Super resolution-assisted deep aerial vehicle detection

Conference Paper · May 2019

DOI: 10.1117/12.2519045

3 authors:

Syeda Nyma Ferdous
West Virginia University
15 PUBLICATIONS **75** CITATIONS
SEE PROFILE

Moktari Mostofa
West Virginia University
16 PUBLICATIONS **129** CITATIONS
SEE PROFILE

N.M. Nasrabadi
West Virginia University
576 PUBLICATIONS **15,500** CITATIONS
SEE PROFILE

Super Resolution-Assisted Deep Aerial Vehicle Detection

Syeda Nyma Ferdous*, Moktari Mostofa* and Nasser M. Nasrabadi*

*Lane Department of Computer Science and Electrical Engineering, West Virginia University;

ABSTRACT

Vehicle detection in aerial imagery has become tremendously a challenging task due to the low resolution characteristics of the aerial images. Super-Resolution; a technique which recovers high-resolution image from a single low-resolution image can be an effective approach to resolve this shortcoming. Hence, our prime focus is to design a framework for detecting vehicles in super resolved aerial images. Our proposed system can be represented as a combination of two sub-systems. The first sub-system aims to use a Generative Adversarial Network (GAN) for getting super resolved images. A GAN consists of two networks: a generator network and a discriminator network. It ensures recovery of photo-realistic images from down-sampled images. The second sub-system consists of a deep neural network (DNN)-based object detector for detecting vehicles in super resolved images. In our architecture, the Single Shot Multi Box Detector (SSD) is used for vehicle detection. The SSD generates fixed-size bounding boxes with predicting scores for different object class instances in those boxes. It also employs a non-maximum suppression step to produce final detections. In our algorithm, our deep SSD detector is trained with the predicted super resolved images and its performance is then compared with an SSD detector that is trained only on the low-resolution images. Finally, we compare the performance of our proposed pre-trained SSD detector on super-resolved images with an SSD that is trained only on the original high resolution images.

Keywords: Super Resolution (SR), SRGAN (SR with Generative Adversarial Network), Single Shot Detector (SSD), Aerial vehicle detection

1. INTRODUCTION

Real-time vehicle detection in aerial imagery is extremely difficult. Predicting small vehicles from these large images are even difficult for human eyes. The reason lies in the nature of data along with computational constraints, low resolution and complex background of these imagery. Recently research community has paid substantial attention in this area as the outcome of this research can help better plan in transportation system, surveillance and reconnaissance.

As these images are taken from top view with varying altitude, the target objects can not contain much information. Also, target objects in these images are very small relative to the entire image that makes it hard to design a detector which distinguishes target from background. Using a deep learning technique; super-resolution, we can accelerate the detection performance.

An extensive study has been conducted for object detection in aerial imagery; however it is still an open problem demanding a high performance solution. Recent approaches having promising performances use convolutional neural network (CNN) based solution to detect objects of interest from aerial imagery. In this work, we propose a new architecture utilizing deep neural network (DNN) that helps to locate vehicles in satellite images ensuring effective performance. Our designed architecture can be divided into two stages. The first stage takes an input image and creates super resolved version of the original image using Generative Adversarial Network (GAN).[1] This network helps to augment image features improving image quality that leads to better object recognition. Also, it ensures to preserve high level features while transferring from low to high resolution domain.The second stage handles target prediction from these super-resolved images using single shot object detection. We further investigate the detection performance on multi-resolution images with their super resolved counterpart.

S.N.F.: E-mail: sf0070@mix.wvu.edu,
M.M.: E-mail: mm0251@mix.wvu.edu,
N.M.N. : E-mail: nasser.nasrabadi@mail.wvu.edu

Our proposed method demonstrates promising performance while tracking vehicles in aerial images. This paper is organized as follows. Section II outlines the review of the previous works related to ours. In Section III, we discuss our proposed framework. Experiment details is and the result is demonstrated in Section IV. Finally, in Section V, we reach the conclusion.

2. RELATED WORK

2.1 Single Image Super Resolution

Here, we will discuss the algorithms related to single image super-resolution (SISR). The algorithms can be categorized into four groups -prediction based models, edge based methods, image statistical models and example-based methods. Among them example-based methods[2–13] are known as more powerful methods which aim at mapping between low and high resolution images. They rely on external datasets. Early method was proposed by Freeman et al.[5,14] In the work of Glasner et al, they use patch redundancy across the scales within the image to model the image super-resolution problem. In Huang et al.,[15] self-similarity based super-resolution (SR) algorithms are expanded by incorporating small transformations and geometric variations. A convolutional sparse coding approach was proposed by Gu et al.[16] In their work, they ignore consistency of the pixels in overlapped patches by working on whole image that helps generate more robust reconstruction of image local structures.

Tai et al.[17] combine the benefits of edge-directed SR with learning-based SR to reconstruct more realistic texture details in super-resolved images. To recover high quality SR, Zhang et al.[18] introduce a multi-scale dictionary method that simultaneously incorporates local and non-local priors. The local prior removes artifacts from target image and the non-local prior helps produce more perceptually satisfying image.

The use of sparse learned dictionaries in combination with neighbor embedding methods obtain improved quality and speed in the work of Timofte et al.[7,19] The authors propose anchored neighborhood regression. Kernel ridge regression (KRR) is adopted to learn a map from input LR images to target SR images in Kim and Kwon[20] works. This idea is based on example pairs of input and output images.In Dai et al.[4] a joint learning of patch-specific regressors is proposed during training. At testing phase, it selects the best regressor which yields the smallest super-resolving error.

Convolutional neural network (CNN) based SR algorithms have attained superior performance. Wang et al.[21] propose a sparse representation in combination with feed forward network architecture . The underlying idea is based on the learned iterative shrinkage and thresholding algorithm (LISTA).[22,23] Dong et al.[24,25] trained a three layer deep convolutional neural network that achieved state of the art SR performance.

To recover perceptually more convincing HR images, Johnson et al.[26] and Bruna et al.[27] use a loss function closer to perceptual similarity. For our paper, we particularly follow the works of Christian Ledig et al.,[28] the authors use a perceptual loss function with MSE loss to generate more realistic SR images.

2.2 Vehicle Detection in Aerial Imagery

Vehicle detection in aerial imagery has been studied a lot in literature. Apart from convolutional neural network, prior works employed other approaches to address this problem. A model proposed in[29] utilized Bayesian Network with handcrafted parameters to identify vehicles in aerial image. To find cars in satellite image, The method presented in[30]applied Mean-shift algorithm utilizing shape information of the targets. The framework discussed in[31,32,32] trains a Dynamic Bayesian Network (DBN) with features preserving region level information.

With the progressive success of deep learning in object detection, many recent works use CNN for vehicle detection in aerial imagery. A fast detector proposed by Carlet and Abayowa[33] modifies YOLOv2 for locating vehicles in aerial imagery.Terrail et al.[34] applies modified faster R-CNN algorithm that achieves a breakthrough performance on aerial vehicle detection. In,[35] Soleimani et al. propose a text-guided detection scheme that utilizes both the visual and textual features for detection. Yang et al.[36]present a framework that uses skip connection to merge lower and higher level features utilizing focal loss function.Li et al.[37–39] propose a framework for multi-oriented vehicle detection. In this framework, a rotatable region proposal network is utilized that learns the orientation of vehicles while performing classification on aerial images and videos.

Though vehicle detection in aerial imagery has been a research focus in recent years, comprehensive study of detection performance on super resolved images has not yet been investigated in most of them. The work presented in[40]gives an overview of detection performance on super resolved images considering multiple resolutions. In this paper, a performance gain is reported in most resolutions for applying super resolution technique in original images. Liujuan Cao et al.[41]proposed a framework that employs super resolution with coupled dictionary learning on the satellite imagery and then a detection algorithm is applied on the generated images.

3. PROPOSED MODEL

In this section, we introduce our method in a detailed way. Firstly, we super resolve image from low resolution input image and then feed it to a detector. In this work, we have proposed a two-stage framework that employs SRGAN followed by single shot detector for target detection in aerial imagery.

3.1 Super-Resolution using GAN

Super-resolution aims at recovering a high resolution super resolved image from a low resolution input image. In,[5] the authors propose that there can be two ways to super resolve low resolution input image: single image based SR and multiple images based SR. In multiple-image super-resolution algorithms,[42] a couple of low-resolution images of the same scene is used as input and then a registration algorithm is employed to find the transformation between them. These algorithms can recover higher resolution details, however, their performances are limited by improvement factors close to two.[43] Single-image super-resolution algorithms, like,[24] usually have a single input. They train a set of low-resolution images along with their high-resolution counterparts to learn a relationship between them. The underlying idea is to use this learned relationship to predict the missing high-resolution details of the input low-resolution images. This idea helps generate high-resolution images far better than their low-resolution inputs.[43]

Recently, Super-resolution using the concept of GAN[1] has achieved the state-of-the art results.For our work, we use the super-resolution method of,[28] which is based on the powerful framework GAN. The algorithm encourages to generate photo-realistic images with high perceptual quality. Using this concept, our goal is to generate high resolution, super resolved image X^{SR} from its low resolution X^{LR} input; however X^{LR} is the low resolution counterpart of its original high resolution X^{HR}. The network use high resolution images during training. Here, we train a feed-forward CNN which generates a function G, parametrized by θ_G. θ_G consists of weights and biases of a deep neural network that is optimized using generative adversarial loss. θ_G is obtained by the following equation:

$$\theta_G = arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(X_n^{LR}), X_n^{HR}), \tag{1}$$

where X_n^{LR} with corresponding X_n^{HR} are training images for n = 1.....N. l^{SR} is perceptual loss designed as weighted combination of several loss functions which is to be minimized to recover the desired characteristics from reconstructed super-resolved image.

3.1.1 Generative Adversarial Network Architecture(GAN)

We train and optimize both discriminator network D_{θ_D}, along with G_{θ_G} to solve the adversarial min-max problem:[1]

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{X^{HR} \sim p_{train}(X^{HR})} \left[\log D_{\theta_D}\left(X^{HR}\right) \right] + \mathbb{E}_{X^{LR} \sim p_G(X^{LR})} \left[\log \left(1 - D_{\theta_D}\left(G_{\theta_G}\left(X^{LR}\right)\right)\right) \right]. \tag{2}$$

The main idea is to train a generative model G so that it can learn to create solutions that are highly similar to real images. Along with this, the discriminator is trained to distinguish the super-resolved image from the real one. So both network is optimized in an alternating manner to find the super-resolved image that looks like original high resolution image.

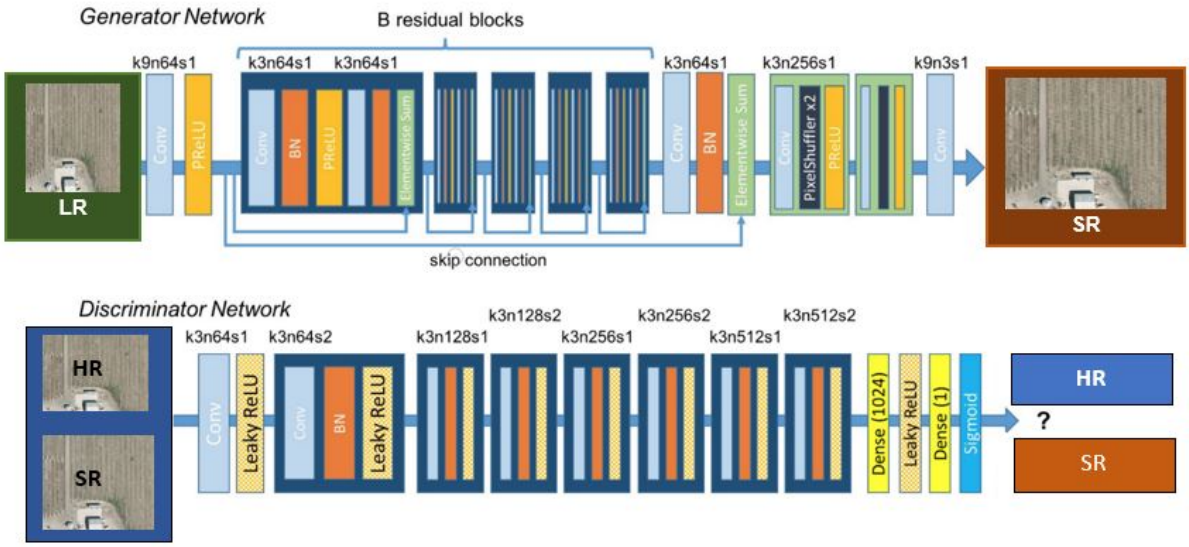


Figure 1: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps(n) and stride (s) indicated for each convolutional layer.

Following Christian Ledig et al.,[28] we use B residual blocks with identical layout in our generator network shown in Figure.1. Each block contains two convolutional layers with 3x3 kernels and 64 feature maps followed by batch-normalization layers[44] and Parametric ReLU[45] as the activation function. Here, the idea of sub-pixel convolution layers is used to increase the resolution of the input image as proposed by Shi et al.[46]

To train discriminator, we use eight convolutional layers with an increasing number of 33 filter kernels increased by a factor of 2 from 64 to 512 kernels as in the VGG network. Here, we have followed the architecture summmarized by Radford et al.[47] We also add two dense layers and a sigmoid activation function at the end of the resulting 512 feature maps to obtain a probability for classification.

3.1.2 Loss functions

In,[28] the authors design a loss function that is assessed perceptually. The loss is calculated as the weighted sum of a content loss ($l^{SR}_{content}$) and an adversarial loss(l^{SR}_{Gen}) component as:

$$l^{SR} = l^{SR}_{cont} + 10^{-3} l^{SR}_{Gen}. \tag{3}$$

They design the content loss using a pre-trained 19 layer VGG network[48] and define VGG loss as the euclidean distance between the feature representations of a reconstructed image and the reference high resolution image that is closer to perceptual similarity.

$$l^{SR}_{VGG/i.j} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}\left(X^{HR}\right)_{x,y} - \phi_{i,j}\left(G_{\theta_G}\left(X^{HR}\right)\right)_{x,y}\right)^2. \tag{4}$$

Here, $\phi_{i,j}$ indicates feature map at j-th convolutional layer followed by i-th maxpooling layer. $W_{i,j}$ and $H_{i,j}$ represent the dimensions of the respective feature maps within the VGG architecture.

In addition, they also add GAN loss which encourages the network to obtain natural looking images. It is defined based on the probability of discriminator, $D_{\theta_D}\left(G_{\theta_G}\left(X^{LR}\right)\right.$ that the reconstructed super-resolved image $G_{\theta_G}\left(X^{LR}\right)$ is an original high resolution image.

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}\left(G_{\theta_G}\left(X^{LR}\right)\right). \tag{5}$$

3.1.3 Object Detection using SSD

Object detection is a computer vision problem which aims to determine the presence of objects in image or video mimicking human-brain. This also identifies the location as well as the type of object. An extensive research has been conducted through decades to solve this problem. Before the emergence of deep neural network, object detection was performed using feature extractors like SIFT,[49] HOG[50] and classifiers such as SVM,[51] AdaBoost,[52] DPM.[53] A notable increase in performance was observed using deep learning based methods because of the robust learning capability of these nets. The state of the art deep learning based methods can be broadly classified into two categories. The first one is region-proposal based and the second one is end-to-end detection based model. RCNN,[54] Fast-RCNN,[55] Faster-RCNN[56]are the pioneer models with high performance of the first group while YOLO,[57] SSD[58] can be considered as significant ones for the second one. The problem with region proposal based approaches is that the detection is performed combining multiple stages making the system incompatible for real-time applications. End-to-end detection based methods alleviate this phenomena by eliminating region proposal stage with a unified architecture for detection and classification.

In our approach, we use SSD (Single Shot Detector) that locates vehicles in an image in real time. Like other object detection algorithms, SSD extracts features from image for generating bounding box with a class label. In contrast to classical region based detectors, it does not need region proposals to perform detection making the process extremely fast.

SSD needs input image and corresponding annotations of objects for detection. The architecture uses the features extracted by VGG-16[59] with six additional layers for object detection. It predicts objects from multiple feature maps of different resolutions organized in a decreasing fashion that ensures detection of variable size objects. For each position in feature map, SSD generates default boxes of varying sizes using multiple scales and aspect ratios with computed confidence score that designates the presence of objects in those boxes. Later, these boxes are matched with the ground truth annotated boxes. Boxes that have high overlap with the ground truth boxes provided by the annotations are considered as good match. A threshold factor named Intersection Over Union (IoU) is applied on the generated boxes. Afterwards, non-maximum suppression algorithm removes the duplicate bounding boxes for the same object.

By default, SSD considers generation of default boxes starting from convolutional layer 4_3. In our method, we have also considered default boxes from convolutional layer 3_3 as the size of our boxes are really tiny. While computing the loss, SSD uses both localization and confidence loss. The loss is calculated using Equation (2):

$$L_{Total} = \frac{1}{N}(L_{conf}(x, c) + L_{loc}(x, l, g)). \tag{6}$$

Here, Localization loss(L_{loc}) measures the distance between ground-truth and predicted box while confidence loss(L_{conf}) indicates the presence of an object in the generated bounding box.

Our proposed architecture is demonstrated in Figure 2. The generator network super resolves the original low resolution image. Discriminator is designed in such a way that it can distinguish between the real and fake images produced by the generator. These two networks benefit from each other generating more realistic images similar to the original image. Our goal is to train the detector with super resolved images and investigate the performance. Testing of our model is performed using Single Shot Detector.

Performance of the detector depends on the design of default boxes from different feature maps, training data and also on the parameters such as IoU, confidence threshold associated with the detection algorithm.

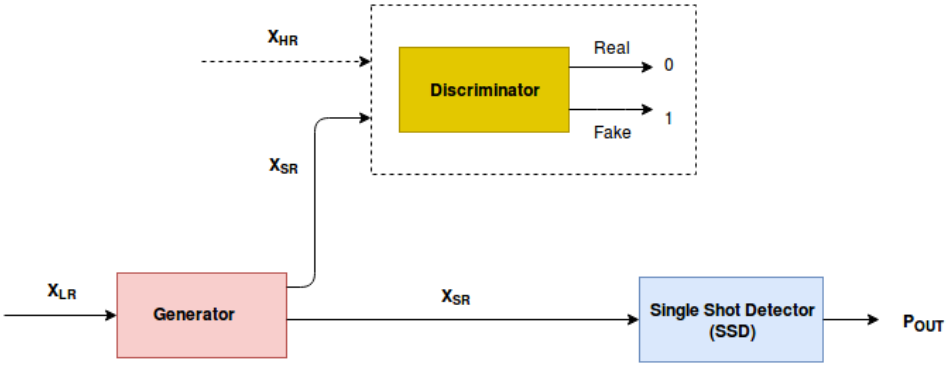


Figure 2: Architecture of super resolution with aerial imagery detection

4. EXPERIMENTS

4.1 Dataset and Implementation Details

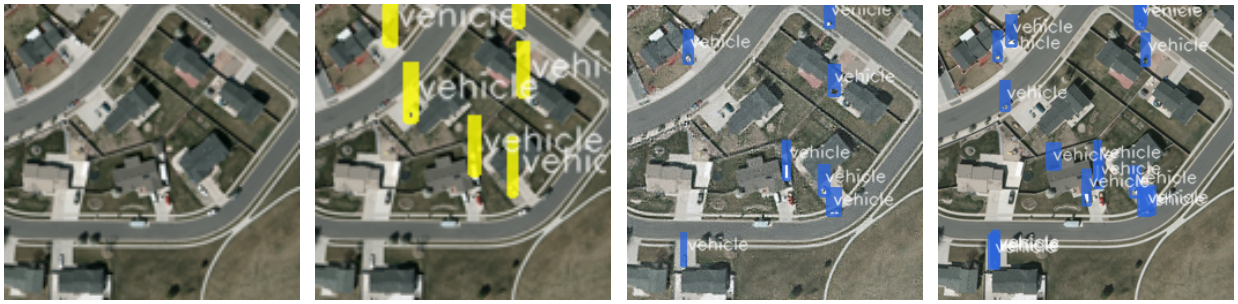
We perform our experiments on VEDAI (Vehicle Detection in Aerial Imagery)[60] aerial dataset. VEDAI images are taken from Utah database. These images contain multi-oriented vehicles with complex background which makes it to be considered as ideal dataset for any aerial image analysis task. The VEDAI dataset contains around 1200 images. For training and testing, we split the dataset into 1100 and 271 images respectively.

In our experiments, we use both the scale factor of 2x and 4x between low- and high-resolution images. For training, network requires low-resolution (LR) image as input. To obtain LR images,the HR images are downsampled using bicubic kernel with downsampling factor of 2 and 4. During implementation, we use the input image of size 128x128 and 256x256 to super-resolve to 256x256 and 512x512 respectively for the scale factor of 2. We further use the input resolution 128x128 to super-resolve to 512x512 for the upscaling factor of 4. To perceive our detection performance in more details, we have shown the detection result of the low, high and super resolved image in Figure 3. More positive objects are detected in Figure 3d, Figure 3h and Figure 3l than Figure 3c, Figure 3g and 3k indicating high resolution detection is more accurate than the corresponding super resolved detection. Figure 3b, Figure 3f and 3j demonstrate poor detection performance with many false detections in low resolution.

As mentioned in section 3.1.1, the generator of our network uses 16 residual blocks. Each of the blocks consists of two convolutional layer with the kernel size of 3x3, stride of 1 and 64 feature maps followed by batch-normalization and parametric-relu activation function. There is no max-pool layer throughout the network. To increase the resolution of the input image by 2 and 4 factor, the network adds one and two subpixel convolutional layer respectively.

Our dataset is not large. That's why we have used different augmentation techniques such as sharpening, flipping to make the model more robust to different image sizes.

We set the network hyperparameters as follows: For super-resolution, we adopt adam optimizer with a momentum of 0.9 and a learning rate of 10^{-4}. The model is trained for 2000 iterations with a batch size of 4. For the detection model, we train SSD architecture with initial learning rate of 0.001 with batch size 16. We optimize the network by Stochastic Gradient Descent (SGD) with a momentum of 0.02 and a weight decay of 0.9. We fine-tune our detection model by setting IoU 0.48, confidence threshold 0.05 and nms 0.45 to achieve the best performance. The entire network is implemented in tensorflow framework on two NVIDIA TITAN XP GPUs. We follow object detection and GAN based solution on Github for our network implementation.



(a) Input(128x128) (b) Detection(128x128) (c) Detection(SR 256x256) (d) Detection(HR 256x256)

(a) Input(256x256) (b) Detection(256x256) (c) Detection(SR 512x512) (d) Detection(HR 512x512)

(e) Input(128x128) (f) Detection(128x128) (g) Detection(SR 512x512) (h) Detection(HR 512x512)

Figure 3: Images illustrating the detection performance on super-resolved images of VEDAI dataset.

5. PERFORMANCE EVALUATION OF THE NETWORK

We have performed our experiments for both upscale factor of 2 and 4 and reported our models performance calculating the total Area Under the Curve(AUC),a metric which falls between 0 and 1, with a higher number indicating better classification performance. In addition, Mean Average Precision (mAP) is used as the metric of evaluation. we have tested our model on 128x128, 256x256 and 512x512 resolutions of the ground truth and their corresponding super-resolved images. The performance of our network for super resolution at upscale factor 2 and 4 is given in Table 1 and Table 2 respectively.

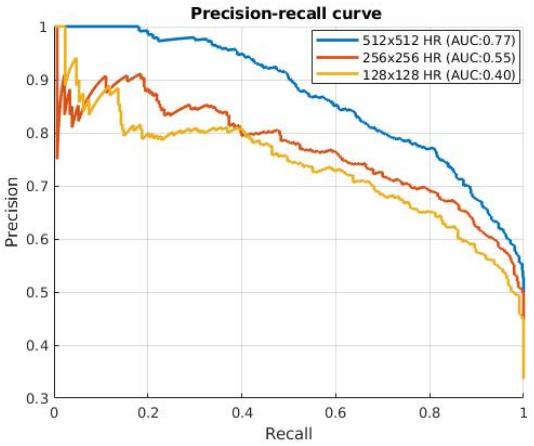
Table 1: Effects of Super-Resolution for upscale factor 2 on SSD performance

Input Resolution	mAP(Input)	SR(2x)	mAP(SR 2x)	Original HR	mAP(HR)
128x128	40.25	256x256	44.34	256x256	55.47
256x256	55.47	512x512	74.56	512x512	77.81

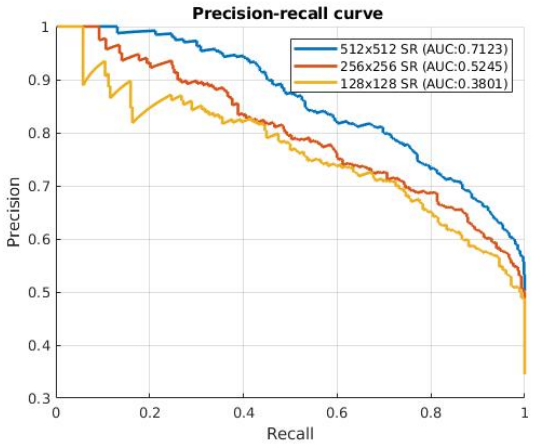
Table 2: Effects of Super-Resolution for upscale factor 4 on SSD performance

Input Resolution	mAP(Input)	SR(4x)	mAP(SR 4x)	Original HR	mAP(HR)
128x128	40.25	512x512	67.41	512x512	77.81

We also include our experimented results calculating AUC for detail explanation. As shown in Figure 4a.,the system performance trained with high resolution image degrades with the decrease in resolution. In this figure, we observe that performance of our proposed scheme is 20% higher in 512x512 resolution than resolution of 256x256 and 30% higher than 128x128 resolution. We investigate the network fine tuned with super-resolved images which is already trained on original high resolution images and report the results in Figure 4(b) that shows a little increase in performance. We can compare the detection results in terms of resolution from Figure 5. We observe that when we perform detection on 512x512 super-resolved images from 256x256 input resolution, the performance gain is close to original 512x512 high-resolution image. But when we experiment on 256x256 super-resolved images from 128x128 input resolution, we don't achieve similar performance gain though in both case the resolution is increased by two factor. In order to see the effect of upscale factor 4 on the detection performance, we conduct our experiment on 512x512 super resolved images where the input resolution is 128x128. During this experiment, We notice a great improvement in detection performance illustrated in Figure 6. We can summarize that super-resolution helps detection mostly in high-resolution.

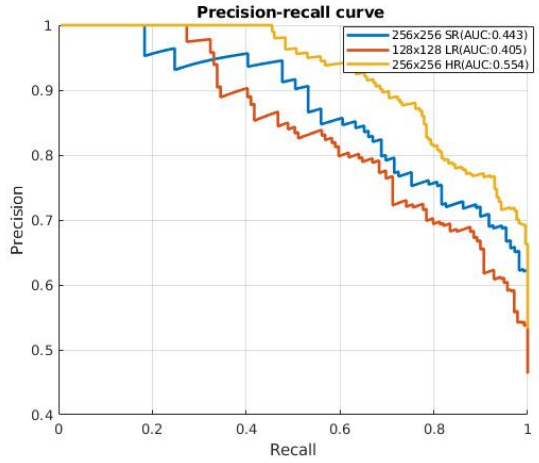


(a) Detection performance on original high resolution aerial images

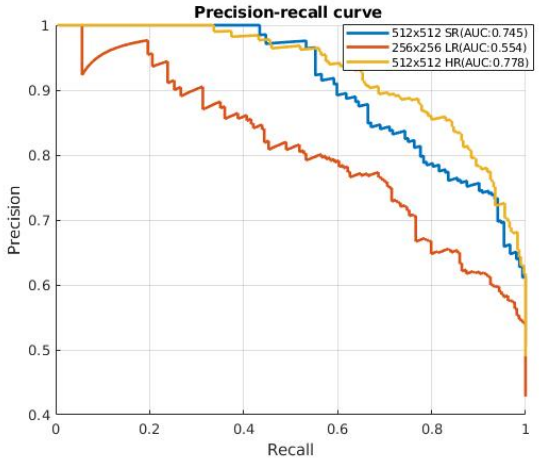


(b) Detection performance of the model fine-tuned with super resolved aerial images

Figure 4: Precision-recall curves for high resolution and super-resolved aerial images



(a) Detection performance on 256x256 super-resolved aerial images with its high resolution and low resolution input image of size 128x128



(b) Detection performance on 512x512 super-resolved aerial images with its high resolution and low resolution input image of size 256x256

Figure 5: Detection performance on super-resolved aerial images for the upscale factor 2

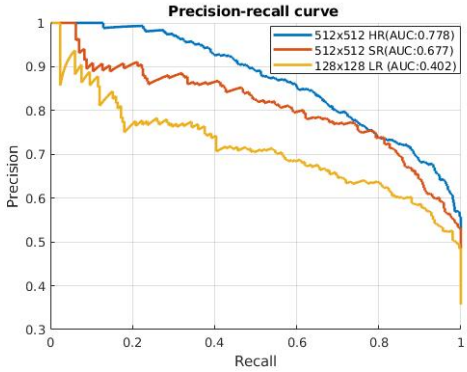


Figure 6: Detection performance on super-resolved aerial images for the upscale factor 4

6. CONCLUSION

This paper presents a model for real-time vehicle detection in aerial image combining two deep networks.The first model augments the features of the low resolution images by super-resolution. During experiments of this stage, we have found that ResNet has a great impact in recovering finer texture details; however it still has limitations to recover missing information from low resolution aerial image. Then, to identify objects from aerial images, the second model utilizes these features and performs detection in real time. Our proposed scheme gives faster object recognition with competitive performance. We have also demonstrated our detection performance in a comparative analysis of multiple resolutions.

In future, we intend to modify the existing architecture that helps to generate better prediction results.Other aerial datasets like COWC,[61] DOTA,[62] X-VIEW[63] can also be explored to get the comparable performances. Furthermore, we can extend our work for vehicle tracking from aerial video. Our proposed system can be rebuilt in such a way that loss optimization of the entire network takes place in combination for increasing the performance of the detector.

ACKNOWLEDGMENTS

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[2] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[3] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, **1**, pp. I–I, IEEE, 2004.

[4] D. Dai, R. Timofte, and L. Van Gool, "Jointly optimized regressors for image super-resolution," in *Computer Graphics Forum*, **34**(2), pp. 95–104, Wiley Online Library, 2015.

[5] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International journal of computer vision* **40**(1), pp. 25–47, 2000.

[6] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3791–3799, 2015.

[7] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE international conference on computer vision*, pp. 1920–1927, 2013.

[8] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE transactions on image processing* **21**(8), pp. 3467–3478, 2012.

[9] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Citeseer, 2008.

[10] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing* **19**(11), pp. 2861–2873, 2010.

[11] F. Taherkhani, N. M. Nasrabadi, and J. Dawson, "A deep face identification network enhanced by facial attributes prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 553–560, 2018.

[12] F. Taherkhani and M. Jamzad, "Restoring highly corrupted images by impulse noise using radial basis functions interpolation," *IET Image Processing* **12**(1), pp. 20–30, 2017.

[13] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*, pp. 711–730, Springer, 2010.

[14] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications* (2), pp. 56–65, 2002.

[15] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , pp. 5197–5206, 2015.

[16] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1823–1831, 2015.

[17] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2400–2407, IEEE, 2010.

[18] K. Zhang, X. Gao, D. Tao, and X. Li, "Multi-scale dictionary for single image super-resolution," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1114–1121, IEEE, 2012.

[19] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian conference on computer vision*, pp. 111–126, Springer, 2014.

[20] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence* **32**(6), pp. 1127–1133, 2010.

[21] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proceedings of the IEEE international conference on computer vision*, pp. 370–378, 2015.

[22] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 399–406, Omnipress, 2010.

[23] V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Multibiometric secure system based on deep learning," in *2017 IEEE Global conference on signal and information processing (globalSIP)*, pp. 298–302, IEEE, 2017.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, pp. 184–199, Springer, 2014.

[25] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence* **38**(2), pp. 295–307, 2016.

[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016.

[27] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," *arXiv preprint arXiv:1511.05666* , 2015.

[28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[29] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," *Image and Vision Computing* **21**(8), pp. 693–703, 2003.

[30] J.-Y. Choi and Y.-K. Yang, "Vehicle detection from aerial images using local shape information," in *Pacific-Rim Symposium on Image and Video Technology*, pp. 227–236, Springer, 2009.

[31] H.-Y. Cheng, C.-C. Weng, and Y.-Y. Chen, "Vehicle detection in aerial surveillance using dynamic bayesian networks," *IEEE transactions on image processing* **21**(4), pp. 2152–2159, 2012.

[32] H. Kazemi, S. Soleymani, F. Taherkhani, S. Iranmanesh, and N. Nasrabadi, "Unsupervised image-to-image translation using domain-specific variational information bound," in *Advances in Neural Information Processing Systems*, pp. 10369–10379, 2018.

[33] J. Carlet and B. Abayowa, "Fast vehicle detection in aerial imagery," *CoRR* **abs/1709.08666**, 2017.

[34] J. O. d. Terrail and F. Jurie, "Faster rer-cnn: application to the detection of vehicles in aerial images," *arXiv preprint arXiv:1809.07628* , 2018.

[35] A. Soleimani, N. M. Nasrabadi, E. Griffith, J. Ralph, and S. Maskell, "Convolutional neural networks for aerial vehicle detection and recognition," in *NAECON 2018-IEEE National Aerospace and Electronics Conference*, pp. 186–191, IEEE, 2018.

[36] M. Y. Yang, W. Liao, X. Li, and B. Rosenhahn, "Vehicle detection in aerial images," *arXiv preprint arXiv:1801.07339* , 2018.

[37] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 684–690, AAAI Press, 2018.

[38] F. Taherkhani, V. Talreja, H. Kazemi, and N. Nasrabadi, "Facial attribute guided deep cross-modal hashing for face image retrieval," in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–6, IEEE, 2018.

[39] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi, "Using deep cross modal hashing and error correcting codes for improving the efficiency of attribute guided facial image retrieval," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 564–568, IEEE, 2018.

[40] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," *arXiv preprint arXiv:1812.04098* , 2018.

[41] L. Cao, R. Ji, C. Wang, and J. Li, "Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[42] G. Polatkan, M. Zhou, L. Carin, D. Blei, and I. Daubechies, "A bayesian nonparametric approach to image super-resolution," *IEEE transactions on pattern analysis and machine intelligence* **37**(2), pp. 346–358, 2015.

[43] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications* **25**(6), pp. 1423–1468, 2014.

[44] B. Normalization, "Accelerating deep network training by reducing internal covariate shift," *CoRR.–2015.–Vol. abs/1502.03167.–URL: http://arxiv. org/abs/1502.03167* , 2015.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[46] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

[47] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434* , 2015.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[49] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, "A comparative study of sift and its variants," *Measurement science review* **13**(3), pp. 122–131, 2013.

[50] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *international Conference on computer vision & Pattern Recognition (CVPR'05)*, **1**, pp. 886–893, IEEE Computer Society, 2005.

[51] L. Wang, *Support vector machines: theory and applications*, vol. 177, Springer Science & Business Media, 2005.

[52] R. E. Schapire, "Explaining adaboost," in *Empirical inference*, pp. 37–52, Springer, 2013.

[53] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence* **32**(9), pp. 1627–1645, 2010.

[54] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[55] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[56] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[57] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[58] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* , 2014.

[60] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery (vedai): a benchmark," *Tech. Rep.* , 2015.

[61] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *European Conference on Computer Vision*, pp. 785–800, Springer, 2016.

[62] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018.

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