Does Deep Super-Resolution Enhance UAV Detection?

Vasileios Magoulianitis, Dimitrios Ataloglou, Anastasios Dimou, Dimitrios Zarpalas, Petros Daras
Information Technologies Institute, Centre for Research and Technology Hellas
6th km Harilaou - Thermi, 57001, Thessaloniki, Greece
{magoulianitis, ataloglou, dimou, zarpalas, daras}@iti.gr

Abstract

The popularity of Unmanned Aerial Vehicles (UAVs) is increasing year by year and reportedly their applications hold great shares in global technology market. Yet, since UAVs can be also used for illegal actions, this raises various security issues that needs to be encountered. Towards this end, UAV detection systems have emerged to detect and further anticipate inimical drones. A very significant factor is the maximum detection range in which the system's senses can "see" an upcoming UAV. For those systems that employ optical cameras for detecting UAVs, the main issue is the accurate drone detection when it fades away into sky. This work proposes the incorporation of Super-Resolution (SR) techniques in the detection pipeline, to increase its recall capabilities. A deep SR model is utilized prior to the UAV detector to enlarge the image by a factor of 2*. Both models are trained in an end-to-end manner to fully exploit the joint optimization effects. Extensive experiments demonstrate the validity of the proposed method, where potential gains in the detector's recall performance can reach up to* 32.4%*.*

1. Introduction

Living in the decade where automation and artificial intelligence have drawn unprecedented attention, has led to contemporary applications in almost every field of science. Unmanned Aerial Vehicles (UAVs) could not be absent from such an evolution and unsurprisingly incorporate cutting-edge technology. Their usage spans from entertaining applications -aerial video captioning- up to military and surveillance services, to monitor a certain region from above or for 3D surface mapping [15]. Although, high-tech UAVs can be very helpful for humans, automating many of their tasks, they can be also used for illegal activities, such as carrying explosives on to targets, area surveillance for espionage, terrorist attacks to buildings/people or disturbing air traffic. As can be inferred, quite recent, this has given rise to research on counter-UAV systems with key aim to detect drones well in advance.

978-1-5386-9294-3/18/$31.00 ©2019 IEEE

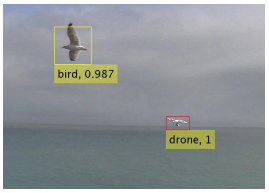

Figure 1. Detection result example of our method from WOSDETC Drone-vs-Bird Challenge, 2019.

A system that senses impending drones –the terms UAV and drone are used interchangeably– from distance, may employ visual, infra-red, acoustics or radar sensory data, with each modality having its advantages and shortcomings. This work examines the drone detection problem under the perspective of visual content. Recent advances of deep learning in computer vision have made it possible to detect objects [12] with high validity, in terms of recall-precision metrics in real-time. Deep Neural Networks (DNNs) learn hidden representations from a bunch of data which are in turn used to recognize the objects which have been trained on [13]. However, the key aspect of this problem is how far in distance the object detector can reach, especially when the object appears distorted from ambient light and occupies merely a few pixels within the image. Hence, it is straightforward to see that the quality of input content, as well as the scale are of utmost importance, because those objects that appear in long distances are more susceptible to distortion effects, thereby challenging the detector performance.

Super-resolution (SR) technique has been studied for many years to increase the image resolution, while preserving fidelity in terms of quality [3, 15]. Recently, much research on the field has been focused on exploiting DNNs that have proved to be superior in performance [19, 1, 6], when compared to old-fashion SR techniques. This work exploits such mechanisms to realize whether SR improves the detection recall of distant drones, when used not solely as a pre-processing step into the object detector's pipeline but being part of the entire optimization process, affecting the weights and the training of the detector module as well. In doing so, small drones that almost fade away into

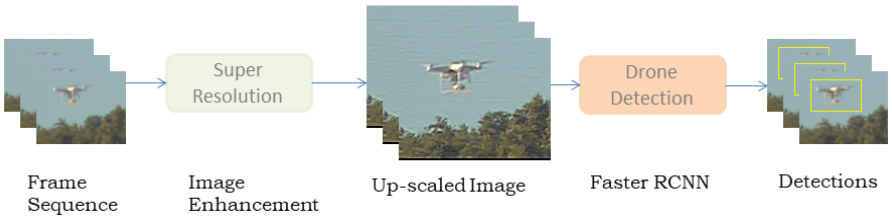

Figure 2. Drone detection pipeline with SR

sky, can be enlarged and further enhanced, so that they can be detectable by the DNN detector. Notably, our method achieved the second place at the WOSDETC Drone-vs-Bird Detection Challenge [1] 2019.

The main contributions of this work can be summarized as:

- Demonstrate how SR can be integrated into the drone detection pipeline so that the two distinct models can be jointly trained for improving the recall performance
- Extensive quantitative comparisons and results are given to validate whether the proposed deep SR-Detection method enhances the detection performance

2. Related Work

The brief literature review is focused on two main areas: super-resolution enhancement and drone detection, as either topics pertain our proposed method for accurate distant drone detection.

2.1. Super-Resolution

Among an abundance of works that revolve around SR techniques, we will only focus on those using the deep learning paradigm since they provide state-of-the-art results. Most of the works learn a mapping between low/high-resolution patches directly from data by using different architectures. For instance, [6] bridges the gap between traditional methods and DL by demonstrating how traditional sparse-coding-based SR techniques can be realized under the CNN perspective. Kim *et al*. [9] propose a very deep architecture by cascading small filters successively, to exploit the contextual information over the large image region. Besides, Ahn *et al*. [1] implement a cascading mechanism upon a residual network to provide a lightweight method. Two interesting works [17, 19] show that the combination of low- and high-level features by using skip connections yielded better performance results, improving the computational burden as well.

[1] https://wosdetc2019.wordpress.com/

2.2. UAV Detection

The problem of drone detection is becoming very popular [2, 13, 7, 5, 14] where different mechanisms are being adopted for tackling the various challenges that occur, such as cluttered background –clouds, ambient light–, similar flying objects –airplanes, birds– [14], as well as large variations in distance and drone type. In general, state-of-the-art generic object detectors, if properly trained on drone data, provide a very elegant solution for drone detection. Recent research on the drone detection problem are extensively using deep learning methods, where DNN-based detectors provide the best performance. For instance, Wu *et al*. [18] modify a popular's network [11] parameters to better fit the drone detection problem. Aker *et al*. [2] address the scarcity of training data for DNNs by proposing an artificial dataset with drones in different flying scenarios and complex background. Saqib *et al*. carry out a study for several object detectors which shows that the VGG-16 [16] performs best as the core DNN in a Faster-RCNN pipeline, at a higher complexity though.

3. Proposed Pipeline

3.1. Super Resolution

In Section 2, some of the ways to enhance the image quality and scale were discussed. In this work, taking into account the best tradeoff between accuracy, in terms of Peak-to-Signal-Noise-Ratio (PSNR), and time complexity, we opt for the Deep Residual CNN with Skip Connection and Network in Network (DCSCN) [19] model. The DCSCN consists of two networks, the feature extraction and the reconstruction one.

3.1.1 Feature Extraction

Most of the deep learning models which perform Single-Image-Super-Resolution (SISR), at their very first step apply a bicubic upscale at the input image before the first convolutional layer and afterwards the network merely improves the local feature representation. It is straightforward to see that applying the convolutional filters on the upscaled

image, results in considerably higher computational complexity and at the expense of fidelity. To this end, the DCSCN model operates on the raw image which is upsampled through the deep network at a later stage. Furthermore, the number of filters is decreased for the sake of complexity and smoother training of the model. Also, the parameters at the subsequent and deeper layers diminish in order to emphasize more on the local, rather than the global features, which is more consequential for the SISR problem. The Parametric Rectified Linear Unit (PReLU) is also selected as activation layer after each convolutional one.

3.1.2 Reconstruction Network

After the feature extraction process, all the intermediate features are concatenated using skip connections, thus resulting in a large feature volume, where 1×1 convolutional layers are employed to reduce its dimensionality, before further processing. The reconstruction network comprise four convolutional layers, following the Network in Network model setting [10]. The reconstruction network learns the residual upscaled image and eventually adds it to the bicubic upsampled input image for improving its PSNR. The entire model is comparably shallow with other deep models, thereby emphasizing more on the learning of local features. Besides, the shallower architecture leads in turn to fewer parameters, casting the model an elegant option to be used prior to the detector, getting them jointly optimized.

This work examines two versions of the DCSCN: The full model parameters and a more compact version (c-DCSCN) with the number of filters truncated to realize the tradeoffs between computational time and recall performance.

3.2. Object Detection

The object detector is based on the well established Faster-RCNN model [12]. This generic object detection architecture is based on a Region Proposals Network (RPN) and enables fast and accurate object detection.

RPN uses the intermediate deep feature representation of the input image to suggest potential object locations. It employs a sliding window mechanism on the feature map to estimate confidence scores about how possible is each region to include an object. Using the official terminology, the RPN evaluates the "objectness" of the region. Also a regression layer, makes the refinement over the initial anchor box position to improve the object localization. As such, the RPN derives k regions that might have objects at some probability. In our work, we retain the h most confident object regions per frame, while with Non-Maximal Suppression (NMS) algorithm, those regions which have high overlaps are merged to a single proposal. The h parameter, which also dictates the maximum number of detections has been set to 64, thus expediting the testing time of the module. The feature volume that corresponds to each of the 64 region proposals is cropped and pooled from the intermediate feature volume such that each region proposal is represented by a $7 \times 7 \times 512$ feature. The batch with all region proposals (here up to $64 \times 7 \times 7 \times 512$) is in turn forwarded to the classifier to assign a confidence on each class and to refine the coordinates of the bounding box, as suggested by the RPN. We chose to replace the backbone network within the Faster-RCNN architecture with MobileNet [8], since it provides a computational inexpensive solution for object detection at high detection performance.

3.3. End-to-end training

To fully leverage the interoperability between the two deep models, by the time they are used sequentially in the detection pipeline, the models are jointly being optimized as one thing. In particular, we train afresh the DCSCN model on generic data (see Section 4.1) and use the pre-trained weights in the end-to-end optimization to favor the training process. In doing so, the DCSCN model's weights are fine-tuned and adjust so to favor the drone detection model which follows. As such, the backward propagation of the gradients from the detector's classifier up to the LR input image affect all the inside networks. In other words, as the training process goes along, the DCSCN model learns to enhance the input image quality and scale, targeted to the drone detection task, while the Faster RCNN learns to take advantage from the enhanced images towards the same task.

4. Experiments

4.1. Training and Testing Setting

4.1.1 Super Resolution Model

For both the feature extraction and reconstruction networks, biases and PReLUs parameters are initialized to zero. A dropout with $p = 0.8$ is applied to all layers to prevent overfitting effects. Moreover, we utilize the Mean Squared Error (MSE) as the objective, where we add the sum of L2 norms of each CNN's weight, scaled by a factor of 10^{-4}, following the original setup [19]. We use Adam optimizer with initial learning rate of 2×10^{-3}. We decrease the learning rate by a factor of two when loss plateaus for five epochs and when it reaches 2×10^{-5} we stop training. The SR model is trained afresh on the Berkeley Segmentation Dataset [4] with data augmentation techniques, such as horizontal and vertical flipping, thus making the training images as many as three times of their initial number. Having been trained, the DCSCN is used pre-trained in the end-to-end optimization (Fig. 2). The input images have size of 1920×1080 and are upsampled by a factor of two from the DCSCN.

Table 1. Comparative results between LR and SR videos

Clip	LR		SR	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
FiveDistantDrones	20.41	99.91	52.88	100
FiveDrones	66.67	100	88.47	100
DoubleDrones	92.26	100	93.15	100
SingleDrone	92.58	99.17	94.12	99.12

4.1.2 Drone Detector

For the RPN, the NMS uses an Intersection over Union (IoU) threshold of 0.7 to merge highly overlapping detections. The Faster-RCNN was trained for $70K$ iterations with an initial learning rate of 10^{-3}, which was reduced by a factor of 10 during the last 20000 iterations. Trainable parameters were optimized using stochastic gradient descend (SGD) with 0.9 momentum and a 4×10^{-5} weight decay. The base CNN (MobileNet) was initialized with weights from a pre-trained model on the ImageNet dataset. The first 5 layers of the Base CNN, as well as all Batch Normalization layers, were kept fixed during training. Among the bunch of video data with UAVs, we found that a balanced sampling of roughly $16K$ frames suffice for training the detector pipeline. Train data comprise a variety of UAVs appearing from close distance up to far away and a commensurate amount of frames with annotated birds, all coming from the WOSDETC challenge and other publicly available datasets, to form a rich bunch of training data. Therefore, the detector is trained to predict three classes; "drone", "bird" and "rest". We found this training setting to assist the detector for more efficiently disentangling between drones and birds, thus increasing its precision. For validation set we use 2814 randomly sampled frames –not successively in sequence– to tune the detector's hyper-parameters. Finally, the detection threshold is set to 0.9 to proclaim "drone" a yielded detection. All experiments were performed on a NVIDIA GeForce TITAN XP with 12GB memory.

4.2. Test Data

To evaluate our initial hypothesis, on how conducive can be the incorporation of the SR technique in the drone detection pipeline, we have selectively pick short video sequences with two types of UAVs that soar around the sky, including many hard cases of UAVs flying scenarios in terms of recall performance. Notably, the bare detector provides adequate results in detecting drones with high precision and fails mainly when drones fly and fade away into sky at distances higher than 200 meters. Hence, in order to assess the improvement offered by the SR model under such circumstances, our experimental data includes mainly flight scenarios where drones are at the verge of getting detected.

Within the bunch of available data, four short video clip sequences have been extracted with difficult challenging scenarios. These scenarios depict drones that typically fly in distances from 120 up to roughly 280 meters. In the *FiveDistantDrones* clip five Phantom-4 drones appear to fly in fixed shape and at large distance. The *FiveDrones* clip shows the same drones to fly in a bit shorter distance. *DoubleDrone* clip shows one Parrot drone which is flying and intermittently another one of the same type, passing through the scene with higher speed. The last clip, *SingleDrone* comprises the same drone flight scenario that flies in distance alone.

4.3. Evaluation

To validate our intuition on the usefulness of the SR image enhancement technique before applying the drone detector, we provide extensive experiments with different settings to draw inferences for each approach.

To begin with, Table 1 demonstrates the recall and precision results, where either the low resolution (LR) images or the enhanced SR ones are used as input to the detector. The precision performance approaches the absolute in all test sequences, since the background is mainly naive in most of the test frames, thereby leaving no space for improvements and yielding only imperceptible differences. Hence, our analysis will be mainly focused on the recall metric, which is the initial subject that this work opts for optimizing.

The first sequence "*FiveDistantDrones*" yields outstanding improvement, by recalling almost 32% more UAVs than the LR input image. The "*FiveDrones*" recall increased by 22%, which shows that although the drones are not flying that far, the bare detector failed to retrieve them in some instances and the SR step improved the whole performance pipeline. In the rest two clip sequences, the bare detector achieves already sufficient results and the margins for large improvements are rather narrow, but we can still notice a small but significant recall increment. Furthermore, the precision performance seems that is not affected from the application of the SR model.

In order to realize tradeoffs between the full version of DCSCN and its more compact version, Table 2 demonstrates that in all clip sequences the full version demonstrates only marginally better recall results. However, the time complexity benefits from the compact version (c-DCSCN), the latter seems to be more worthy for real-time applications, since it needs much less computational effort –81.25% faster–, while on the other hand, the mean loss in

Table 2. Comparative results between DCSCN and c-DCSCN

Clip	DCSCN		c-DCSCN	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
FiveDistantDrones	52.88	100	52.37	100
FiveDrones	88.47	100	88.34	100
DoubleDrones	93.15	100	92.89	100
SingleDrone	94.12	99.17	93.93	99.12

Table 3. Comparative results between linear and bicubic upsampling methods

Clip	Linear		Bicubic	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
FiveDistantDrones	25.77	100	36.33	100
FiveDrones	61.78	100	74.17	100
DoubleDrones	85.24	100	88.71	100
SingleDrone	90.41	99.29	92.01	99.17

Table 4. Time complexity vs mean recall ratio gain tradeoffs. The mean gain in recall ratio for the four tested clip sequences

Model	Frames per Sec (FPS)	Mean Recall Gain (%)
DCSCN	0.32	14.17
c-DCSCN	0.58	13.9

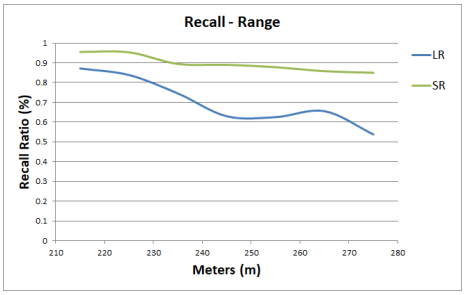

Figure 3. Recall Ratio - Distance diagram for the *FiveDistantDrones* clip

recall ratio for all the clip sequences is only 1.9% (Table 4).

One could argue that the image upscaling by a factor of two enlarges the drone size, as it appears on the image, which eventually helps the detector to find it more easily. To renounce this allegation, we performed an experiment, where we utilized two common and computationally inexpensive methods for upscaling the input image. Interestingly, comparing the recall results between Tables 1 and 3, for the "FiveDistantDrones" sequence the results are significantly improved –not as much as with SR–, but for the remaining three sequences, we can see that the recall performance drops instead, possibly due to distortion effects induced by the upsampling. Only the Bicubic method yields an improvement at the "FiveDrones" sequence but drops in performance for the rest as well. Therefore, the SR module usage before the detector is beneficial, if the recall factor is the aim for a UAV detection system which operates with visual data. The superiority of the usage of SR over LR is illustrated also in Fig. 3, where the recall factor for higher

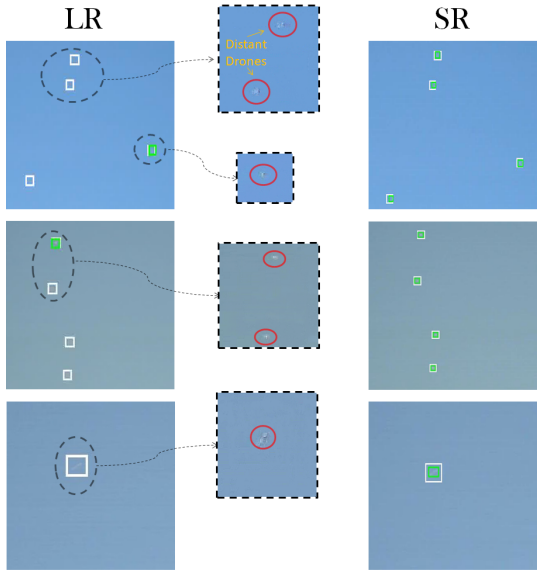

Figure 4. Qualitative results. White boxes refer to the ground truth (implies the UAV existence) and the green boxes correspond to the detections. Red circles indicate the distant drones in zoomed-in areas. **Best viewed in color**

distance flying UAVs is significantly improved. The impact on the full SR detection pipeline is also visualized in Fig. 4.

5. Conclusion

This work presents a UAV detection pipeline that embeds the SR technique, prior to the drone detector for jointly training both modules. In doing so, the entire pipeline benefits from learning to enhance the input LR image and in turn to improve the recall capabilities of the system. This is meant to retrieve drones that fly far away where typical state-of-the-art detectors fail in such scenarios. The impact of the SR model that performs the SISR is consequential for UAV detection at longer ranges and has been validated by extensive experimentation.

References

[1] N. Ahn, B. Kang, and K.-A. Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018. 1, 2

[2] C. Aker and S. Kalkan. Using deep networks for drone detection. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. 2

[3] S. Anwar, S. Khan, and N. Barnes. A deep journey into super-resolution: A survey. *CoRR*, abs/1904.07523, 2019. 1

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 3

[5] W. Budiharto, A. A. Gunawan, J. S. Suroso, A. Chowanda, A. Patrik, and G. Utama. Fast object detection for quadcopter drone using deep learning. In *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, pages 192–195. IEEE, 2018. 2

[6] 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):295–307, 2015. 1, 2

[7] M. Farhadi and R. Amandi. Drone detection using combined motion and shape features. In *IEEE International Workshop on Small-Drone Surveillance, Detection and Counteraction Techniques*, 2017. 2

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[9] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[10] M. Lin, Q. Chen, and S. Yan. Network in network. *International Conference on Learning Representations*, 2014. 3

[11] 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*, pages 779–788, 2016. 2

[12] 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*, pages 91–99, 2015. 1, 3

[13] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein. A study on detecting drones using deep convolutional neural networks. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017. 1, 2

[14] A. Schumann, L. Sommer, J. Klatte, T. Schuchert, and J. Beyerer. Deep cross-domain flying object classification for robust uav detection. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. 2

[15] H. Shakhatreh, A. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani. Unmanned aerial vehicles: A survey on civil applications and key research challenges. *arXiv preprint arXiv:1805.00881*. 1

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[17] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4799–4807, 2017. 2

[18] M. Wu, W. Xie, X. Shi, P. Shao, and Z. Shi. Real-time drone detection using deep learning approach. In *International Conference on Machine Learning and Intelligent Communications*, pages 22–32. Springer, 2018. 2

[19] J. Yamanaka, S. Kuwashima, and T. Kurita. Fast and accurate image super resolution by deep cnn with skip connection and network in network. In *International Conference on Neural Information Processing*, pages 217–225. Springer, 2017. 1, 2, 3