detection network.

3.3. Implementation Details

The proposed method was implemented on the PyTorch framework with an Ubuntu 20.04 computer with 16 GB RAM and a Nvidia Titan XP graphics processor. A gradient penalty coefficient of 10 was selected for training Wasserstein GAN, while the total loss function weights, μ_1, μ_2, and $\mu_{3,}$ were assigned values of 0.95, 10, and 0.1, respectively. The network was trained for a total of 250 epochs.

4. Results

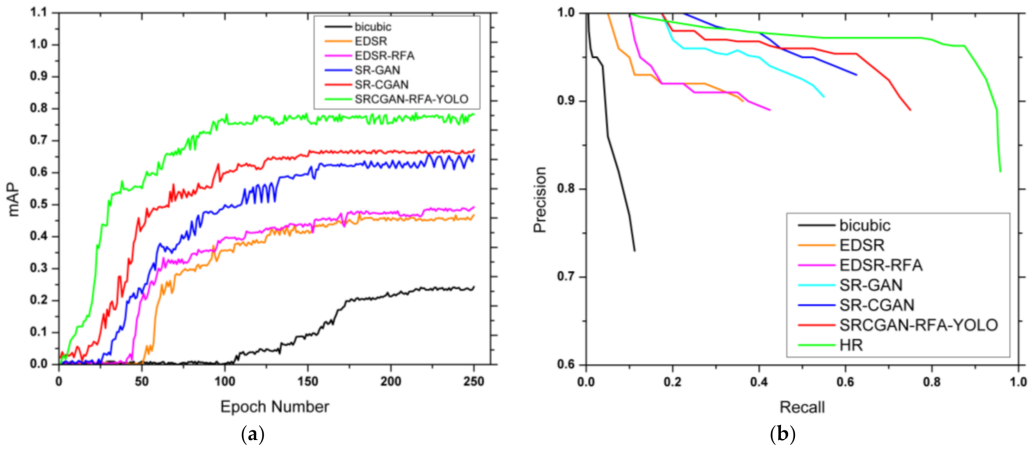
In this section, we discuss and evaluate the results achieved by the proposed novel SRCGAN-RFA-YOLO network. Initially, we focus on the main task of object detection and compare the results with the prevailing methods reported in this research. In Section 4.2, we compare the results of our approach with state-of-the-art object detectors in terms of AP for an IoU of 0.10. Lastly, we discuss the concept of transfer learning, using a pre-trained network to perform image super-resolution and object detection on other aerial datasets, such as Draper Satellite Image Chronology [41] and VAID [42].

4.1. Improvement in Image Quality and the Detection Accuracy

The SR results of various methods in terms of image quality metrics (PSNR and SSIM) are shown in Figure 9a. The LR version is for a scale factor of 16, which corresponds to a GSD of 80 cm/pixel. The image SR methods, EDSR and EDSR-RFA, when used alone, do not achieve high image quality as seen in Figure 9a, while the cyclic approaches, i.e., SRCGAN and SRCGAN-RFA-YOLO, generate good quality, super-resolved images with PSNR of 28.13 dB and 30.41 dB, respectively. The zoomed images in Figure 9b show the quality of recovered details for the various methods. Figure 9 shows that the cyclic models performed best by reconstructing the images with SSIM of 0.86 and 0.93 for SRCGAN and SRCGAN-RFA-YOLO, respectively. The proposed method of SRCGAN-RFA-YOLO achieved the best results especially for recovering texture information within the low-resolution image, as seen in Figure 9a, due to the improvements introduced in Section 3.2.

Figure 9. Comparison of the results for SR with a scale factor of 16 using EDSR-RFA, SR-CGAN, and SRGAN-RFA-YOLO. The LR version is 80 cm/pixel while the HR image is 5 cm/pixel. (**a**) SR images of different methods and their image quality metrics in terms of PSNR and SSIM; (**b**) zoomed sections for two different locations.

Finally, we compare training performance and precision/recall curves of the proposed SRCGAN-RFA-YOLO method compared with bicubic interpolation, EDSR, EDSR-RFA, SR-GAN, and SR-CGAN for 250 epochs. The training performance of our methods surpasses all the other reported methods in terms of AP (See Figure 10a); at the same time, our methods achieved higher AP within a few epochs, which shows that the proposed method converges fast and that the training is stable. Precision/recall curves also show the superior performance of our method in Figure 10b. The green line in Figure 10b represents the precision/recall curve for the HR image. The precision/recall curve of the proposed SRCGAN-RFA-YOLO method is shown in red. Compared to EDSR and EDSR-RFA, our method approached the performance achieved by the HR images, thus offering superior detection performance for low-resolution images.



(**a**) (**b**)

Figure 10. Performance evaluation of various methods used in this study based on AP and precision/recall curves with a YOLOv3 detector and an IoU of 0.10. (**a**) AP of the various methods during training; (**b**) precision versus recall curves.

The detection performance was measured in terms of precision, recall, and F1 score, as shown in Table 5.

Table 5. Detection performance on ISPRS Potsdam dataset for a scale factor of 16 and IoU of 0.10.

Method	Precision	Recall	F1 Score
HR	0.96	0.96	0.96
Bicubic	0.20	0.02	0.03
EDSR	0.24	0.02	0.03
EDSR-RFA	0.31	0.02	0.04
SR-GAN	0.86	0.48	0.61
SR-CGAN	0.90	0.56	0.69
SRCGAN-RFA-YOLO	**0.87**	**0.84**	**0.86**

As shown in Table 5, SRCGAN-RFA-YOLO reported an F1 score of 0.86 for an IoU of 0.10 for the ISPRS Potsdam dataset. Compared to the detection F1 score of EDSR-RFA, our method reported an increase in F1 score by a factor of 20.5.

4.2. Performance against Other Object Detectors

In this section, the performance of novel SRCGAN-RFA-YOLO is tested with other state-of-the-art object detectors by replacing the detector network in the method with RetinaNet [15], SSD (VGG16) [14], EfficientDet [54], and Faster R-CNN [10]. To avoid method bias, we trained all the methods using the HR images, using an input image size of 512 × 512 pixels on the ISPRS Potsdam datasets. The detection performance of SRCGAN-RFA-YOLO is compared with state-of-the-art object detectors for scale factors of 8 and 16 as shown in Table 6. For HR images, the detection performance is almost similar for all object detectors, but as the scale factor increases, the proposed achieves the highest AP values, especially for a scale factor of 16 where the GSD is 80 cm/pixel. The single image inference time shows that the proposed method, while providing the best AP for higher scaling factors, has a relatively low inference time, i.e., 33.67 ms per image.

The super-resolved images generated by the proposed SRCGAN-RFA-YOLO were used as inputs to the object detectors mentioned above, and the detection performance in terms of AP for an IoU of 0.10 is shown in Table 7.

Table 6. The impact of scale factor on the detection AP with an IoU of 0.10.

Method	HR (AP)	SF = 8 (AP)	SF = 16 (AP)	Time (ms)
YOLOv3	96.83	0.713	24.34	**27.73**
RetinaNet	92.41	0.687	14.78	91.64
SSD (VGG16)	**97.05**	0.761	41.31	32.45
EfficientDet	96.97	0.754	42.23	51.24
Faster R-CNN	96.93	0.727	39.72	94.19
SRCGAN-RFA-YOLO	96.57	**0.897**	**78.67**	33.67

SF is the scale factor. The best results are shown in bold font while the second-best results are underlined.

Table 7. The impact of changing the detection networks on AP with an IoU of 0.10 (for a scale factor of 16).

Method	HR	Bicubic	EDSR-RFA	SRCGAN-RFA-YOLO
YOLOv3	96.83	24.34	49.31	78.67
RetinaNet	92.41	14.78	29.84	64.72
SSD (VGG16)	**97.05**	41.31	69.98	87.54
EfficientDet	96.97	**42.23**	69.17	88.56
Faster R-CNN	96.93	39.72	**70.26**	**88.81**

The best results are shown in bold font while the second-best results are underlined.

As shown in Table 7, YOLOv3, SSD (VGG16), EfficientDet, and Faster R-CNN achieved similar detection performance on HR, which shows that the HR image performance is not dependent upon the choice of detector while for SR images (for a scale factor of 16), the best results were achieved by Faster R-CNN with AP of 88.81%. Compared to the bicubic interpolation, the most significant change in AP (~54%) was achieved by the YOLOv3 detector because SSD (VGG16), EfficientDet, and Faster R-CNN achieved comparatively high AP values for bicubic interpolated images.

Higher AP values were achieved by SSD (VGG16) and EfficientDet on bicubic interpolated images mainly because of the built-in image augmentation scheme within the training process. The input images and the boundary boxes are resampled for different scale values and aspect ratios [14,54]. Similarly, Faster R-CNN also achieved improved AP because of the introduction of the region proposal network (RPM), which uses attention mechanisms to improve detection accuracy. The proposed framework allows integrating different object detectors; thus, using object detectors such as SSD (VGG16), EfficientDet, and Faster R-CNN will improve the overall AP by 10% as shown in Table 7.

4.3. Applying Transfer Learning on Different Remote Sensing Datasets

This section discusses the concept of transfer learning by performing object detection on different remote sensing datasets, using the parameters learned by the novel SRCGAN-RGA-YOLO on the ISPRS Potsdam dataset. The two datasets considered for this task are Draper Satellite Image Chronology [41] and VAID [42]. Figure 11 shows the results of image super-resolution on the datasets. To compare the performance, we downscaled the images to match the GSD of 80 cm/pixel: the VAID dataset has a GSD of 12.5 cm/pixel, thus, a scale factor of 6 was used to generate LR images; for Draper Satellite Image Chronology, the GSD was 10 cm/pixel, thus, a scale factor of 8 was used to generate LR images.

The results in Figure 11 depict the significance of the proposed method to perform small object detection on low-resolution satellite images, using a pre-trained network in an end-to-end manner.

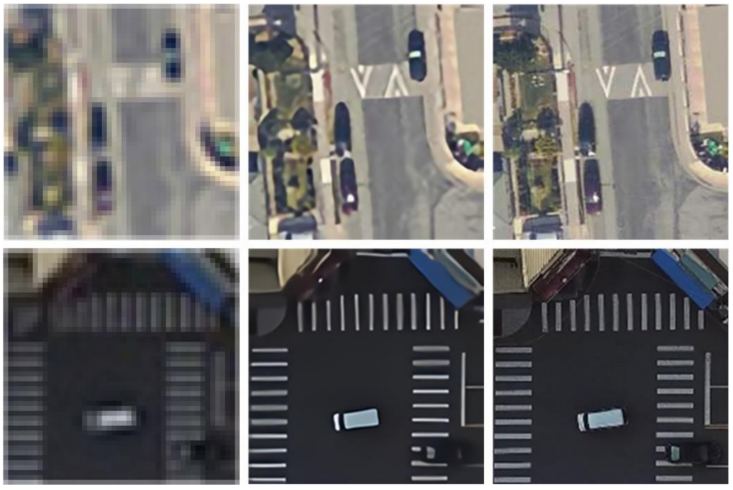


Figure 11. Object reconstruction using SRCGAN-RFA-YOLO having a scale factor of 8 from Draper Satellite Image Chronology (**top**) and VAID dataset (**bottom**) using the parameters learned from the Potsdam dataset.

5. Discussion

Analyzing other image tiles from the same dataset, we measured the detection performance for various methods, including SRCGAN-RFA-YOLO. We further investigated the impact of IoU on the detection performance, and Table 8 shows the variation in AP for three different values of IoU, i.e., 0.10, 0.25, 0.50. The best results are highlighted in bold while the second-best results are underlined.

Table 8. Comparison of AP corresponding to different IoU for methods explored in this study (Bold = best results; Underline-second-best results).

Method	IoU = 0.10	IoU = 0.25	IoU = 0.50
HR	96.83	95.71	85.64
Bicubic	24.34	19.52	12.24
EDSR	46.81	42.68	31.12
EDSR-RFA	49.31	45.13	36.84
SR-GAN	65.53	61.18	<u>48.82</u>
SR-CGAN	<u>67.19</u>	<u>63.97</u>	48.63
SRCGAN-RFA-YOLO	**78.67**	**72.28**	**58.76**

It is worth noticing that SRCGAN-RFA-YOLO achieved the best AP for all three IoU values, SRCGAN achieved the second-best AP for an IoU of 0.10 and 0.25, while SR-GAN achieved second best AP for an IoU of 0.50. For small object detection, a low threshold value of 0.10 is selected as the IoU, and the proposed method achieved an AP value of 78.67% that is approximately 32% more than the AP achieved by the EDSR method when used with the YOLO detector. Furthermore, using the YOLO network's loss function to train our SRCGAN-RFA increased the AP by 11.5%, highlighting the superior performance of the proposed method.

Compared to bicubic interpolation-based detection, the proposed method achieved an increase of 54.3% AP for an IoU of 0.10 while, compared to the EDSR approach, AP value increased by 31.8%. These increases in AP are very significant because the spatial resolution of the LR image is 80 cm/pixel at a scale factor of 16, and for such low-resolution images, the proposed method achieved state-of-the-art performance.

In Figure 12, for the LR image, the YOLO network did not detect any vehicles, while residual feature aggregation-based methods successfully detected all five vehicles while the

SRCGAN-RFA-YOLO achieved the best performance in terms of image quality confidence score and AP.

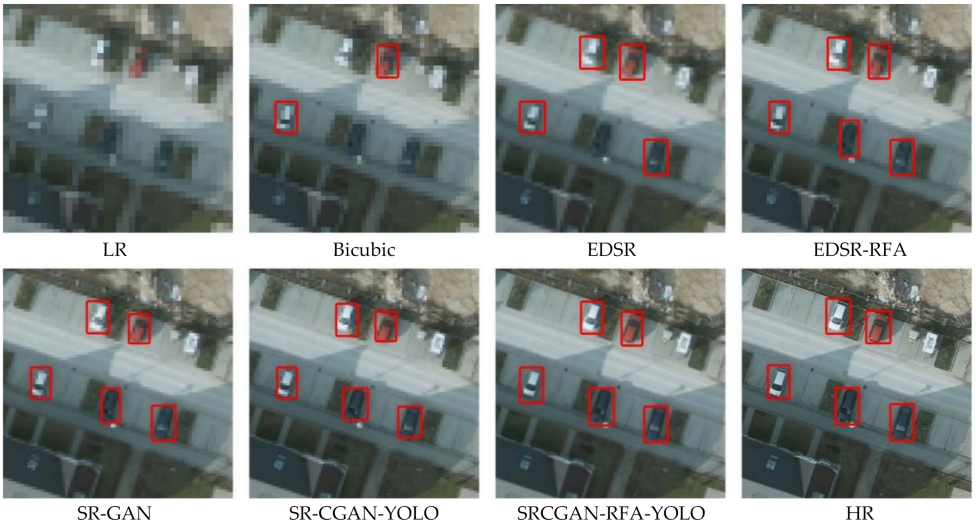


Figure 12. Detection examples using YOLOv3 as detector network.

The proposed framework allows integrating different object detectors; thus, using object detectors like SSD (VGG16), EfficientDet, and Faster R-CNN will improve the overall AP by 10% as shown in Table 7.

The results of transfer learning on both datasets are shown in Figure 11. Although our method improved the images, especially the vehicles within the LR images, the complex textures were not generated by our approach because of the lack of features; however, the super-resolved images achieved improved object detection.

The novelty of the method is threefold: first, a framework is proposed, independent of the final object detector used in the research, i.e., YOLOv3 could be replaced with Faster R-CNN or any object detector to perform object detection; second, a residual feature aggregation network was used in the generator, which significantly improved the detection performance as the RFA network detected complex features; and third, the whole network was transformed into a cyclic GAN, which improved the training, test, and validation accuracy, also evident from Figure 10a.

The detection performance of the proposed method was measured on an independent dataset created from selected images from the Draper Satellite Image Chronology and VAID dataset (images downscaled to match a GSD of 80 cm/pixel, i.e., low-resolution). The detection performance was measured in terms of precision, recall, and F1 score as shown in Table 9.

Table 9. Detection performance on an independent dataset with a GSD of 80 cm/pixel (IoU of 0.10). (Bold = best results).

Method	Precision	Recall	F1 Score
HR	0.96	0.95	0.96
Bicubic	0.27	0.05	0.09
EDSR	0.24	0.07	0.11
EDSR-RFA	0.25	0.10	0.14
SR-GAN	0.70	0.44	0.54
SR-CGAN	0.82	0.59	0.69
SRCGAN-RFA-YOLO	**0.88**	**0.80**	**0.84**

As shown in Table 9, SRCGAN-RFA-YOLO reported an F1 score of 0.84 for an IoU of 0.10 for an independent dataset, using the network parameters learned from the LR–HR

image pairs from the ISPRS Potsdam dataset. Figure 13 shows the examples of detection by SRCGAN-RFA-YOLO.

Figure 13. Detection on an independent dataset for an IoU of 0.10. **Green**—True Positive, **red**—False Positive, and **blue**—False Negative.

The performance on the independent dataset shows the effectiveness of the proposed method. As evident in Figure 13, misclassified examples include vehicles under shadow or vehicles embedded in a similar color background. However, using the pre-trained network, SRCGAN-RFA-YOLO achieved an F1 score of 0.84.

6. Conclusions and Future Directions

Object detection for low-resolution images is challenging, and generic detectors perform poorly when used directly on low-resolution images. In the case of satellite images, when dealing with low-resolution images, the object of interest (vehicles in our case) occupies few pixels, and we have shown that using super-resolution before the detection task significantly increases the detection accuracy. Our method targets the challenging condition where the spatial resolution of the remote sensing images is low, and in the case of most remote sensing satellite images, the image resolution is not very high due to the limitation of the imaging payload. The proposed improvements in the object detection network include using a residual feature aggregation-based super-resolution method (EDSR-RFA) to gather capture complex features and improve the model by employing a cyclic GAN-based (SRCGAN) approach and using an object detection network (YOLOv3) for the training of the network and detection tasks. The results show that compared to object detection on bicubic interpolated images, our methods improved the AP by 54.3%, and compared to the state-of-the-art EDSR-RFA method, the proposed SRCGAN-RFA-YOLO achieved a 29.4% increase in AP. Furthermore, transfer learning showed that our method increases the image quality significantly when applied to other datasets, using the parameters learned from the dataset used in this study.

Compared to other object detectors (YOLOv3 and SSD), the inference time of the proposed method is relatively high (33.67 ms) since the residual feature aggregation ensures the collection of complex features, thereby increasing the network parameters. The next step is to combine all the proposed improvements into a single framework and deploy it as a generic object detector for low-resolution images. Furthermore, we plan to train the network on big datasets, such as BigEarth [55], and release the pre-trained networks for object detection tasks in aerial and remote sensing images.

Author Contributions: S.M.A.B. proposed the concept and performed experiments. S.M.A.B. wrote the manuscript; Y.W. supervised, reviewed, and provided suggestions on the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62071384) and the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2019JM-311).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available at ISPRS Potsdam (https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/), VAID (https://vision.ee.ccu.edu.tw/aerialimage/) and Draper Satellite Image Chronology (https://www.kaggle.com/c/draper-satellite-image-chronology).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cao, C.; Wang, B.; Zhang, W.; Zeng, X.; Yan, X.; Feng, Z.; Liu, Y.; Wu, Z. An Improved Faster R-CNN for Small Object Detection. *IEEE Access* **2019**, *7*, 106838–106846. [CrossRef]
2. Stuparu, D.G.; Ciobanu, R.I.; Dobre, C. Vehicle detection in overhead satellite images using a one-stage object detection model. *Sensors* **2020**, *20*, 6485. [CrossRef]
3. Guan, L.; Wu, Y.; Zhao, J. SCAN: Semantic context aware network for accurate small object detection. *Int. J. Comput. Intell. Syst.* **2018**, *11*, 951–961. [CrossRef]
4. Yao, Q.; Hu, X.; Lei, H. Multiscale Convolutional Neural Networks for Geospatial Object Detection in VHR Satellite Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 23–27. [CrossRef]
5. Alganci, U.; Soydas, M.; Sertel, E. Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images. *Remote Sens.* **2020**, *12*, 458. [CrossRef]
6. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
7. Robinson, Y.H.; Vimal, S.; Khari, M.; Hernández, F.C.L.; Crespo, R.G. Tree-based convolutional neural networks for object classification in segmented satellite images. *Int. J. High. Perform. Comput. Appl.* **2020**. [CrossRef]
8. Uehara, K.; Nosato, H.; Murakawa, M.; Nakamura, R.; Miyamoto, H.; Sakanashi, H. Multi-channel higher-order local autocorrelation for object detection on satellite images. *Int. J. Remote Sens.* **2020**, *41*, 752–771. [CrossRef]
9. Zhang, W.; Wang, S.; Thachan, S.; Chen, J.; Qian, Y. Deconv R-CNN for small object detection on remote sensing images. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 2483–2486.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 10–12 December 2015.
11. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
12. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804:02767.
13. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004:10934.
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
15. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
16. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H. IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]
17. Xu, Z.; Shi, H.; Li, N.; Xiang, C.; Zhou, H. Vehicle Detection under UAV Based on Optimal Dense YOLO Method. In Proceedings of the 2018 5th International Conference on Systems and Informatics, ICSAI 2018, Nanjing, China, 10–12 November 2018; pp. 407–411.
18. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Ma, Y.; Piao, C. Uav-yolo: Small object detection on unmanned aerial vehicle perspective. *Sensors* **2020**, *20*, 2238. [CrossRef] [PubMed]
19. Froidevaux, A.; Julier, A.; Lifschitz, A.; Pham, M.T.; Dambreville, R.; Lefèvre, S.; Lassalle, P.; Huynh, T.L. Vehicle Detection and Counting from VHR Satellite Images: Efforts and Open Issues. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Virtual Symposium, 26 September–2 October 2020; pp. 256–259.
20. Pham, M.T.; Courtrai, L.; Friguet, C.; Lefèvre, S.; Baussard, A. YOLO-fine: One-stage detector of small objects under various backgrounds in remote sensing images. *Remote Sens.* **2020**, *12*, 2501. [CrossRef]
21. Wang, J.; Lin, Y.; Guo, J.; Zhuang, L. SSS-YOLO: Towards more accurate detection for small ships in SAR image. *Remote Sens. Lett.* **2021**, *12*, 122–131. [CrossRef]
22. Courtrai, L.; Pham, M.T.; Lefèvre, S. Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks. *Remote Sens.* **2020**, *12*, 3152. [CrossRef]
23. Hui, Z.; Li, J.; Gao, X.; Wang, X. Progressive perception-oriented network for single image super-resolution. *Inf. Sci.* **2021**, 769–786. [CrossRef]

24. Bashir, S.M.A.; Wang, Y.; Khan, M. A Comprehensive Review of Deep Learning-based Single Image Super-resolution. *arXiv* **2021**, arXiv:2102.09351v1.
25. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In *Lecture Notes in Computer Science*; Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8692 LNCS.
26. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
27. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In *Lecture Notes in Computer Science*; Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9906 LNCS, pp. 391–407.
28. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
29. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
30. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
31. Sajjadi, M.S.M.; Scholkopf, B.; Hirsch, M. EnhanceNet: Single Image Super-Resolution through Automated Texture Synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 4491–4500.
32. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent back-projection network for video super-resolution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3897–3906.
33. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015.
34. Xue, S.; Qiu, W.; Liu, F.; Jin, X. Wavelet-based residual attention network for image super-resolution. *Neurocomputing* **2020**, *382*, 116–126. [CrossRef]
35. Zhang, L.; Ma, J. Salient Object Detection Based on Progressively Supervised Learning for Remote Sensing Images. *IEEE Trans. Geosci. Remote. Sens.* **2021**. [CrossRef]
36. Ferdous, S.N.; Mostofa, M.; Nasrabadi, N. Super resolution-assisted deep aerial vehicle detection. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Baltimore, MD, USA, 15–17 April 2019; p. 1100617.
37. Rabbi, J.; Ray, N.; Schubert, M.; Chowdhury, S.; Chao, D. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. *Remote Sens.* **2020**, *12*, 1432. [CrossRef]
38. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Computer Vision—ECCV 2018 Workshops. ECCV 2018*; Leal-Taixé, L., Roth, S., Eds.; Springer: Cham, Switzerland, 2019; pp. 63–79. ISBN 978-3-030-11021-5.
39. Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Lu, T.; Jiang, J. Edge-Enhanced GAN for Remote Sensing Image Superresolution. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5799–5812. [CrossRef]
40. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *I-3*, 293–298. [CrossRef]
41. Draper Draper Satellite Image Chronology. Available online: https://www.kaggle.com/c/draper-satellite-image-chronology/overview (accessed on 8 May 2021).
42. Li, C.Y.; Lin, H.Y. Vehicle detection and classification in aerial images using convolutional neural networks. In Proceedings of the VISIGRAPP 2020—15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Valetta, Malta, 27–29 February 2020; pp. 775–782.
43. Shermeyer, J.; Van Etten, A. The effects of super-resolution on object detection performance in satellite imagery. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
44. Wei, Z.; Liu, Y. Deep Intelligent Neural Network for Medical Geographic Small-target Intelligent Satellite Image Super-resolution. *J. Imaging Sci. Technol.* **2021**. [CrossRef]
45. Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
46. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
47. Freeman, W.T.; Jones, T.R.; Pasztor, E.C. Example-based super-resolution. *IEEE Comput. Graph. Appl.* **2002**, *22*, 56–65. [CrossRef]
48. Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; Wu, G. Residual Feature Aggregation Network for Image Super-Resolution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual Conference, 14–19 June 2020.

49. Audebert, N.; Le Saux, B.; Lefèvre, S. Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sens.* **2017**, *9*, 368. [CrossRef]
50. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, Australia, 6–11 August 2017.
51. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
52. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein GANs. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5767–5777.
53. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2223–2232.
54. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual Conference, 14–19 June 2020; pp. 10781–10790.
55. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 5901–5904.