Received: 12 May 2021 | Revised: 10 November 2021 | Accepted: 29 November 2021

DOI: 10.1111/exsy.12930

ORIGINAL ARTICLE

Expert Systems WILEY

Improved detection of small objects in road network sequences using CNN and super resolution

Iván García-Aguilar[1] | Rafael Marcos Luque-Baena[1,2] | Ezequiel López-Rubio[1,2]

[1]Department of Languages and Computer Science, University of Málaga, Málaga, Spain

[2]Biomedical Research Institute of Málaga (IBIMA), Málaga, Spain

Correspondence
Iván García-Aguilar, Department of Computer Languages and Computer Science, University of Málaga, Bulevar Louis Pasteur, 35, Málaga 29071, Spain.
Email: ivangarcia@uma.es

Funding information
Universidad de Málaga; European Regional Development Fund (ERDF); Autonomous Government of Andalusia (Spain), Grant/Award Numbers: SNGJ5Y6-15, UMA18-FEDERJA-084; Ministry of Science, Innovation and Universities, Grant/Award Number: RTI2018-094645-B-I00

Abstract

The detection of small objects is one of the problems present in deep learning due to the context of the scene or the low number of pixels of the objects to be detected. According to these problems, current pre-trained models based on convolutional neural networks usually give a poor average precision, highlighting some as CenterNet HourGlass104 with a mean average precision of 25.6%, or SSD-512 with 9%. This work focuses on the detection of small objects. In particular, our proposal aims to vehicle detection from images captured by video surveillance cameras with pre-trained models without modifying their structures, so it does not require retraining the network to improve the detection rate of the elements. For better performance, a technique has been developed which, starting from certain initial regions, detects a higher number of objects and improves their class inference without modifying or retraining the network. The neural network is integrated with processes that are in charge of increasing the resolution of the images to improve the object detection performance. This solution has been tested for a set of traffic images containing elements of different scales to check the efficiency depending on the detections obtained by the model. Our proposal achieves good results in a wide range of situations, obtaining, for example, an average score of 45.1% with the EfficientDet-D4 model for the first video sequence, compared to the 24.3% accuracy initially provided by the pre-trained model.

KEYWORDS

convolutional neural networks, object detection, small scale, super-resolution

1 | INTRODUCTION

Currently, object detection is one of the most popular computer vision applications of deep learning. The proliferation of this type of application is caused by the increase of video data obtained from different sources, as well as the improvement in the computational power of the hardware, thereby facilitating the successful accomplishment of this task. The area of road network management can be considered as one of the potential application domains of this type of technology because there are many objects to detect (vehicles), the scenarios are very heterogeneous due to the position of the camera, angle, orientation and distance to the road, and there is a huge number of already installed traffic cameras with the possibility of capturing relevant information.

The proposed solution has as its primary objective the detection of reduced size elements (objects) using convolutional neural networks. Small elements are those which occupy a small region within the entire image. At present, some pre-trained models have significant intrinsic problems that need to be addressed. Object detection models determine the number of features by aggregating information from the raw pixels across the layers of the convolutional network. Most of them reduce the resolution of the images in intermediate layers. This fact causes the loss of the features of small objects, which disappear during the processing carried out by the network, thus avoiding their detection. The low detection rate of this type of element is also caused by the generic background clutter in the images to be inferred, thus making the task of detecting new elements more complex due to many potential object locations. The small objects to be detected have simple shapes that cannot decompose into smaller parts or features. Another point to consider is the quality of the image since there are video surveillance systems with low image quality. Poor quality images negatively affect the performance of object detection methods, because these types of sequences, especially in the case of traffic videos, are dense, with small vehicles and partially occluded. On the other hand, some objects are similar in feature, shape, colour or pattern. Therefore, existing pre-trained high-quality models cannot distinguish among them accurately.

The small object detection problem is relevant in many areas. Direct applications include the following: the enhancement in the classification and detection of elements through images captured by satellites, the improvement in video surveillance through the treatment of images provided by security cameras established at high points, and the increase in the detection of pedestrians or traffic. This article focuses on the last of these direct applications. Thanks to this new solution, it will be possible to detect a larger number of elements and improve the confidence of each detection without the need to retrain the model so that this solution may be advantageous for automated traffic control systems.

Currently, there are some models aimed at the detection of these elements through two workflows. The first of them follows the usual flow in which a series of candidate regions are generated to perform the classification of each proposed area subsequently. The second method establishes the detection of objects as a regression or classification problem to adopt a framework to achieve final results. Among the methods based on regional proposals, we find mainly R-CNN, Faster R-CNN and Mask R-CNN, while in regression methods, we can mention SSD and YOLO (Zhao et al., 2019). As an example, we find the model called *EfficientDet*, which has an average detection rate of 51% for medium-sized elements. In contrast, smaller elements are detected in only 12% of the cases. It should also be noted that there are no datasets devoted explicitly to the detection of small objects, which is why most of the attained detections correspond to larger elements. When applying existing models such as *Faster-RCNN* it is found that it misses several small objects in part due to the size of the anchor boxes of the elements detected in the image.

The proposal put forward in this article is based on the design, implementation and subsequent testing of a technique based on convolutional neural networks, capable of detecting small-scale elements and improving the class inference, without the need to retrain the pre-trained model. To achieve this goal, the hyper-parameters of the network have been modified to reduce false negatives. Subsequently, super-resolution processes have been applied to the input image to generate a series of images centred on each of the detected elements to increase its resolution. Then the model yields an enhanced inference about the object class. Using this method, implemented in conjunction with image pre-processing through a denoising filter, small object detection and class inference are significantly improved.

Additionally, it has been necessary to create a new dataset to perform the relevant tests to evaluate the qualitative improvement attained by the newly developed technique. Most existing datasets are not adequate, such as *KITTI* (Geiger et al., 2013), which is composed of a series of images collected by cameras located on mobile vehicles. This set of images contains several vehicles. However, the size of these images is large, and it should be noted that not all images include vehicles, and in other cases, the number of vehicles is small. It should be noted that nowadays, traffic-oriented video surveillance systems produce sharp and clear images. With this in mind, it can therefore be established that there are no appropriate public datasets on which to perform tests, so a new dataset of vehicle detection data has been developed from a series of traffic videos captured by cameras and surveillance systems intended for that purpose. Three test sets filmed in different locations and with a series of challenges such as image quality, light interference, motion blur and light interference, among others, have been compiled. This dataset consists of 476 images with a total of 14,557 detections. These images contain a set of challenges to be solved, as shown in Figure 1.

Under the premises described above, it can be concluded that there are significant shortcomings in the state of the art methods for the detection of small-scale elements, given the low detection rate of current models and the few performance-enhancing procedures.

The rest of this article is organized as follows. Section 2 sets out the related work. Throughout Section 3, the improvements developed are detailed, explaining in depth the implemented workflow. Section 4 includes the performed tests along with their respective results. Finally, in Section 5, the conclusions and the future works to be developed according to the proposed solution are outlined.

2 | RELATED WORK

Given the context of our proposal, there are pre-trained models such as, for example, *YOLO v4* (Bochkovskiy et al., 2020). This network uses as a feature extractor CSPDArknet53 for the GPU version, spatial pyramid pooling known as SSP, and Path Aggregation Network (PAN) to speed up the inference process. Head uses the same design for the *YOLO* v3 network. The model divides the image into a grid to subsequently predict the class labels for each bounding box. Another frequently used model is *CenterNet* Duan et al. (2019). This model detects each object as a triplet instead of a pair to improve the accuracy in detections and the element class, thus presenting an efficient solution that explores the visual

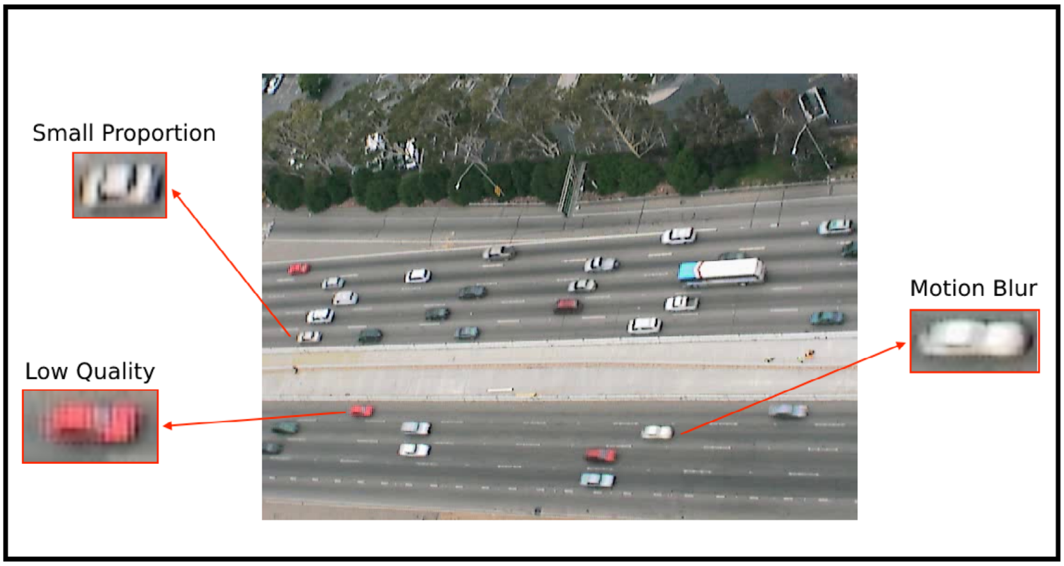

FIGURE 1 Main challenges to be solved

patterns within each cropped region with minimal cost. According to previous studies, it has been established that methods based on regional proposals obtain a higher hit rate when detecting elements in an image. However, in the area of small object detection, the average accuracy rates are poor, with 24.3% for *YOLO v4* and 28.9% for *CenterNet* model.

In the field of small object detection, there are numerous works focused on pedestrian detection. Works such as the one established by Han et al. (2020) determine a novel deep small-scale sense network (named as SSN). The proposed architecture generates some proposal regions that are more effective for detecting small-scale pedestrians. On the other hand, the loss function based on cross-entropy loss increases small-scale elements that are difficult to detect. The method extracts convolutional features with a *VGG* network that are transferred to the scale-wise proposal network. Another work to be highlighted is that of Lai et al. (2020), based on a one-stage method that improves the detector performance. It reconstructs the network with the deconvolutional method, thus combining the up-sampling method and deconvolution parameters into a deconvolutional layer. Moreover, it adjusts the prior anchor parameter by the *k*-means clustering algorithm. Both presented works improve the detection of small objects applied to the field of pedestrian detection. In the field of vehicles detection, we can highlight works such as the one proposed by (Kim et al., 2019) that recognizes the direction of the vehicle in the input image. However, as the distance from the vehicle increases, the image quality deteriorates, resulting in poorer accuracy in detecting small elements. In this field, there are no significant advances focused exclusively on improving the detection of small vehicles on urban roads, so we put forward the proposal presented in this paper.

The articles cited above are based on the modification of the internal structure of the network. Our proposal, unlike the previous related studies, employs the pre-trained models without modifying their structures, so it does not require re-training the network to improve the detection rate of the elements, thus enabling the improvement of the mean average precision (*MAP*) of small elements captured by video-surveillance systems located at high points. We take advantage of the fact that detection models based on convolutional neural networks perform well in the high-resolution domain. We use super-resolution to transfer the classification task to a pre-trained network in high-resolution images.

Image super-resolution processes have shown high potential in the field of image restoration and interpolation. They establish a series of techniques to increase the resolution of an input image compared to the conventional pixel-wise interpolation algorithms. There are several advances and developments for super-resolution applied to a single image. For example, (Dong et al., 2014) proposed a network based on super-resolution known as *SR-CNN*. This network establishes an end-to-end mapping between images given as input with low resolution and those processed by the model. This method jointly optimizes all layers, improving restoration quality. The network is composed of the extraction part, application of a non-linear mapping, and reconstruction. As an improvement of this, the model proposed by (Kim et al., 2015a) is developed and is known as *VDSR*. This model uses a deep convolutional network inspired by *VGG-net*, using 20 weight layers, and contextual information is exploited efficiently with global skip connection. However, convergence speed becomes a critical issue during training. Other advances such as the one proposed by, (Kim et al., 2015b) use a deeply recursive convolutional network with a deep recursive layer to improve performance without introducing new parameters for the additional convolutions called *DRCN*. However, as a negative point, learning this network is very difficult with a standard gradient descent method due to gradient explosion/fading. To mitigate the training difficulty, the authors propose recursive supervision and hop connection. One of the models for the application of super-resolution in real-time corresponds to the work presented by Dong et al. (2016). *FSRCNN* introduces a deconvolution layer at the end of the network to perform up-sampling. The non-linear mapping step in SRCNN is replaced by three steps in FSRCNN, namely the shrinking, mapping and expanding. Finally, the smaller filter sizes and a deeper network structure provide better performance and are tens of times faster than other models. In particular, *FSRCNN-s* which can be implemented in real-time on a generic CPU. The article presented by Cao and Chen (2019) provides a study of some of the super-resolution methods mentioned above, thus establishing a comparison to identify the one that gives the best results.

These processes remain at a primitive stage and are complex. There are significant advances, mainly focused on the creation of algorithms that perform these functions effectively. One of the main approaches in this field is presented by Xing et al. (2019), as they proposed a solution by using a generative adversarial network to recover small objects from low-resolution images to high resolution to get better detection performance, achieving a MAP of the 68.38%. Other work proposed by Mostofa et al. (2020) established a network called *Joint-SRVDNet* that generates high-resolution images of vehicles from low-resolution aerial images. First, images are up-scaled by a factor of 4× using a multi-scale generative adversarial network (*MsGAN*). The network jointly learns discriminative and hierarchical features of targets and produces optimal super-resolution results.

Although these approaches are related to the one presented in this paper, they fail to propose a scheme based on current convolutional neural network models. Thanks to the super-resolution processes as well as the pre-processing stage, our proposal finds out new visual elements. In this paper, we hypothesize an improvement in the detection of small-scale objects using super-resolution (*SR*) techniques for the enhancement in the inference provided by models such as *CenterNet HourGlass*, specifically for sequences obtained by video surveillance cameras set at high points.

3 | METHODOLOGY

Next, our small object detection system is presented. Figure 2 shows the global workflow of our proposed framework. As can be seen in the figure, we start from a pre-trained model that will establish a series of initial detections based on the detected vehicles. On each of these detections, a super-resolution process is applied to generate a new sub-image in which the elements contained in it have a higher number of pixels. With this implementation, it is possible to detect more elements close to the initial detections that were not located at the beginning, thereby improving the accuracy for more reliable detections without the need to modify or retrain the initial object detection network. Finally, a transformation process will be necessary to translate the coordinates. Next, each of the parts that make up the proposal presented will be explained in more detail.

In what follows, the methodology of our proposal is explained in detail. As set out in point 1 of the workflow, let us consider a deep learning neural network for object detection which takes an image $\mathbf{X}$ as input and returns a set of detections S as output:

$$S = \mathcal{F}(\mathbf{X}) \tag{1}$$

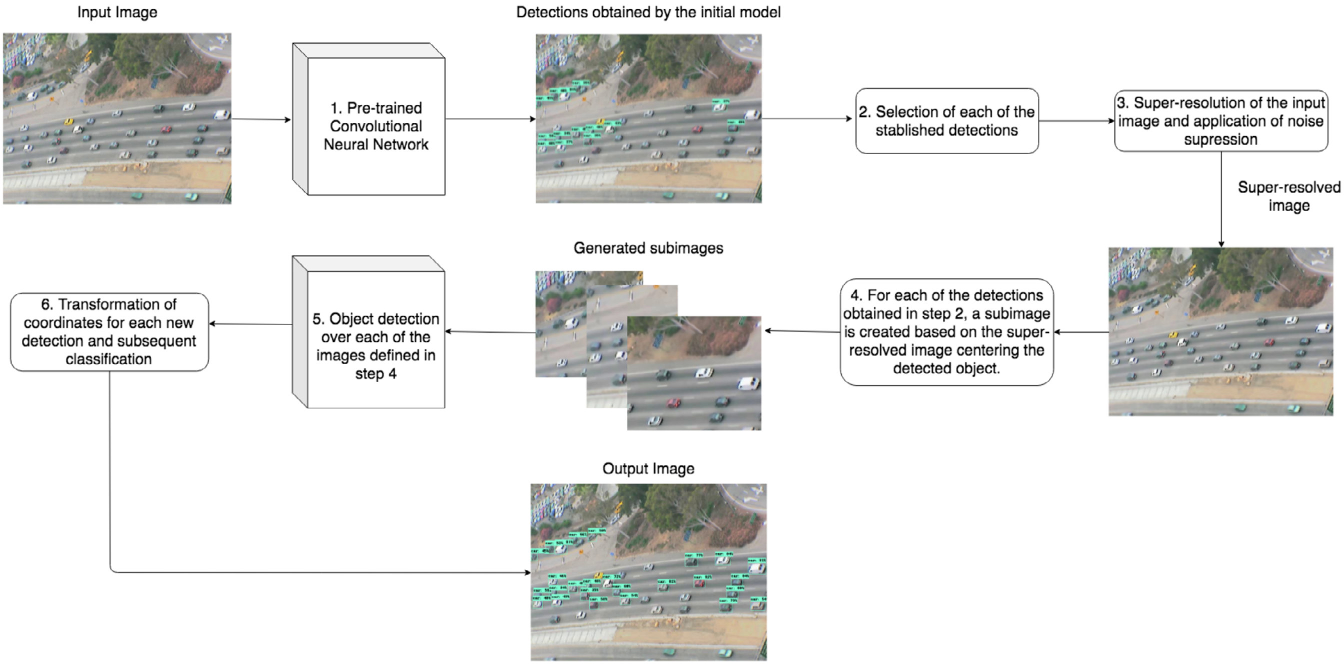
Input Image
1. Pre-trained Convolutional Neural Network
Detections obtained by the initial model
2. Selection of each of the stablished detections
3. Super-resolution of the input image and application of noise supression
Super-resolved image
4. For each of the detections obtained in step 2, a subimage is created based on the super-resolved image centering the detected object.
Generated subimages
5. Object detection over each of the images defined in step 4
6. Transformation of coordinates for each new detection and subsequent classification
Output Image

FIGURE 2 Workflow of the proposed technique

$$S=\{(a_i,b_i,c_i,d_i,q_i,r_i)\mid i\in\{1,\ldots,N\}\}\tag{2}$$

where N is the number of detections, $(a_i,b_i)\in\mathbb{R}^2$ are the coordinates of the upper left corner of the i-th detection within the image $\mathbf{X}$, $(c_i,d_i)\in\mathbb{R}^2$ are the coordinates of the lower right corner of the i-th detection within $\mathbf{X}$, q_i is the class label of the detection, and $r_i\in\mathbb{R}$ is the class score of the detection. The higher r_i, the more confidence that an object of class q_i is actually there. It is assumed that the origin of the coordinate system of all images is at the centre of the image. Next, step 2 is performed based on the selection of each of the initially established detections. Given a low-resolution input image $\mathbf{X}_{\mathrm{LR}}$ as shown, for example, in Figure 3, our first step is to process it with the object detection network to yield a set of tentative detections S_{LR}:

$$S_{\mathrm{LR}}=\mathcal{F}(\mathbf{X}_{\mathrm{LR}})\tag{3}$$

Subsequently, as detailed in the third step of the workflow, a new image is generated by applying super-resolution and denoising processes to the initial image given as input.

This process starts by applying a super-resolution network $\mathcal{G}$ to the input low-resolution image $\mathbf{X}_{\mathrm{LR}}$ to yield a super-resolved version $\widetilde{\mathbf{X}}_{\mathrm{HR}}$:

$$\widetilde{\mathbf{X}}_{\mathrm{HR}}=\mathcal{G}(\mathbf{X}_{\mathrm{LR}})\tag{4}$$

After that, a denoising procedure $\mathcal{D}$ is applied to $\widetilde{\mathbf{X}}_{\mathrm{HR}}$ to obtain the noise-removed, high-resolution version $\mathbf{X}_{\mathrm{HR}}$ of the input image $\mathbf{X}_{\mathrm{LR}}$:

$$\mathbf{X}_{\mathrm{HR}}=\mathcal{D}\left(\widetilde{\mathbf{X}}_{\mathrm{HR}}\right)\tag{5}$$

Figure 3 shows the process for the generation of the sub-images with which the pre-trained model will perform the inference to detect a higher number of vehicles. As shown in the image, the first step is to start with a low-resolution image given as input on which the pre-trained model will establish the initial detections. The input image is processed with super-resolution obtaining a new one with a higher number of pixels. Finally, the *Buades et al*. proposal, based on Gaussian noise reduction, will be applied to improve the number of detections as well as their accuracy. Once the new super-resolved image is generated and subsequently processed, based on each of the initial detections, a new sub-image will be generated, starting from the same one, according to the input size required by the model to perform the inference process.

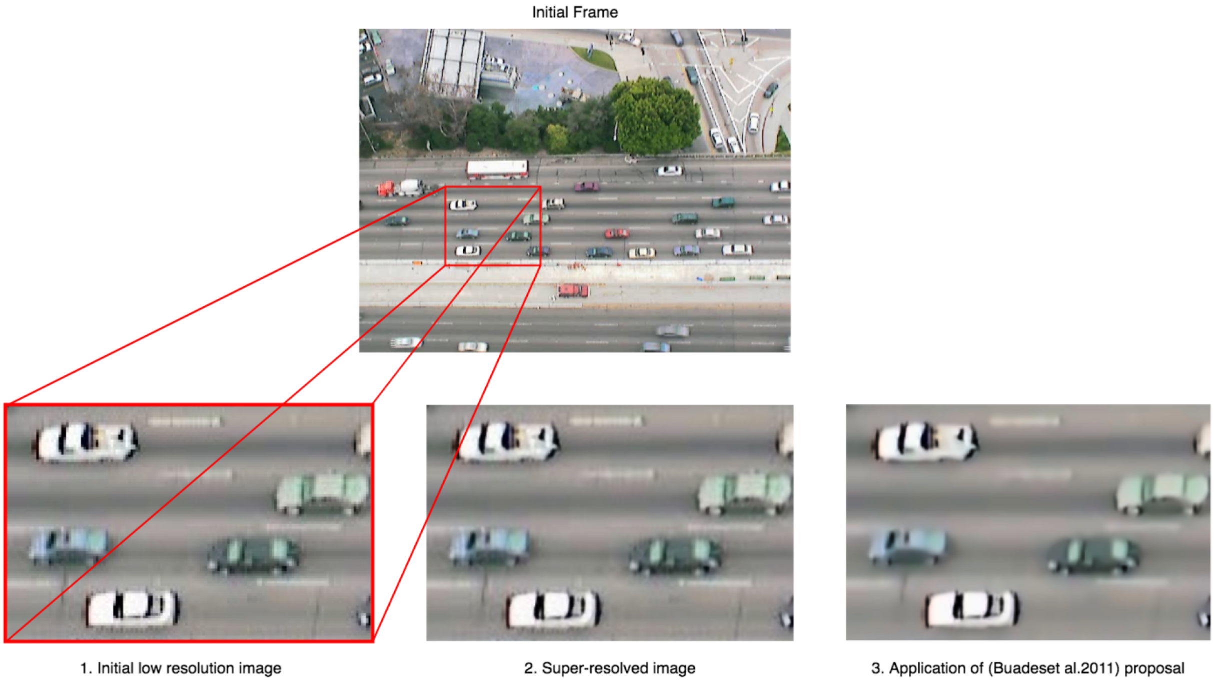
Initial Frame
1. Initial low resolution image
2. Super-resolved image
3. Application of (Buadeset al.2011) proposal

FIGURE 3 Frame processing

A super-resolution deep network is used to obtain a high-resolution version $\mathbf{X}_{\mathrm{HR}}$ with zoom factor Z of the low-resolution input image $\mathbf{X}_{\mathrm{LR}}$ as shown in point 2 of the Figure 3. To implement this super-resolution process, the OpenCV library has been used. There are some pre-trained models $\mathcal{G}$ for the execution of processes that increase the initial resolution of an image. The considered models are the following:

- SRCNN: Image SR using convolutional neural networks (Dong et al., 2015).
- FSRCNN: Fast super-resolution convolutional neural network (Dong et al., 2016).
- VDSR: Image SR using very deep convolutional networks (Kim et al., 2016).
- DRCN: Image SR using deeply-recursive convolutional networks (Kim, Lee, & Lee 2015c).

To evaluate the accuracy of SR, we used the peak signal-to-noise ratio *PSNR* (Yang et al., 2014) as shown in Table 1. PSNR has been used as a fidelity measurement. It is the ratio between the maximum possible power of an original signal and the power of corrupting noise that affects the fidelity of its representation. Based on the results, we can state that one of the models that provide better results is the one called *DRCN* followed by *VDSR*. However, we must take into account the processing time required by each of them. For this purpose, Table 2 shows the times required to evaluate the data set named as *Set 5*, consisting of five images (baby, bird, butterfly, head and woman) commonly used for testing performance of image super-resolution models. In the context of vehicle detection in road sequences, super-resolution response and object detection must be performed in real-time. Hence, despite the good results provided by the models DRCN and *VDSR* it is not possible to apply them to the context presented in this article. In view of this, the model named *FSRCNN* has been chosen. The proposal is not excluding for these models. According to the context in which the proposal is to be applied, it will be convenient to apply *SR* models independently of time, to obtain a super-resolved image with the highest possible image quality, thus obtaining a higher detection rate. Figure 4 shows the reconstructed high-resolution *HR* image of 'frame 1' from sequence 1, when the up-sampled scale is x2 from the various super-resolution models established for their study.

Subsequently, to improve the detections by the network, noise elimination is performed. This process $\mathcal{D}$ is performed through the use of the non-local denoising algorithm[1] making use of various computational optimizations (Buades et al., 2011). This method is applied when Gaussian white noise is expected (point 3 of the Figure 3). After an empirical study, it was concluded that this type of processing improves the number of detected elements.

Then, for each detection in S_{LR}, a sub-image $\mathbf{X}_i$ with the same size as $\mathbf{X}_{\mathrm{LR}}$ is extracted from $\mathbf{X}_{\mathrm{HR}}$. The sub-image $\mathbf{X}_i$ is centred at the centre of the detection:

$$\mathbf{y}_i = \left(\frac{a_i + c_i}{2}, \frac{b_i + d_i}{2} \right) \tag{6}$$

TABLE 1 Average PSNRs for scale factors of x2, x3, and x4 on Sequence 1, Sequence 2 and Sequence 3 (higher is better)

	Sequence 1			Sequence 2			Sequence 3		
	x2	x3	x4	x2	x3	x4	x2	x3	x4
Bicubic	30.42	27.64	26.72	31.27	28.09	26.90	27.65	24.72	23.90
SRCNN (Dong et al., 2015)	32.57	30.12	28.58	34.06	30.86	28.82	30.45	27.42	25.84
FSRCNN (Dong et al., 2016)	33.15	**31.05**	28.79	34.73	**31.88**	29.17	31.43	28.33	26.15
VDSR (Kim et al., 2016)	33.25	30.72	28.98	31.89	31.40	**29.37**	31.47	**28.73**	**26.42**
DRCN (Kim et al. 2015c)	**34.22**	30.61	**29.15**	**35.80**	31.10	29.33	**32.56**	27.72	26.31

Note: The best results are marked in bold.

TABLE 2 Average time for scale factors of x2, x3, and x4 on Dataset Set 5 (lower is better)

	Set 5		
	x2	x3	x4
Bicubic	**0.028**	**0.041**	**0.065**
SRCNN	2.190	2.230	2.190
FSRCNN	*0.068*	*0.027*	*0.018*
VDSR	0.130	0.130	0.120
DRCN	1.540	1.550	1.540

Note: The best results are marked in bold and italic indicates the second best performance.

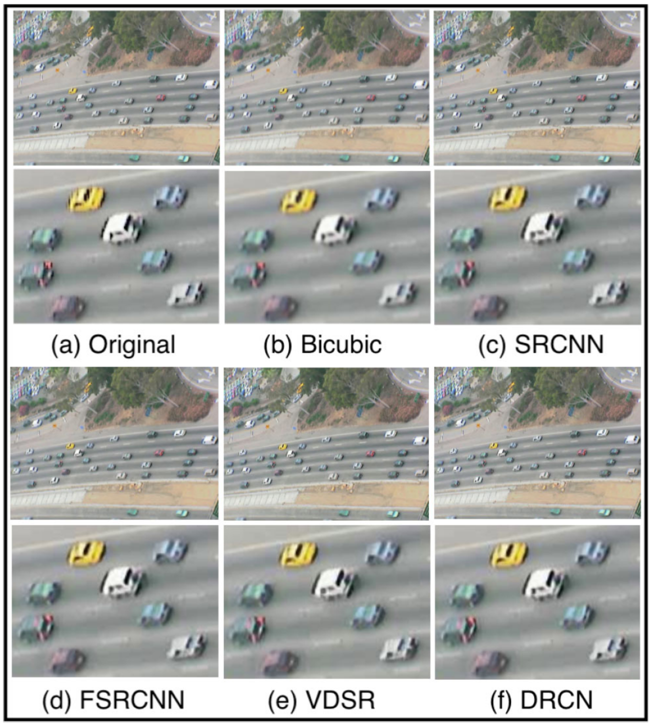

FIGURE 4 Visual evaluation for a scale factor of x2 on the 'frame 1' image from Sequence 1. From left to right: Original, Bicubic, SRCNN, FSRCNN, VDSR and DRCN

$$\hat{\mathbf{y}}_i = Z\mathbf{y}_i \tag{7}$$

where $\mathbf{y}_i$ is the centre of $\mathbf{X}_i$ expressed in coordinates of $\mathbf{X}_{\mathrm{LR}}$, while $\hat{\mathbf{y}}_i$ is the centre of $\mathbf{X}_i$ expressed in coordinates of $\mathbf{X}_{\mathrm{HR}}$. The object detection network is applied to $\mathbf{X}_i$ in order to yield a new list of detections:

$$S_i = \mathscr{F}(\mathbf{X}_i) \tag{8}$$

$$S_i = \left\{ \left(\widetilde{a}_{i,j}, \widetilde{b}_{i,j}, \widetilde{c}_{i,j}, \widetilde{d}_{i,j}, \widetilde{q}_{i,j}, \widetilde{r}_{i,j} \right) \mid j \in \{1, \ldots, N_i\} \right\} \tag{9}$$

where N_i is the number of detections for sub-image $\mathbf{X}_i$.

As indicated in step 5 of the workflow, each of the generated sub-images $\mathbf{X}_i$ is passed on to the object detection model to improve the inference or detect elements not initially spotted.

The detections of S_i are expressed in coordinates of $\mathbf{X}_i$, so that they must be translated to coordinates of $\mathbf{X}_{\mathrm{LR}}$. The equation to translate a point $\widetilde{\mathbf{h}}$ in coordinates of $\mathbf{X}_i$ to coordinates $\mathbf{h}$ of $\mathbf{X}_{\mathrm{LR}}$ is:

$$\mathbf{h} = \mathbf{y}_i + \frac{1}{Z}\widetilde{\mathbf{h}} \tag{10}$$

Therefore, the list of detections for sub-image $\mathbf{X}_i$ expressed in coordinates of $\mathbf{X}_{\mathrm{LR}}$ is:

$$S_i = \left\{ \left(a_{i,j}, b_{i,j}, c_{i,j}, d_{i,j}, q_{i,j}, r_{i,j} \right) \mid j \in \{1, \ldots, N_i\} \right\} \tag{11}$$

$$\left(a_{i,j}, b_{i,j}\right) = \mathbf{y}_i + \frac{1}{Z}\left(\widetilde{a}_{i,j}, \widetilde{b}_{i,j}\right) \tag{12}$$

$$\left(c_{i,j}, d_{i,j}\right) = \mathbf{y}_i + \frac{1}{Z}\left(\widetilde{c}_{i,j}, \widetilde{d}_{i,j}\right) \tag{13}$$

$$q_{i,j} = \widetilde{q}_{i,j} \tag{14}$$

$$r_{i,j} = \widetilde{r}_{i,j} \tag{15}$$

At this point, the coordinates of the detected elements will correspond to the vehicles composing the initial image. For each initial detection *i* established by the pre-trained *CNN*, a new super-resolved and processed sub-image $\mathbf{X}_i$ is generated on which to infer. The sub-image $\mathbf{X}_i$ may contain a set of vehicles, resulting in multiple detections being established for the same element, see Figure 5.

After that, all the detections coming from the detection set S_i are processed to estimate whether they correspond to new objects, or they match some detection already present in S_{LR}. To this end, the intersection over union (IOU) measure is computed for each pair of detections D_j and D_k:

$$\mathrm{IOU} = \frac{\mathrm{Area}\left(D_j \cap D_k\right)}{\mathrm{Area}\left(D_j \cup D_k\right)} \tag{16}$$

The detections D_j and D_k are judged to correspond to the same object whenever $IOU > \theta$, where θ is a tunable parameter. The final set of detections is comprised of those detections that persist after filtering the matching detections. This point finally corresponds to step 6 of the workflow.

In Figure 6, it can be seen how effectively, after performing the operation known as IOU, the multiple detections established for the same element have been filtered, selecting the one with the highest score as the final solution. At the end of this process, an image with a higher number of detections and score inference of each element will be obtained.

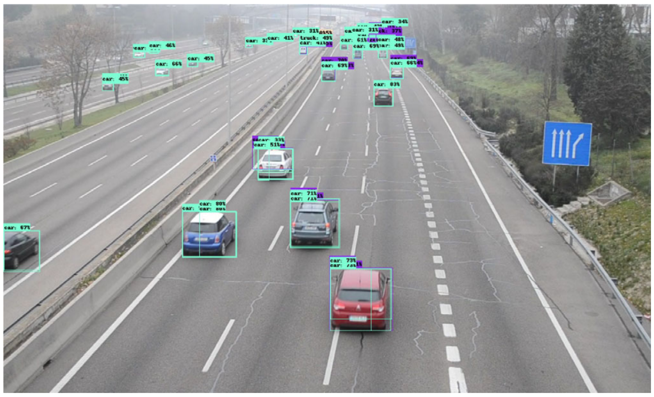

FIGURE 5 Video frame before filtering matching detections

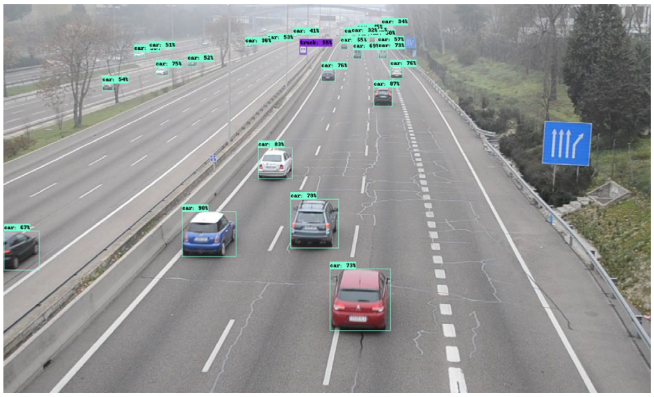

FIGURE 6 Final result after filtering matching detections

TABLE 3 Selected values of the hyperparameters

Parameter	Value
Maximum number of detections	100
Minimum percentage of inference	0.3
IOU threshold θ	0.1

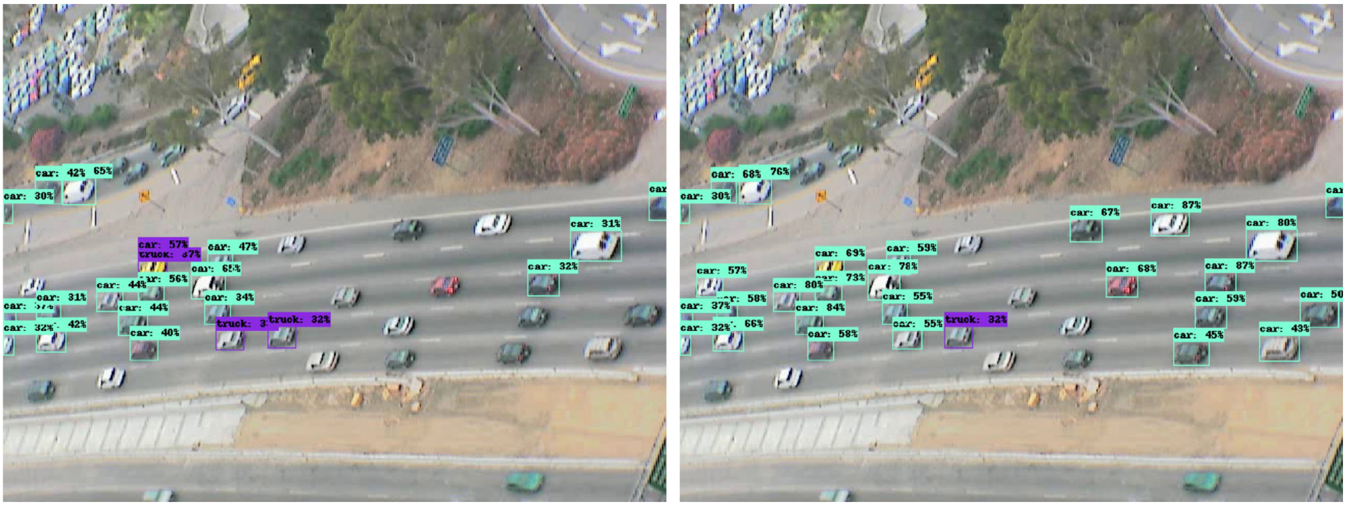

FIGURE 7 Frame 1 of the first video denoted as sb-camera1-0820am-0835am. The left side shows the results obtained by the unmodified model while the right side shows the detections after applying our proposal

4 | EXPERIMENTAL RESULTS

In order to carry out the tests on the data set designed for the experiments, it is necessary to establish a series of parameters such as the maximum number of detected objects, the minimum inference to take into account an object as well as the threshold θ for the IOU index to eliminate simultaneous detections for the same object. These parameters are established in Table 3 and selected according to the best results obtained after an empirical study.

Regarding the dataset used for the tests, as stated Section 1, three videos from video surveillance systems were manually labelled, all of which pointed to roads with a large number of vehicles, video 1 is named as *sb-camera1-0820am-0835am*, video 2 as *sb-camera3-0820am-0835am* and video 3 as *sb-camera4-0820am-0835am*. As can be seen in the name of each of the videos, they are formed by the number of the camera referred to as well as the time that the images were acquired. This dataset consists of 476 images with a total of 14,557 detections. Four specific categories related to the detection of vehicles on roads have been selected. Namely, the labels *car*, *truck*, *motorcycle* and *bus* have been chosen to analyse the results. The dataset is composed of a series of images with specific interesting characteristics for the study of the obtained results, mainly highlighting the following aspects:

- The vehicles set in each frame occupy a small area of the image, thus qualifying them as small size elements. In datasets such as *KITTI*, we have a large number of vehicles. However, the size of the vehicles is much larger compared to the dataset manually developed for this article.
- There is an imbalance of classes. There are more abundant categories of elements such as cars versus motorcycles which appear in a limited number of frames.

Next, a series of examples are shown in Figures 7–12 in which the qualitative results obtained with the proposed technique are shown for some instances of the considered dataset.

To check the results obtained by the proposed technique, in addition to the images coming to the selected videos, the technique has been tested on a series of previously available ground truth frames from other benchmark videos as can be seen in Figures 10–12. To obtain quantitative performance results, the evaluation process developed by *COCO*[2] has been used. The standard measure for object detection tasks is the mean average precision (mAP), which takes into account the bounding box prediction and the class inference for each object. Thus, the bounding box of a prediction is considered a correct detection depending on the overlap with the ground truth bounding box, as measured by the

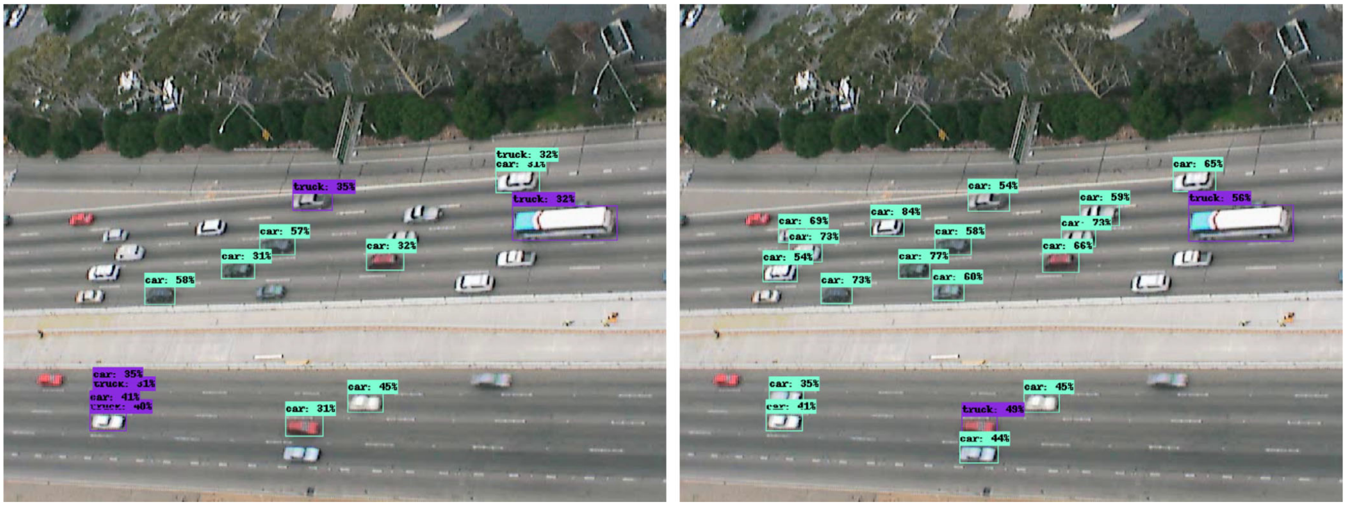

FIGURE 8 Frame 6 of the second video denoted as sb-camera3-0820am-0835am. The left side shows the results obtained by the unmodified model while the right side shows the detections after applying our proposal using CenterNetHourGlass104Keypoints

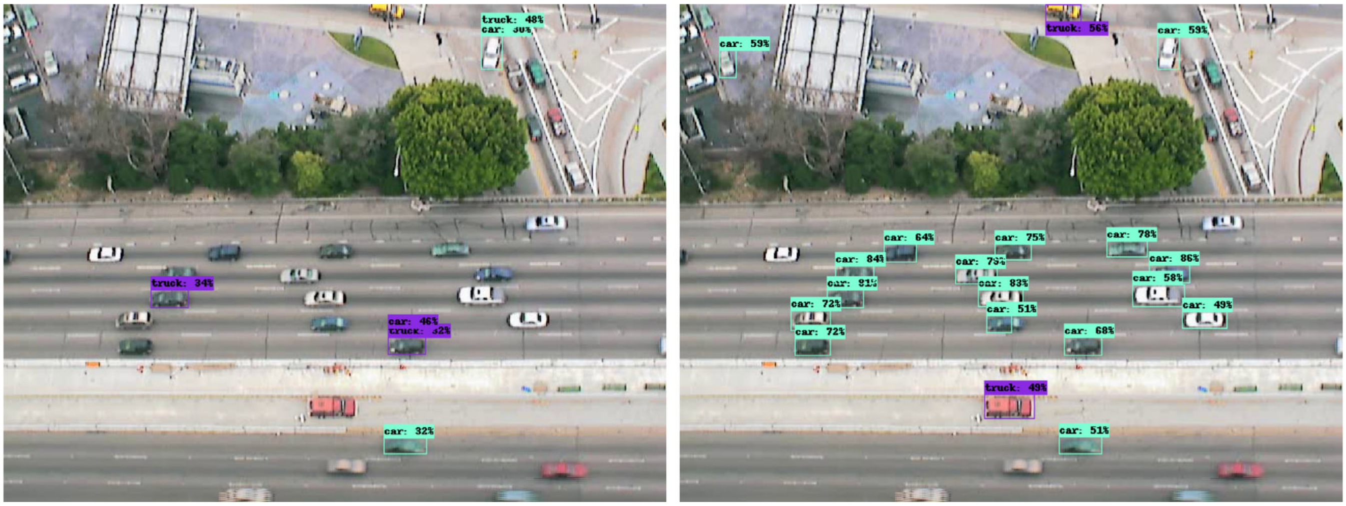

FIGURE 9 Frame 136 of the third video denoted as sb-camera4-0820am-0835am. The left side shows the results obtained by the unmodified model while the right side shows the detections after applying our proposal using CenterNetHourGlass104Keypoints

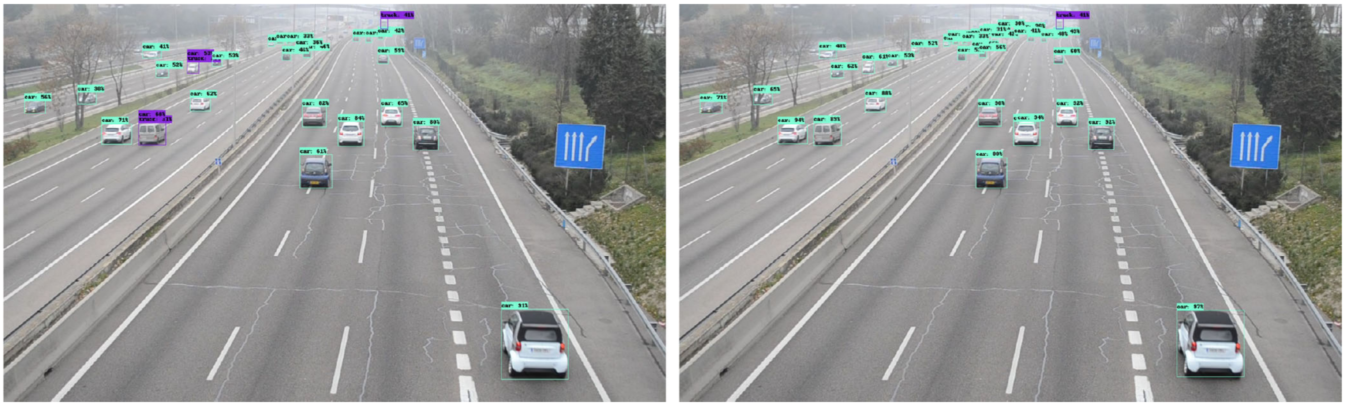

FIGURE 10 Frame 5837 of the M-30-HD dataset (Guerrero-Gomez-Olmedo et al., 2013). The left side shows the results obtained by the unmodified model while the right side shows the detections after applying our proposal using CenterNetHourGlass104Keypoints

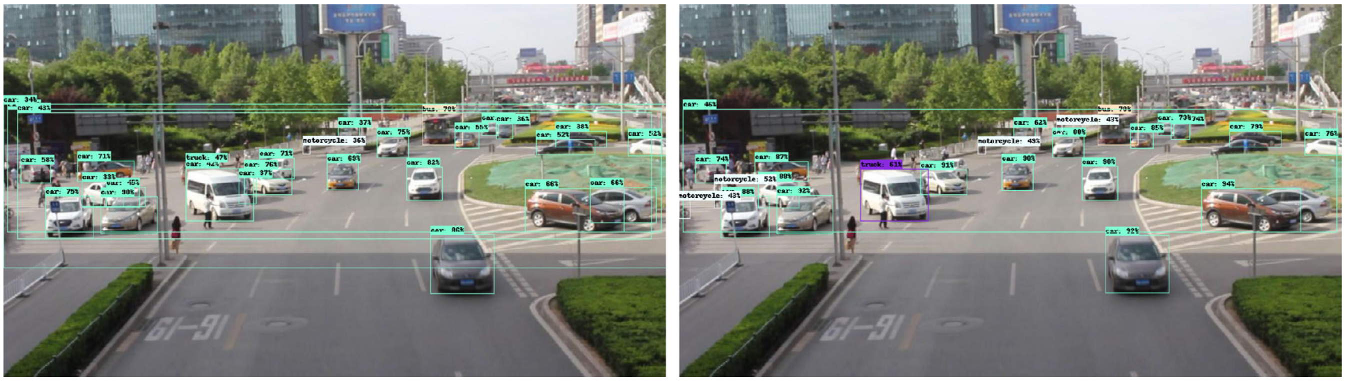

FIGURE 11 Frame 2 of the UA-DETRAC MVI_39311 dataset (Lyu et al., 2017, 2018; Wen et al., 2020). The left side shows the results obtained by the unmodified model while the right side shows the detections after applying our proposal using CenterNetHourGlass104Keypoints

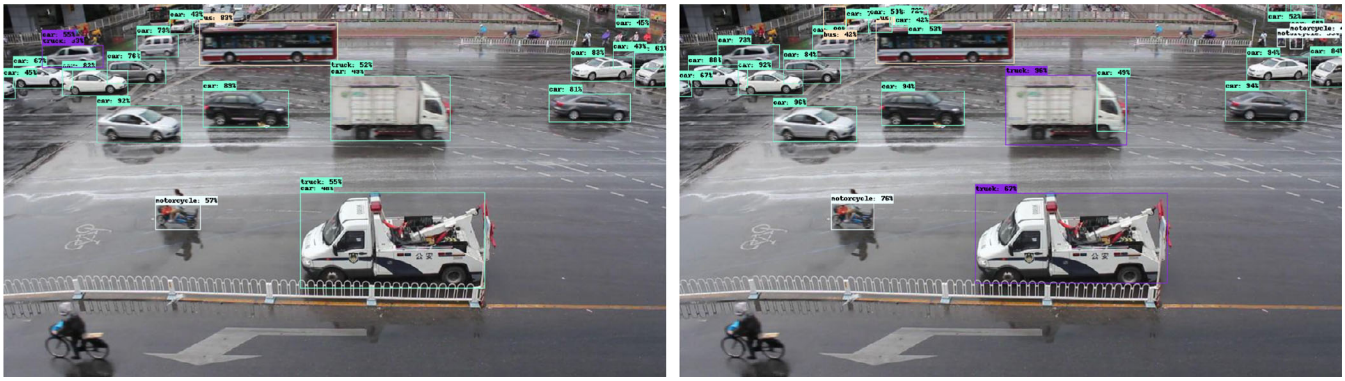

FIGURE 12 Frame 1 of the UA-DETRAC MVI_40903 dataset (Lyu et al., 2017, 2018; Wen et al., 2020). The left side shows the results obtained by the unmodified model while the right side shows the detections after applying our proposal using CenterNetHourGlass104Keypoints

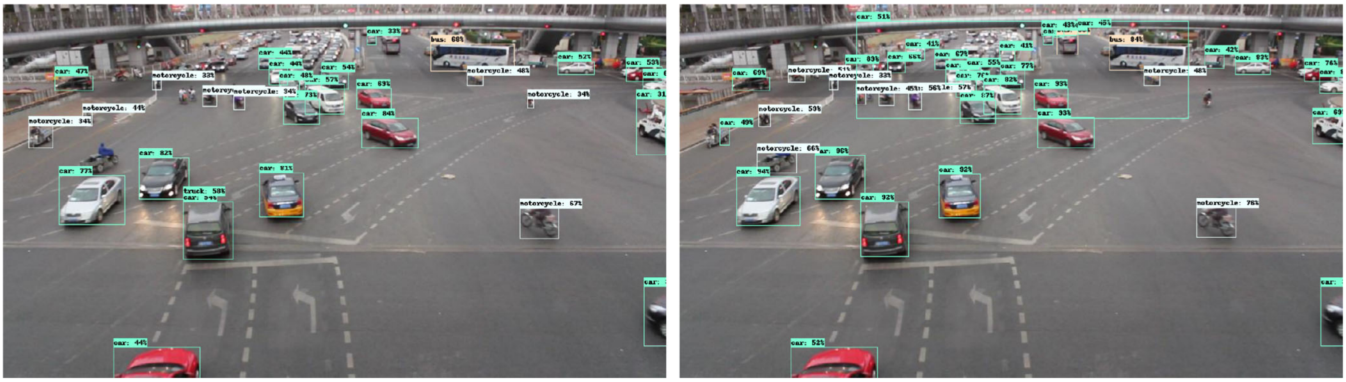

FIGURE 13 Frame 5 of the UA-DETRAC MVI_40852 dataset (Lyu et al., 2017, 2018, Wen et al., 2020). The left side shows the results obtained by the unmodified model while the right side shows the detections after applying our proposal using CenterNetHourGlass104Keypoints

intersection over the union known as *IoU*. In addition, the class inference must be correct for that element. As output, several mAPs measures are obtained, divided by the object size and the quality of the inference class. According to the tests carried out, our technique solves a series of problems depending on the context of the image. Firstly, it detects many elements that were not detected a priori by the model in the first pass. Secondly, in the initial inference, we can observe in certain areas that the class of the elements is not correct, and after applying our technique, this problem is solved.

To test our proposal, three video sequences have been selected from *US Highway 101 Dataset* [https://www.fhwa.dot.gov/publications/research/operations/07030/index.cfm]. These videos have been captured by highway video surveillance systems from a high perspective and belong to the U.S. Department of Transportation, see Figures 7–9. In order to demonstrate the present improvement through the use of this proposal, we have also tested a series of frames from the dataset named as *M-30-HD* presented by Guerrero-Gomez-Olmedo et al. (2013), see Figure 10, as well as the dataset *UA-DETRAC-MVI* submitted by Wen et al. (2020), thus obtaining Figures 11–13. We have made use of the datasets mentioned above, since they are captured by video surveillance systems, capturing a large number of vehicles on urban roads and state highways. On the other hand, these video surveillance systems are placed in high points, which capture elements of various sizes, including small-scale ones, to evaluate the increase in the number of vehicles as well as the score of each of these using the proposal.

However, it should be noted that our solution is not infallible, and its performance depends on the context in which it is applied. For example, in Figure 13 we can see how the number of elements is increased. However, we also obtain false positives such as the one in the centre of the image or the incorrect inference of the motorcycle located at the centre-left side. This problem can be alleviated by increasing the minimum percentage of inference to take into account a detection.

To carry out the tests, several variants of either the *EfficientNet* or *CenterNet* models have been used. For *EfficientNet*, some variants have been defined, which range from zero to seven, each representing variants for efficiency and accuracy according to a given scale. Thanks to this scaling heuristic, it is possible for the base model denoted as *D0* to outperform the models at each scale, also avoiding an extensive search in the hyperparameter space. At first, one may get the impression that *EfficientNet* constitutes a continuous family of models defined by an arbitrary choice of scale; however, other aspects such as resolution, width and depth are contemplated. According to resolution, models with low scale such as *D0* or *D1* apply zero-padding near the boundaries of specific layers, thereby wasting computational resources. As for depth and width, the size of the channels must be a multiple of eight. For *CenterNet*, three variants have been used, each of them with a defined architecture, featuring as main differences the size of the image given as input as well as the way of processing each of the detections. It has been tested with variants based on keypoints as well as traditional models. These pre-trained models have been extracted from Tensorflow Model Zoo.[3] It is important to note that no training process using traffic sequences is performed on these pre-trained models.

The performance of these initial models and the proposed methodology is quantitatively compared. Since the *car* class is majority, the performance in the detection of cars is considered as the measure to compare. For this purpose, only class number 3 has been considered, corresponding in *COCO* to the *car* element, thus obtaining Tables 4–6. In each of these tables, the mAP is determined for both the initial

TABLE 4 Results obtained for the first video

Video 1—sb-camera1-0820am-0835am					
Model	**IoU = 0.50:0.95 |area = all**	**IoU > 0.50 |area = all**	**IoU > 0.75 |area = all**	**IoU = 0.50:0.95 |area = small**	**IoU > 0.50| area = medium**
CenterNet HourGlass104 Keypoints	0.244/**0.361**	0.420/**0.628**	0.263/**0.381**	0.258/**0.375**	0.193/**0.283**
CenterNet MobileNetV2 FPN Keypoints	0.083/**0.161**	0.191/**0.358**	0.053/**0.121**	0.085/**0.163**	0.081/**0.122**
CenterNet Resnet101 V1 FPN	0.273/**0.350**	0.512/**0.656**	0.248/**0.318**	0.274/**0.355**	0.244/**0.273**
EfficientDet D3	0.159/**0.336**	0.276/**0.585**	0.169/**0.364**	0.159/**0.343**	0.154/**0.247**
EfficientDet D4	0.236/**0.442**	0.424/**0.790**	0.237/**0.459**	0.243/**0.451**	0.168/**0.298**
EfficientDet D5	0.245/**0.378**	0.433/**0.672**	0.252/**0.389**	0.251/**0.395**	0.158/**0.164**

Note: On the left is the mAP obtained by the base model and on the right of each box is the mAP obtained by our proposal (higher is better). The best results are marked in bold.

TABLE 5 Results obtained for the second video

Video 2—sb-camera3-0820am-0835am					
Model	**IoU = 0.50:0.95 |area = all**	**IoU > 0.50 |area = all**	**IoU > 0.75 |area = all**	**IoU = 0.50:0.95 |area = small**	**IoU > 0.50 |area = medium**
CenterNet HourGlass104 Keypoints	0.090/**0.186**	0.176/**0.363**	0.074/**0.154**	0.092/**0.191**	0.043/**0.112**
CenterNet MobileNetV2 FPN Keypoints	0.117/**0.158**	0.309/**0.390**	0.047/**0.083**	0.119/**0.159**	0.039/**0.085**
CenterNet Resnet101 V1 FPN	0.111/**0.188**	0.242/**0.403**	0.072/**0.124**	0.111/**0.191**	**0.096**/0.093
EfficientDet D3	0.092/**0.294**	0.177/**0.590**	0.073/**0.224**	0.091/**0.297**	0.105/**0.171**
EfficientDet D4	0.236/**0.345**	0.480/**0.689**	0.171/**0.261**	0.237/**0.350**	0.232/**0.244**
EfficientDet D5	0.178/**0.292**	0.345/**0.588**	0.145/**0.219**	0.178/**0.295**	0.207/**0.221**

Note: On the left is the mAP obtained by the base model and on the right of each box is the mAP obtained by our proposal (higher is better). The best results are marked in bold.

TABLE 6 Results obtained for the third video

Video 3—sb-camera4-0820am-0835am					
Model	**IoU = 0.50:0.95 |area = all**	**IoU > 0.50 |area = all**	**IoU > 0.75 |area = all**	**IoU = 0.50:0.95 |area = small**	**IoU > 0.50 |area = medium**
CenterNet HourGlass104 Keypoints	0.074/**0.219**	0.137/**0.394**	0.072/**0.217**	0.075/**0.227**	0.054/**0.060**
CenterNet MobileNetV2 FPN Keypoints	0.060/**0.102**	0.159/**0.237**	0.026/**0.064**	0.066/**0.108**	0.009/**0.014**
CenterNet Resnet101 V1 FPN	0.039/**0.082**	0.075/**0.160**	0.028/**0.065**	0.040/**0.084**	**0.016**/0.015
EfficientDet D3	0.029/**0.163**	0.050/**0.292**	0.022/**0.164**	0.029/**0.168**	0.012/**0.040**
EfficientDet D4	0.129/**0.280**	0.236/**0.518**	0.118/**0.264**	0.130/**0.287**	0.068/**0.151**
EfficientDet D5	0.056/**0.204**	0.099/**0.384**	0.051/**0.184**	0.058/**0.207**	0.006/**0.082**

Note: On the left is the mAP obtained by the base model and on the right of each box is the mAP obtained by performing our proposal (higher is better). The best results are marked in bold.

detections as well as the results obtained after applying our proposal. Columns second to fourth refer to detections of objects of any size. The second column corresponds to the mAP of the elements detected with an *IoU* between 50% and 90%. The third column reports detections above 50% of *IoU*, while the fourth column corresponds to the detections above 75% of this metric. Additionally, results are presented for small-sized objects with an *IoU* between 50% and 95% and medium-sized items with an *IoU* above 50%, in the fifth and sixth columns, respectively.

Metrics obtained for large-sized objects have been omitted since there are no samples in the established dataset. As shown in each of the tables obtained for the images that make up the defined dataset, there is a clear improvement in the mAP measure.

To better illustrate the increase in the number of detections, the *CenterNet HourGlass104 Keypoints* model has been selected for this test as it is one of the best performing models compared to the rest of the models used for this experiment. For each of the sets of images that make up the videos from which the dataset has been made, the improvement obtained in the number of elements by the described technique compared to the results initially provided by the model has been graphically represented in Figures 14–16.

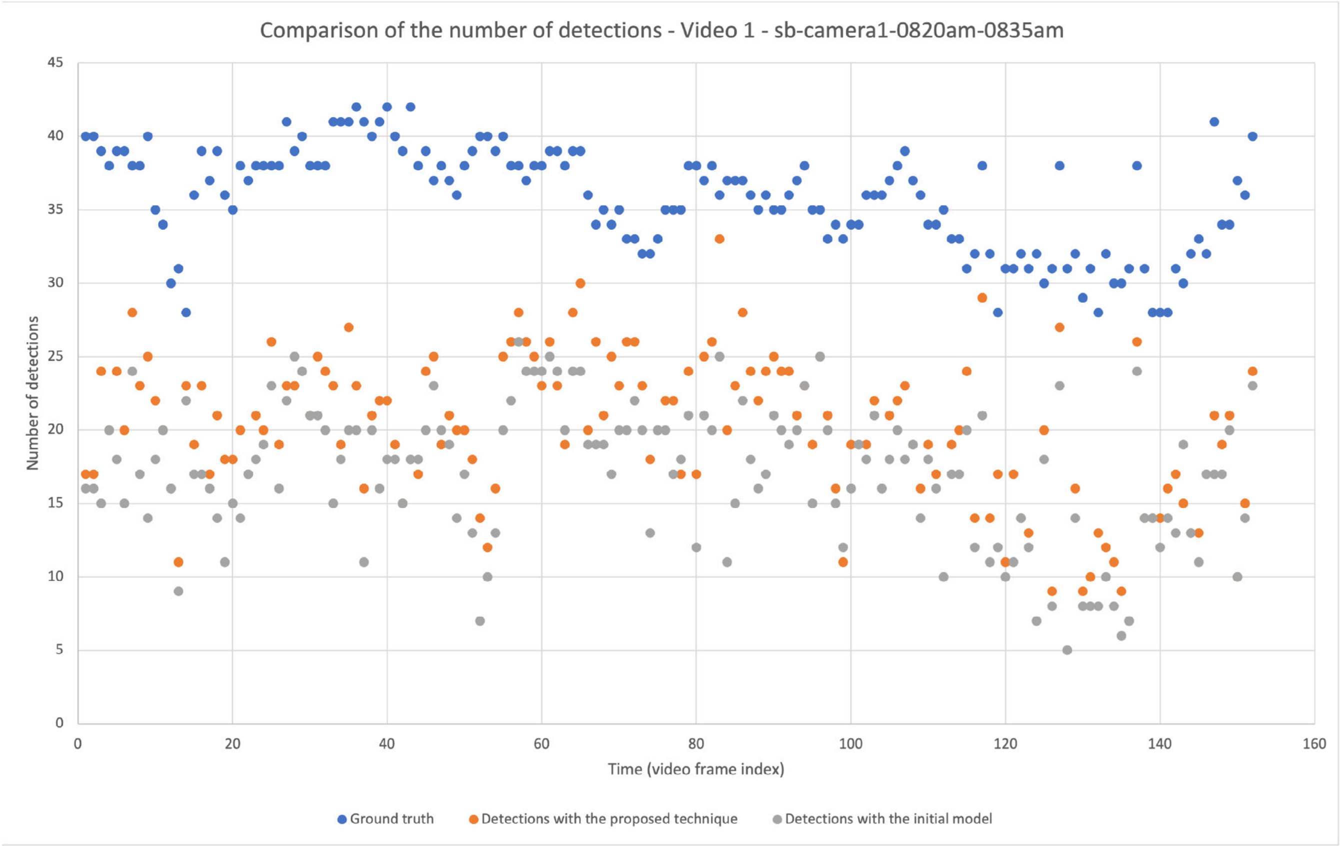

FIGURE 14 Comparison of the number of detections for the first sequence

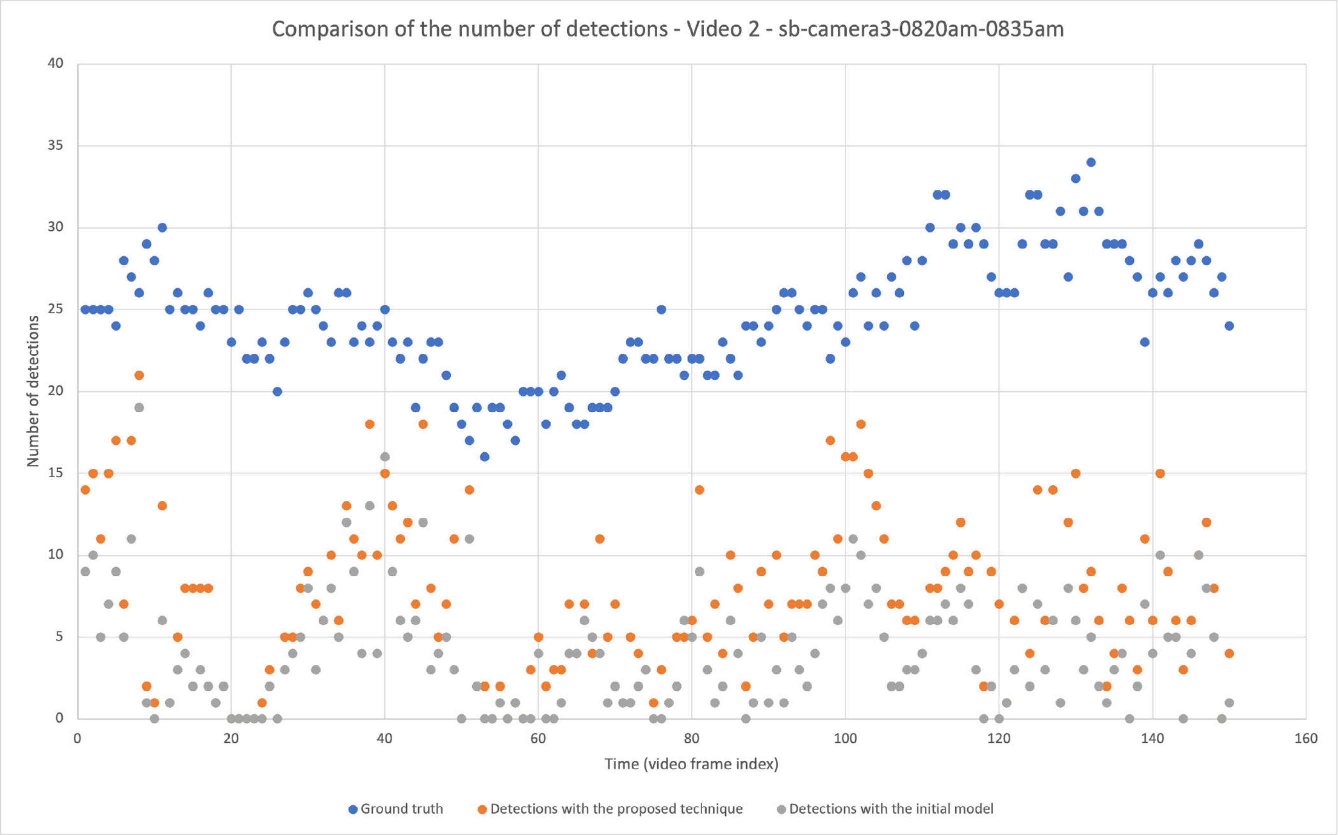

FIGURE 15 Comparison of the number of detections for the second sequence

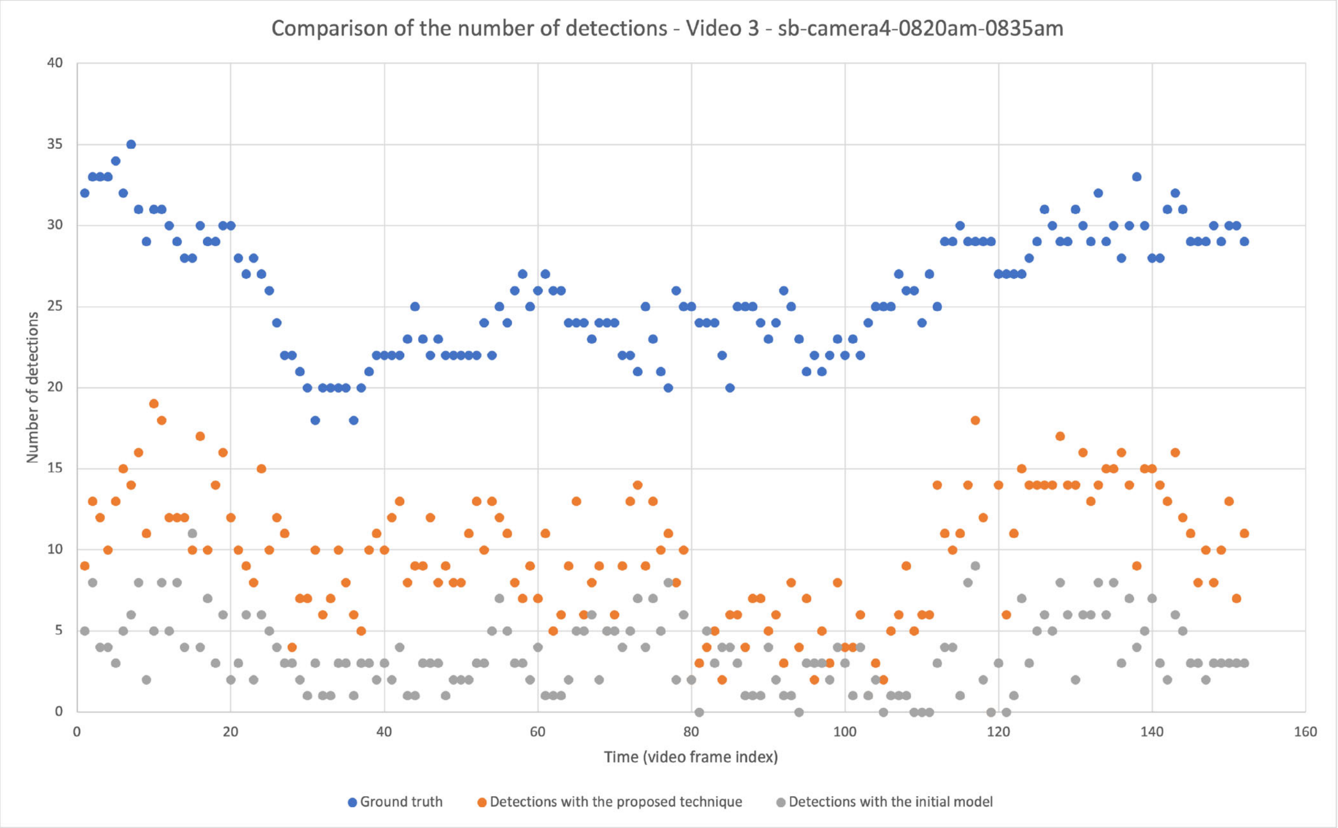

FIGURE 16 Comparison of the number of detections for the third sequence

According to Tables 4–6, the application of the technique described in this work not only improves the accuracy of the elements initially detected by the model but also detects objects that were not identified a priori. As can be seen in Figures 14–16, the proposed technique results in an increase in the number of detected elements. In blue, the Ground Truth of the vehicles that make up the selected sequences from the dataset called *U.S. Highway 101 Dataset*, shown in blue, are annotated. In these figures, we can see how the number of vehicles detected by the pre-trained model initially, represented in grey colour, is low because the model does not detect vehicles due to the small size they present and the conditions of the scene. However, from the results obtained by our proposal, represented in orange colour, we can say that indeed, by applying super-resolution in conjunction with the processing of the initial image given as input and generate the various sub-images on which the model will subsequently infer the result improves, thereby detecting a greater number of vehicles. This fact, as previously mentioned, is due to the increase in the number of pixels that make up the vehicles located in the scene, leading to better results.

5 | CONCLUSIONS

Throughout this paper, a new technique has been developed which employs deep convolutional neural networks for super-resolution to detect small-scale visual elements and improve the reliability of the estimation of their classes. The method makes a first pass of the object detection deep neural network. Then it super-resolves the video frame. After that, a window is defined for each detected object, which has the detected object at its centre, and an object detection is carried out for that window. This object detection results in more accurate estimations of the characteristics of the previously detected objects, as well as the detection of additional objects that were missed by the object detection network in the first pass.

A quantitative comparison of our approach has been carried out. The results obtained show that the application of super-resolution processes does improve the detections made on an image since, based on the metrics obtained using the *COCO* evaluator. According to the diverse object detection models used, we can highlight the increase in the *MAP* obtained after applying the proposal presented. Establishing for example the model *EfficientDet D3* whose *MAP* is increased by 17.6% on average, going from the accuracy of 9.1%–29.7%. This improvement also applies to medium and large elements, obtaining an overall improvement of 17.1% for this model, thus increasing the number of elements detected in the image given as input, improving also the class score inferred by the model. Other models such as *CenterNet HourGlass104 Keypoints* obtain an increase of 12.3% for small objects and 12% for elements independent of their size. The main advantage of our proposal is that it is not necessary to modify the object detection network, thus avoiding the modification of the intermediate layers that make up the model or the need for retraining.

The quantitative object detection results are enhanced. Qualitative results further confirm the advantages of our proposal since the improvement of the reliability of the object detections and the discovery of additional objects can be observed.

Let us remember that, in all cases, the images that make up the model on which the results have been validated correspond to video surveillance systems placed in high points. Regarding future lines to extend the research reported in this article, it should be noted that a novel approach is envisaged on which work is already underway. It consists of reducing the number of windows that need to be processed during the object detection. This enhancement might result in a reduction of the computational load associated with the execution of our method.

Another approach that is being pursued is based on a proposed methodology to automatically adapt an object detection method to a specific scene with small objects with no human intervention. The super-resolution strategy presented in this paper is first applied to find detection examples from the scene even if the object detection method is not able to detect them from the original image. The new examples are then used as training data in order to perform a fine-tuning process from a general-purpose pre-trained object detection method to improve its performance on that specific scene.

ACKNOWLEDGEMENTS

This work is partially supported by the Ministry of Science, Innovation and Universities of Spain [grant number RTI2018-094645-B-I00], project name Automated detection with low-cost hardware of unusual activities in video sequences. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project UMA18-FEDERJA-084, project name Detection of anomalous behaviour agents by deep learning in low-cost video surveillance intelligent systems. All of them include funds from the European Regional Development Fund (ERDF). It is also partially supported by the University of Málaga (Spain) under grants B1-2019_01, project name Anomaly detection on roads by moving cameras, and B1-2019_02, project name Self-Organizing Neural Systems for Non-Stationary Environments. The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. The authors acknowledge the funding from the Universidad de Málaga. I.G.-A. is funded by a scholarship from the Autonomous Government of Andalusia (Spain) under the Young Employment operative program [grant number SNGJ5Y6-15]. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X GPUs. Funding for open access charge: Universidad de Málaga / CBUA.

CONFLICT OF INTEREST

No conflict of interest has been declared by the authors.

ENDNOTES

[1] http://www.ipol.im/pub/algo/bcm_non_local_means_denoising/.

[2] https://github.com/cocodataset/cocoapi/blob/master/PythonAPI/pycocotools/cocoeval.py.

[3] https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Iván García-Aguilar https://orcid.org/0000-0001-5476-6704
Rafael Marcos Luque-Baena https://orcid.org/0000-0001-5536-1805
Ezequiel López-Rubio https://orcid.org/0000-0001-8231-5687

REFERENCES

Bochkovskiy, A., Wang, C., & Liao, H. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. CoRR, abs/2004.10934. https://arxiv.org/abs/2004.10934

Buades, A., Coll, B., & Morel, J.-M. (2011). Non-local means denoising. *Image Processing On Line*, *1*, 208–212. https://doi.org/10.5201/ipol.2011.bcm_nlm

Cao, F., & Chen, B. (2019). New architecture of deep recursive convolution networks for super-resolution. *Knowledge-Based Systems*, *178*, 98–110. https://doi.org/10.1016/j.knosys.2019.04.021

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision—Eccv 2014* (pp. 184–199). Springer International Publishing.

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks.

Dong, C., Loy, C. C., & Tang, X. (2016). Accelerating the super-resolution convolutional neural network. CoRR, abs/1608.00367. http://arxiv.org/abs/1608.00367

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). CenterNet: Keypoint triplets for object detection. CoRR, abs/1904.08189. http://arxiv.org/abs/1904.08189

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR).*, *32*, 1231–1237.

Guerrero-Gomez-Olmedo, R., Lopez-Sastre, R. J., Maldonado-Bascon, S. & Fernandez-Caballero, A. (2013). Vehicle tracking by simultaneous detection and viewpoint estimation. In IWINAC 2013, Part II, LNCS 7931 (pp. 306–316).

Han, B., Wang, Y., Yang, Z., & Gao, X. (2020). Small-scale pedestrian detection based on deep neural network. *IEEE Transactions on Intelligent Transportation Systems*, *21*(7), 3046–3055. https://doi.org/10.1109/TITS.2019.2923752

Kim, J., Lee, J., Song, K., & Kim, Y.-S. (2019). Vehicle model recognition using SRGAN for low-resolution vehicle images. doi: https://doi.org/10.1145/3357254.3357284

Kim, J., Lee, J. K., & Lee, K. M. (2015a). Accurate image super-resolution using very deep convolutional networks. CoRR,abs/1511.04587. http://arxiv.org/abs/1511.04587

Kim, J., Lee, J. K., & Lee, K. M. (2015b). Deeply-recursive convolutional network for image super-resolution. CoRR, abs/1511.04491. http://arxiv.org/abs/1511.04491

Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks.

Lai, Y., Sun, F., & Liu, H. (2020). Small object detection base on YOLOv3 for pedestrian recognition. In 2020 5th international conference on control and robotics engineering (ICCRE) (p. 235–241). doi: https://doi.org/10.1109/ICCRE49379.2020.9096492

Lyu, S., Chang, M.-C., Du, D., Li, W., Wei, Y., & Del Coco, M. (2018). UA-DETRAC 2018: Report of AVSS2018 & IWT4S challenge on advanced traffic monitoring. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1–6).

Lyu, S., Chang, M.-C., Du, D., Wen, L., Qi, H., Li, Y. (2017). UA-DETRAC 2017: Report of AVSS2017 & IWT4S Challenge on Advanced Traffic Monitoring. In Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on (pp. 1–7).

Mostofa, M., Ferdous, S. N., Riggan, B. S., & Nasrabadi, N. M. (2020). Joint-srvdnet: Joint super resolution and vehicle detection network. CoRR, abs/2005.00983. https://arxiv.org/abs/2005.00983

Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M., Qi, H., & Lyu, S. (2020). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. Computer Vision and Image Understanding, *193*, 102907.

Xing, C., Liang, X., & Bao, Z. (2019). A small object detection solution by using super- resolution recovery. In 2019 IEEE 7th international conference on computer science and network technology (ICCSNT) (p. 313–316). doi: https://doi.org/10.1109/ICCSNT47585.2019.8962422

Yang, C.-Y., Ma, C., & Yang, M.-H. (2014). Single-image super-resolution: A benchmark. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision—ECCV 2014* (pp. 372–386). Springer International Publishing.

Zhao, Z., Zheng, P., Xu, S., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

AUTHOR BIOGRAPHIES

Iván García-Aguilar received the BSc degree in Computer Science, and the MSc in Software Engineering and Artificial Intelligence, from the University of Málaga, Málaga, Spain, in 2019 and 2020, respectively. Currently he is an Associate Researcher in the University of Málaga. His research interests include object detection using convolutional neural networks, image processing with super-resolution processes and artificial intelligence algorithms.

Rafael Marcos Luque-Baena received the MS and PhD degrees in Computer Engineering from the University of Málaga, Spain, in 2007 and 2012, respectively. He moved to Mérida, Spain in 2013, as a Lecturer at the Department of Computer Engineering in the Centro Universitario de Mérida, University of Extremadura. Currently, he has come back to the University of Málaga (2016) and has a teaching position as Associate Professor in the Department of Languages and Computer Science. He also keeps pursuing research activities in collaboration with other universities. His current research interests include visual surveillance, image/video processing, neural networks and pattern recognition.

Ezequiel López-Rubio received his MSc and PhD (honours) degrees in Computer Engineering from the University of Málaga, Spain, in 1999 and 2002, respectively. He also received his MSc in Social and Cultural Anthropology and his PhD in Philosophy of Science from the Spanish Distance Education University, Madrid, Spain, in 2010 and 2020, respectively. He joined the Department of Computer Languages and Computer Science, University of Málaga, in 2000, where he is currently a Full Professor of Computer Science and Artificial Intelligence. His technical interests are in deep learning, pattern recognition, image processing and computer vision.

How to cite this article: García-Aguilar, I., Luque-Baena, R. M., & López-Rubio, E. (2021). Improved detection of small objects in road network sequences using CNN and super resolution. *Expert Systems*, 1–17. https://doi.org/10.1111/exsy.12930