AIMS Mathematics, 9(6): 16015–16030.
DOI: 10.3934/math.2024776
Received: 06 March 2024
Revised: 18 April 2024
Accepted: 28 April 2024
Published: 07 May 2024

http://www.aimspress.com/journal/Math

Research article

AI-based outdoor moving object detection for smart city surveillance

Yahia Said[1,*] and Amjad A. Alsuwaylimi[2]

[1] Department of Electrical Engineering, College of Engineering, Northern Border University, Arar 91431, Saudi Arabia
[2] Department of Information Technology, Faculty of Computing and Information Technology, Northern Border University, Rafha 91911, Saudi Arabia

* **Correspondence:** Email: Yahia.said@nbu.edu.sa.

Abstract: One essential component of the futuristic way of living in "smart cities" is the installation of surveillance cameras. There are a wide variety of applications for surveillance cameras, including but not limited to: investigating and preventing crimes, identifying sick individuals (coronavirus), locating missing persons, and many more. In this research, we provided a system for smart city outdoor item recognition using visual data collected by security cameras. The object identification model used by the proposed outdoor system was an enhanced version of RetinaNet. A state of the art object identification model, RetinaNet boasts lightning-fast processing and pinpoint accuracy. Its primary purpose was to rectify the focal loss-based training dataset's inherent class imbalance. To make the RetinaNet better at identifying tiny objects, we increased its receptive field with custom-made convolution blocks. In addition, we adjusted the number of anchors by decreasing their scale and increasing their ratio. Using a mix of open-source datasets including BDD100K, MS COCO, and Pascal Vocab, the suggested outdoor object identification system was trained and tested. While maintaining real-time operation, the suggested system's performance has been markedly enhanced in terms of accuracy.

Keywords: moving objects detection; smart cities; deep learning; artificial intelligence
Mathematics Subject Classification: 68T07, 68T45

1. Introduction

By lowering the cost of computing resources, smart cities aim to improve and sustainably increase

the efficiency of the urban environment. A large number of cameras were placed in various locations around the urban areas for the aim of surveillance. All sorts of municipal problems, including traffic management, crime prevention, identifying diseases (coronavirus), energy consumption reduction, etc., may be better addressed with the utilization of the massive amounts of visual data collected by these cameras. The strength of smart cities' infrastructures and data centers determines how well they function. As a result, developing fast and accurate data processing algorithms is crucial. Data processing in smart cities primarily serves to identify dangers, such as people, cars, motorcycles, firearms, and a plethora of other items.

Smart cities prioritize public safety and security. Public areas, transit hubs, and important infrastructure are frequently monitored by surveillance cameras. Systems powered by deep learning can aid with the detection and monitoring of intruders, suspicious items, or illegal cars, among other possible dangers, in real time. Object detection in real conditions such as surveillance using traditional approaches is generally error-prone and requires human involvement. By automating the process of object detection, deep learning algorithms can make surveillance systems more efficient and less reliant on human oversight. This mechanization lessens the burden on human operators and allows for quicker responses to security problems. Many computer vision tasks, such as object identification and fault diagnosis [1,2], have been tremendously improved by deep learning approaches, especially convolutional neural networks (CNN). These algorithms excel at accurately recognizing moving objects in surveillance systems because they can learn intricate patterns and characteristics from raw pixel data.

Deep learning models for object detection tasks have been made easier by the availability of large-scale annotated datasets like ImageNet and COCO (common objects in context). There is a wealth of material for building reliable surveillance systems in these datasets, which include a wide variety of images and videos shot in various settings.

Modern hardware accelerators, such as tensor processing units (TPUs) and powerful graphics processing units (GPUs), have made deep learning model training and inference faster. This makes it possible to evaluate high-resolution surveillance footage in real-time or near-real-time, which speeds up the process of detecting moving objects in busy areas.

Smart city projects like urban planning, pedestrian flow analysis, and intelligent transportation systems may work in tandem with object detection systems powered by deep learning. These technologies can help with better urban administration and allocation of resources by giving real-time data regarding the movement of people and vehicles.

Deep learning models have recently become the basis of the majority of object identification and detection methods. A CNN and a region proposal method were used to create the region-based convolutional neural network (R-CNN) [3], which greatly improved the object identification model's performance. Subsequently, other object identification models were suggested using the RCNN's architecture, including the Fast R-CNN [4], the faster R-CNN [5], the region-based fully convolutional network (R-FCN) [6], and the mask R-CNN [7]. Although all of those types are incredibly accurate detectors, they are quite sluggish and have a complicated construction. Redesigning the object detection model to utilize a single network and do away with the region proposal process might lead to greater performance. To detect objects, Yolo was the first to deploy a single CNN [8]. Yolo approaches object detection as a regression issue and finds a solution. Even though it was lightning quick, the Yolo model had trouble recognizing little things and had poor detection precision overall. Subsequently, further Yolo versions [9–11] were suggested to better improve its functionality. The

single-shot multi-box detector (SSD) [12] was later suggested as a way to strike a balance between accuracy and speed. Using faster R-CNN's specified anchors on a single CNN is what the SSD does. In order to identify items at various sizes, it suggests using multi-scale feature maps. By striking a superior balance between speed and precision while operating in real-time, the SSD outperformed the Yolo. One major issue with single-network detection models is their poor accuracy due to a lack of balanced classes. In order to address this issue, RetinaNet [13] suggested reducing the cross-entropy loss function's emphasis on high-accuracy classes and increasing it on intriguing scenarios when classes have poor accuracy. “Focal loss” describes the novel loss function. In order to achieve a more balanced global accuracy throughout the whole dataset, this rescaling strategy enables the network to zero in on classes with lower accuracy. Compared to detectors built on the region proposals network, the RetinaNet's single-network detectors are just as precise. Indoor object detection [14] and traffic sign identification [15] are only two of the numerous uses for the suggested object detection models.

The importance of outdoor moving object detection to smart city surveillance systems prompted us to examine it in this work. The most significant difficulty with outdoor moving object detection is that all objects appear small due to the distance separating them from the acquisition camera. Other difficulties include complicated backgrounds, inter-class variation, intra-class variation occlusion, and the object's movement. The given work is motivated by these problems. Constructing an object detection system capable of surmounting the obstacles encountered was the primary goal.

In this paper, we present OMOD-RetinaNet, an improved version of the RetinaNet model for outdoor object recognition in smart city surveillance. In order to tackle the given issues, RetinaNet was enhanced to function better. The most notable modifications were as follows: (1) Increasing feature collection for tiny object recognition by decreasing the kernel size of the ResNet 101 model's first convolution layer [14]; (2) Expanding the detecting layers' receptive fields by using a receptive field module; (3) Modify the anchor scales so they can detect tiny items; (4) Optimizing the focus loss settings for outstanding detection performance.

It is well-known that training a CNN requires an enormous quantity of data. To that end, we suggest merging three open-source datasets to increase the quantity of training data. For outdoor moving object detection, the existing datasets include a large number of classes that are irrelevant. To train the network, we used datasets including 13 classes—people, dogs, cats, cows, sheep, horses, birds, cars, trucks, buses, bicycles, motorbikes, and trains—as positive examples of outside moving things in an urban setting. The datasets that have been suggested are the BDD100K dataset [16], the MS COCO dataset [17], and the pascal voc dataset [18]. On an Nvidia GTX960 GPU, the OMOD-RetinaNet achieved a mAP50 of 71.18% and a processing speed of 26 FPS during the test on the suggested dataset combination, proving its efficiency.

The following are the primary contributions to this paper:

- Proposing to include a feature for the identification of outside moving items in smart city surveillance systems.
- Enhancing the RetinaNet object detection model for outdoor moving object detection by adjusting the kernel size of the ResNet101 backbone’s first convolution layer to gain useful information about tiny objects, incorporating a receptive field module to expand the detection layer's receptive field, and adjusting the anchor sizes, scales, and aspect ratios to pick up on tiny objects.
- Increasing the amount of training data by merging three publicly available datasets.

The remainder of the paper is organized as follows. Section 2 is reserved for detailing and discussing related works. The proposed approach is presented with details in Section 3. In Section 4,

the experimental results are presented and discussed. Section 4 contains the conclusions and future work.

2. Related works

Researchers are drawn to generic object detection as a fascinating computer vision challenge because of its usefulness in many applications. Outdoor moving object detection is a specialized subtask of general object detection that we're dealing with here. There have been several proposals for works that aim to attain the utmost performance, but there is still room for development in this area.

Based on the SSD model, Ning et al. [19] provided an outdoor object detection model. To enhance the classification accuracy of the SSD model, inception modules were used in place of SSD layers [20]. The inception module made use of additional approaches, such as residual connections and batch normalization, to improve speed. Additionally, non-maximum weighting was used instead of non-maximum suppression. A 78.6% mAP, achieved after training and evaluation on the Pascal Voc 2007 dataset [14], demonstrated the efficacy of the suggested enhancements.

For the benefit of the visually impaired, an outside object detector was suggested in [21]. It is on the SSD model that the suggested outside object detector is built. The fundamental aim was to combine the feature fusion approach with a features pyramid structure. To train and test the suggested approach, a dataset named BLIND was also acquired. By gathering photos from varying distances, the dataset provides several sizes of the same thing, which might be useful for visually impaired individuals. In the end, this will allow for the construction of a high-performance detector with a high degree of scale invariance. The suggested dataset yielded a mAP of 75.4%, a 1.7% improvement over the performance of the first SSD model.

Wang et al. [22] suggested an object detecting approach for outside monitoring. Identifying pedestrians, a crucial target for surveillance systems, was the primary objective. The suggested technique relies on a spatial attention module cascaded with the R-CNN model. By combining static and dynamic data, the suggested module trained the network to zero down on pedestrian locations. The suggested approach was tested on two separate datasets: the open-source DukeMTMC dataset [23] and a dataset developed specifically for this study [24]. Despite extensive testing, the presented findings reveal sluggish processing time and poor accuracy, casting doubt on the efficacy of the suggested method.

To identify minute items in aerial photographs of rural areas, a CNN was suggested for use in remote sensing [25]. At its core, the proposed detector consists of three distinct phases. The initial step involves creating a landscape mask. At this point, we're using the VGG 16 model for feature extraction [26]. The second step is to develop an object detection model specifically for detecting coarse litter. At this point, we used a binary classifier to distinguish between trash and terrain. A higher sensitivity to litter than to terrain was programmed into the classifier during training. Class activation is the basis of the third stage, which consists of a convolution layer, an average pooling layer, a softmax layer, and a VGG 16 convolution block. The second stage's localization findings were fine-tuned using it. Results showed promise for the suggested method with an average accuracy of 57.5% when tested on the trushnet dataset [27].

Wu et al. [28] proposed a method that blends position-based spatial mapping with object-based feature matching to improve the accuracy and efficiency of numerous cameras' location and identification. To begin, the item of interest is mapped and matched using a uniform spatial constraint approach inside the overlapped region of several camera targets. To find a match, we look at the target

object's color characteristics. Second, we use homologous transformation to bring in the you only look once (YOLO) object identification technique, which can identify objects inside the overlapped camera region. Using the YOLO object identification technique as a foundation, a multi-camera positioning system is developed. On the test set, the YOLOv5 algorithm achieved a maximum mAP accuracy of 97.2%, according to the results. The YOLOv5 algorithm achieves a maximum mAP accuracy of 51.6% at a reasoning speed of 10 ms. The YOLOv5 algorithm's objective loss function, classification loss function, and GloU loss function have average values of 0.001, 0.01, and 0.015, respectively. Within 10 cm, the DukeMTMC-reID dataset maintains an error probability of YOLO of more than 96.5%. In the OTB dataset, the error probability of YOLO within 9.5cm is still greater than 95%. The YOLO positioning system achieves a maximum accuracy of 0.74 when the target item is in the way.

Yadav et al. [29] provided a technique that uses state of the art computer vision algorithms to recognize and track several objects in films, both in real-time and in recorded footage. This object identification and tracking pipeline combines YOLO, a high-performance convolutional neural network for object recognition, with DeepSORT, an algorithm for splitting object instances and matching detections across frames based on motion and appearance. Using DeepSORT's reliable object tracking in conjunction with YOLO's fast object detection, the study found that accurate and immediate object monitoring was achieved. Areas such as object identification, traffic control, and video surveillance stand to benefit greatly from the proposed method's use since it enhances automation and situational awareness. Future developments in computer vision and artificial intelligence may be possible thanks to the results presented here, which open the door to further research and practical use of these technologies.

A wide variety of approaches to outdoor object detection have been suggested in the aforementioned papers. Outdoor item detection in urban settings is the primary goal of this study. Overcoming the obstacles of urban space was the goal of the suggested endeavor. Additional information on the suggested method will be presented in the section that follows.

3. Proposed approach

For the purpose of detecting outside moving objects, we suggest enhancing the current RetinaNet model. On the MS COCO dataset, RetinaNet performs state of the art object identification, earning it the title of top object detector. We investigate the use of the RetinaNet for the studied task, but the achieved results were poor and do not attend the desired performance. The outdoor moving objects detection faces the following challenges:

- Interclass variation: The classes of the outdoor moving object present a high inter-class variation. For example, there is a big difference between a cat, a person, and a car. Such variation makes it hard to balance the performance of the detector across the presented classes.
- Intra-class variation: In the outdoor scene, an object of the same class can appear in different colors, positions, and sizes. As an example, we take the "person" class which presents many variations in size (young vs adult), and in position (front, side, back).
- Occlusion: An object in urban space can be partially or totally occluded (occluded by another person or another object).
- Moving object and fixed sensor: Surveillance systems are based on data collection through the mounted cameras in the urban space where most relevant objects are moving such as persons,

vehicles, and animals. Thus, it is hard to detect the object while it is changing its position continuously.

- Small objects: Taking into account the distance separating the acquisition camera from the snaped object in the surveillance system, all objects are looking small even small objects will be smaller than usual.

To build a high-performance outdoor moving object detector, it is important to handle the mentioned challenges. As a solution, we propose to improve the RetinaNet object detection model. We call the improved model OMOD-RetinaNet. The RetinaNet model was designed to solve the problem of class imbalance which is perfect for our task. Then, it still has more challenges to handle.

As a solution for the remaining problems, we start by improving the backbone. The ResNet101 model was proposed as a backbone for the OMOD-RetinaNet. This backbone and its variant ResneNet152 were the deepest neural network with 110 layers and 152 layers respectively. It was the winner of the image classification and object detection in the ILSVERC 2015. The high performance was achieved thanks to the residual block used instead of convolution layers. A residual block contains convolution layers (CONV), activation layers (RELU), and batch normalization layers (BN) with a skip connection between its input and output. The residual block used in ResNet 101 is illustrated in Figure 1.

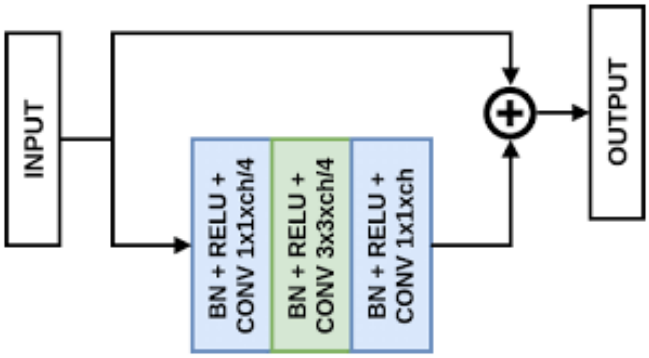

Figure 1. Residual block.

Very deep neural networks can be constructed using the suggested residual block, and their complexity will not explode. Plus, it gets rid of the vanishing gradient issue that arises during training of deep neural networks. In outdoor motion, objects tend to be little, and CNN tend to store small object information in their base layers. We update the original 7x7 kernel size of the first convolution layer with a cascade of 3 filters with a 3x3 kernel size to make ResNet101 appropriate for outdoor moving object identification. To begin, we recommend using a stride of 1, and for the subsequent filters, we recommend using a stride of 2. Although processing performance is negatively impacted, a smaller kernel size improves accuracy and enables the detection of tiny objects. The processing speed will remain unaffected, and real-time processing is still within reach, because the suggested application is built to run on high-performance devices.

Our next goal is to enhance the RetinaNet's detecting capabilities. We provide a receptive field module (RFM) as a solution to the tiny item detection problem by increasing the detection level's receptive field. The RFM was influenced by the human visual system. The object detection system's identification capabilities will be improved by expanding the receptive field. Layers of 1x1 and 3x3

convolution, 3x1 and 1x3 convolution, and 3x3 dilates convolution with varying rates make up the RFM. The suggested RFM is shown in Figure 2.

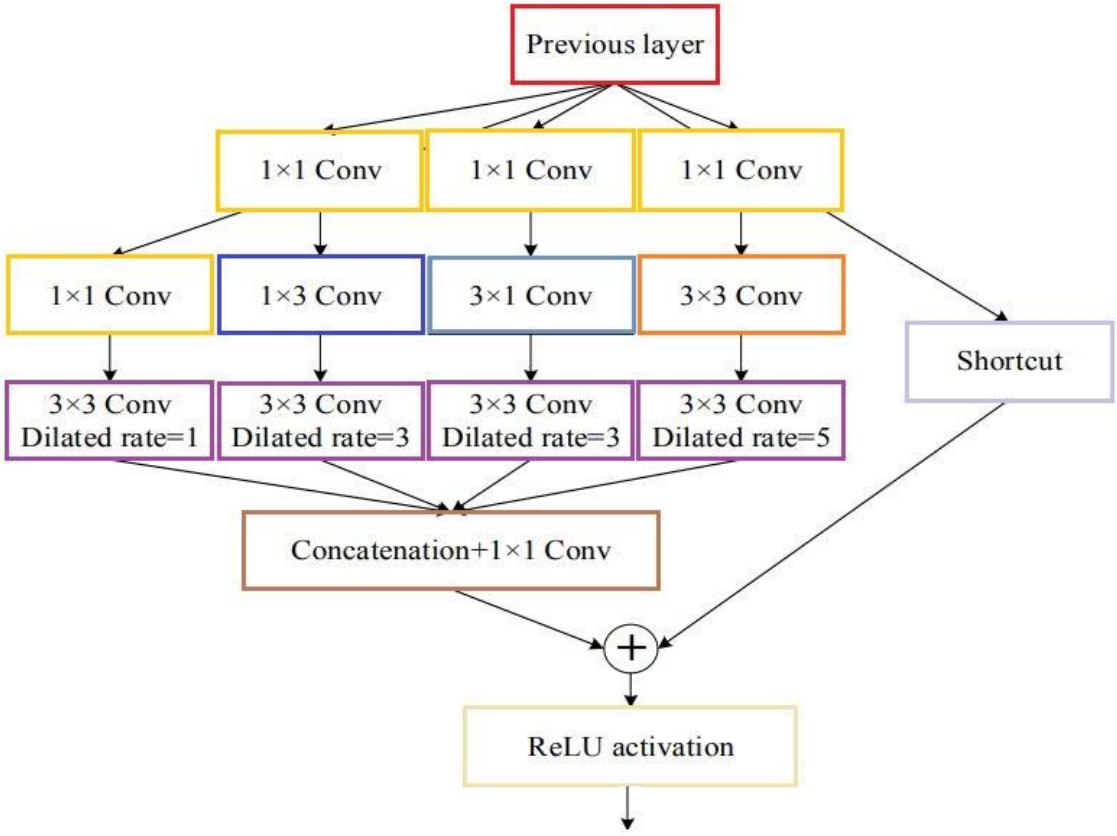

Figure 2. Proposed receptive field model.

The 1x1 convolution layers were used to reduce the input channel of the feature map. The receptive field was expanded in multiple directions using the 1x3 and 3x1 convolution layers. The dilate convolution layers were proposed by the deep lab [30], which are used to expand the receptive field and to extract multi-scale features while using small kernel sizes and without exploding the number of parameters. The dilated convolution generates large-scale feature maps with rich spatial information while maintaining high-resolution. A skip connection similar to the residual blocks proposed by ResNet was applied in the RFM. The proposed module was designed to detect small objects without increasing the computation complexity of the model.

Another optimization was applied to RetinaNet to enhance its performance for outdoor moving object detection. The anchors are the main component of the object detection system. In RetinaNet, the anchors are preselected manually. So, the anchor scales and ratios must be chosen wisely for each studied task. For outdoor moving object detection, the anchors must be suitable for small object detection. The original sizes of anchors at each layer of the 5 detection layers are {32×32, 64×64, 128×128, 256×256, 512×512}, the original three ratios are {1:2, 1:1, 2:1} and the original three scales are $\{2^0, 2^{1/3}, 2^{2/3}\}$. At each point of the feature map, RetinaNet predicts 9 anchor boxes on the basis of three scales by three aspect ratios. The prediction of bounding boxes is based on the matching

strategy which works as follows. The tested anchor is deemed to be in good agreement with the ground truth box if its intersection over union (IoU) value is higher than 0.5. A negative match is defined as an IoU value below 0.4 between the tested anchor and the ground truth box. We shall match the tested anchor to the ground truth box 1 if its IoU value is greater than the IoU value of the identical anchor and ground truth box 2. Detecting tiny objects is not a good fit for the original anchors. Therefore, we propose to change the parameters of the anchor sizes, scales, and aspect ratios. Also, we add a shallower detection layer. The new sizes of the anchors are {8×8, 16×16, 32×32, 64×64, 128×128, 256×256}, the new aspect ratios are {1:3, 1:2, 1:1, 2:1, 3:1}, and the new scales are $\{2^0, 2^{1/2}\}$. According to the new configurations, at each point of the feature map, the OMOD-RetinaNet predicts 10 bounding boxes. The main changes of the anchor scales and aspect ratios are illustrated in Figure 3.

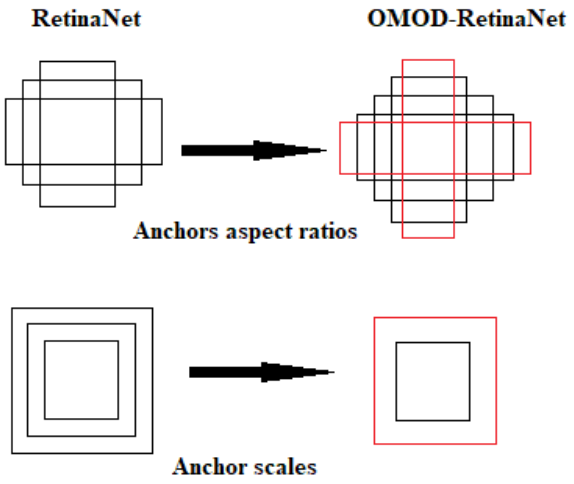

Figure 3. Anchor scales and aspect ratios modifications in OMOD-RetinaNet.

Modifying the settings of the focus loss was the last optimization done on the RetinaNet. The cross-entropy function, with minor adjustments, is the basis of the focal loss. The cross-entropy function is defined as Eq (1), where p_i is the estimated probability for a class *i*.

$$CE(p_i) = -\log(p_i). \tag{1}$$

The RetinaNet introduces the use of a coefficient α_i to balance the foreground-background classes. The α_i is defined as Eq (2),

$$\alpha_i = \begin{cases} \alpha, & \text{if y} = 1, \\ 1-\alpha, & \text{otherwise.} \end{cases} \tag{2}$$

During training, RetinaNet incorporates a modulation factor that raises the proportion of loss for difficult-to-recognize classes while lowering it for easily recognized classes, directing the detection model's attention to these more challenging classes. The modulation factor is defined as $(1-p_i)^\gamma$, where γ has a fixed value. The focal loss function is presented as Eq (3),

$$CE(p_i) = -\alpha_i(1-p_i)^\gamma \log(p_i). \tag{3}$$

The large value of γ makes the model focus more on hard to recognize classes. In the studied task, the images are characterized by a complex background. So, we increase the original value of α_i from 0.25 to 0.5 and the value of γ from 2 to 3. We shall demonstrate the effect of the suggested changes in the experimental findings. To sum up, we propose to improve the RetinaNet model for outdoor moving objects detection. The first improvement was to use the ResNet101 as a backbone and modify the kernel size of its first convolution layer to extract more relevant features to detect small objects. The second improvement was to add the RFM to increase the receptive field in the detection layers. The third improvement was to optimize the anchor scales, sizes, and aspect ratios to detect small objects. Finally, we modify the parameters of the focal loss separate the complex background from the foreground, and balance the detection precision of all classes. The architecture of the OMOD-RetinaNet is presented in Figure 4.

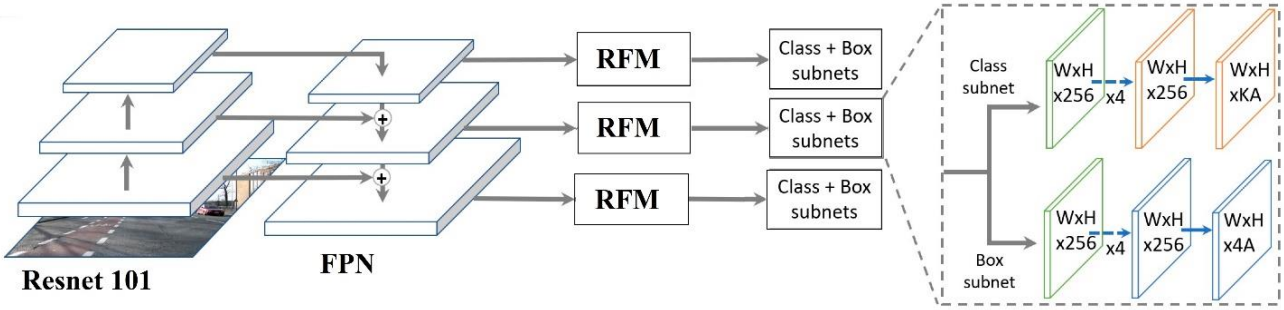

Figure 4. Proposed architecture of the OMOD-RetinaNet.

4. Experiments and results

This work's experimental environment is a desktop running Linux on top of 32 GB of RAM, an Nvidia GTX960 GPU, and an Intel i7 CPU. The OMOD-RetinaNet model that was suggested was built using the TensorFlow deep learning system. We utilized the open cv package for picture editing and visualization. As a learning algorithm, the adam optimizer was employed.

We suggest merging three open-source datasets to train the OMOD-RetinaNet model on adequate data to determine the target classes. One such dataset is the Pascal Vocabulary, which has been under development since 2005 and has undergone updates all the way up to 2012. Among its 20 offered classes, it found usage in object identification, instance segmentation, and object recognition. For the purpose of the object detection challenge, Pascal Voc 2007 and 2012 both provide picture sets. Among the 9963 photos in Pascal Voc 2007 are 24640 annotated objects, whereas among the 11530 images in Pascal Voc 2012 are 27450 annotated objects. There are a total of 21493 photos when the two databases are combined. The classes that are taken into consideration for this project are: automobile, train, bus, motorcycle, bicycle, human, bird, cat, cow, dog, horse, and sheep. The other classes are seen as negative instances. Second, Microsoft gathered the MSCOCO dataset in 2014 for a variety of uses, including object identification, instance segmentation, and keypoint estimation, among others. Eighty-one categories were included in the data collection. With over 200,000 photos, the MSCOCO 2019 collection is the biggest MSCOCO dataset to date. The MS COCO dataset utilizes identical classes as the pascal voc dataset. The BDD100K dataset, created by the artificial intelligence research team at Berkley, is the third dataset. Presented in 10 categories, the dataset was created for item recognition in

an urban setting. There are 120,000,000 pictures in the collection, produced by 100,000 films that run for 40 seconds at 30 frames per second. In order to identify moving objects in the outdoors, only the following classifications were considered: vehicle, truck, bus, bicycle, motorbike, train, and human. The suggested datasets provide a diverse variety of training pictures, which could improve the model's ability to generalize. Outdoor moving things were categorized into fourteen groups: humans, dogs, cats, cows, sheep, horses, birds, cars, trucks, buses, bicycles, motorbikes, and trains.

Our proposed model's performance evaluation metric is the mean average precision, which was originally suggested by the Pascal Vocabulary dataset and later refined by the MSCOCO dataset by including the IoU threshold. To measure how well the suggested model works, we use the mean average precision (mAP_{50}) and apply an IoU criterion of 0.5.

Because of the big amount of training data, the model was trained for 50 epochs, each epoch with 10000 steps. The model takes 2 days to be trained. By testing the model, high performance was achieved with mAP_{50} of 71.18%. Table 1 summarizes the archived average precisions (AP) per class.

Table 1 provides valuable insights into the performance of the object detection model across different object classes. The object classes "Cat" and "Person" achieve the highest AP percentages at 77.65% and 77.54%, respectively. These classes exhibit excellent model performance, indicating that the object detection model can accurately detect and localize instances of cats and persons in images. AP percentages vary across different object classes, highlighting differences in the model's ability to detect and classify different types of objects. While some classes achieve high AP percentages (e.g., "Cat", "Person", "Train"), others exhibit lower performance (e.g., "Dog", "Sheep", "Bus"). The performance of the object detection model for specific object classes has practical implications for applications such as smart city surveillance, autonomous driving, and image classification. Classes with higher AP percentages are more reliably detected by the model, making them crucial for accurate object recognition and scene understanding. Classes with lower AP percentages may indicate areas for improvement in the object detection model. For example, classes such as "Dog" and "Sheep" exhibit relatively lower performance, suggesting that the model may struggle to accurately detect these objects. Optimizing the model architecture, training data, or object detection algorithms may help improve performance for these classes. The performance of individual object classes contributes to the overall effectiveness of the object detection model. Understanding the performance characteristics of each class allows for targeted optimization efforts to enhance model performance across the board. When deploying the object detection model in real-world scenarios, it's essential to consider the performance of individual object classes and prioritize optimization efforts accordingly. Classes with higher AP percentages may require less attention, while classes with lower performance may warrant additional refinement and tuning.

Table 1. Achieved APs per class.

Class	Person	Dog	Cat	Caw	sheep	Bird	Horse	car	Truck	Bus	bicycle	motorcycle	Train
AP (%)	77.54	60.74	77.65	72.64	68.72	72.19	68.35	70.45	75.28	63.39	76.23	67.82	74.35

We compare the outcomes of our OMOD-RetinaNet with the top models for outdoor objects detection. In Table 2, we see how our model compares to top-tier models using mAP50 and the training

and testing datasets.

Table 2. Comparison against state of the art models for outdoor objects detection.

Model	mAP_{50} (%)	Train/ test data	Number of classes	Number of images
I-SSD [19]	78.6	Pascal voc 2007	20	9963
FP SSD [21]	75.4	BLIND	7	8,900
VGG+CAM [25]	57.5	trushnet	5	21,000
RetinaNet	69.56	Pascal voc+MSCOCO+BDD100K	13	349963
OMOD-RetinaNet	71.18	Pascal voc+MSCOCO+BDD100K	13	349963

The mAP_{50} indicates the overall performance of each object detection model in accurately detecting and localizing objects in images. The highest performing model is "I-SSD," achieving an mAP_{50} of 78.6%, followed by "OMOD-RetinaNet" at 71.18%. Conversely, "VGG+CAM" exhibits the lowest mAP50 at 57.5%. These variations in model performance highlight differences in the effectiveness of the object detection algorithms and architectures employed by each model.

The choice of training and testing datasets significantly impacts model performance. Models trained on diverse and comprehensive datasets, such as "RetinaNet" and "OMOD-RetinaNet," which utilize a combination of Pascal VOC, MS COCO, and BDD100K datasets, tend to achieve higher mAP50 percentages compared to models trained on smaller or less diverse datasets. For example, "VGG+CAM" trained on the "trushnet" dataset achieves a lower mAP50, indicating the importance of dataset diversity and representativeness in training robust object detection models.

The number of object classes present in the training data also influences model performance. Models trained on datasets with a larger number of classes may face greater complexity and challenges in object recognition, potentially affecting their mAP_{50}. However, the relationship between the number of classes and model performance may vary depending on the specific characteristics of the dataset and the capabilities of the object detection model.

The size of the training dataset, as indicated by the number of images, plays a crucial role in model training and generalization. Models trained on larger datasets, such as "RetinaNet" and "OMOD-RetinaNet" with 349,963 images, demonstrate higher mAP_{50} compared to models trained on smaller datasets. The availability of a large and diverse training dataset allows models to learn robust features and patterns, leading to improved performance in object detection tasks.

The findings from the table have implications for model selection and deployment in practical applications. Object detection models with higher mAP_{50} percentages, trained on diverse datasets with a large number of images and classes, are better suited for real-world scenarios requiring accurate and reliable object recognition. Consideration of these factors is essential when choosing an object detection model for specific applications such as smart city surveillance, autonomous vehicles, and image analysis.

As some state of the art models present a better performance than the OMOD-RetinaNet, there is a big difference in the amount of testing data and the number of classes. These factors directly affect the performance of the model, where bigger testing data results in decreasing the precision and fewer classes increase the precision of the model. For example, our testing data is more than 600000 images and the

pascal voc2007 is 4000 images. Also, if we compare the number of classes, our dataset presents 13 classes and the BLIND dataset presents only 7 classes.

When it comes to processing speed, the suggested model obtained 26 FPS, which is ideal for processing in real-time. When taking into account the object's velocity and the frequency of the surveillance camera, the attained speed will adhere to the requirements of real-time processing. The processing speed can be enhanced by using a higher performance platform with a better GPU.

To test the generalization power of the proposed model, we test it using new images that do not belong to the dataset which we collect from the internet. The result is shown in Figure 5. The achieved results prove that the OMOD-RetinaNet have a good generalization power to detect outdoor moving objects at a new environment with a complex background and presents many challenges such as occluded objects and intra-class variation.

Figure 5. Result of the OMOD-RetinaNet on new images.

The reported results show a big balance between class precisions. So, all the achieved precisions are in a compressed range of values. This proves the efficiency of the OMOD-RetinaNet and the impact of the proposed improvements. The modification of the parameters of the focal loss was very effective for class precision balancing. Reducing the kernel size of the ResNet101 helped to extract rich information for small object detection. The proposed anchor scales and aspect ratios have improved the detection task. Also, the proposed RFM was very important to expand the receptive field in the detection layer without increasing the computation complexity. Overall, the proposed OMOD-RetinaNet has improved the detection precision with more than 2% compared to the original RetinaNet.

An ablation study was performed to show the impact of the proposed contributions on the baseline model. For each ablation study, we maintain the proposed modification and manipulate the desired parameter. First, we investigated the impact of the proposed residual blocks which replace the 7x7 filter by 3 filters with 3x3 kernel. Table 3 summarizes the achieved results.

Table 3. Ablation study on the kernel size of the first residual block.

	mAP_{50} (%)
7x7 kernel	70.31
3x3 kernel x 3	71.18

Second, the impact of the RFM was investigated by testing the proposed model with the original setting and with the proposed RFM. Table 4 presents the achieved results. Through this ablation study, it is obvious that the proposed RFM has a wide impact on the performance of the proposed model for outdoor object detection.

Table 4. Ablation study on the RFM.

	mAP_{50} (%)
original	69.91
RFM	71.18

5. Conclusions

In the near future, smart cities will be a new lifestyle. To optimize the services of a smart city it must have an artificial intelligence system able to manipulate and process a huge amount of data daily. Most of the intelligent systems are based on surveillance data. In this work, we propose to build an automatic outdoor moving object detector. Outdoor moving objects are the main elements of the smart city for many tasks such as crime prevention, detecting diseases (coronavirus), finding lost peoples or animals, etc. The proposed outdoor moving objects detector was based on the RetinaNet object detector. RetinaNet was improved to make it suitable for outdoor moving objects detection. For the studied task, objects are small and moving. So, we improve RetinaNet in this way by optimizing the existing parts and adding other parts. We suggest starting by making the first convolution layer of the backbone use a smaller kernel. Second, in order to make it work for detecting small objects, we adjusted the anchor sizes, scales, and aspect ratios. Third, in order to obtain a proper class balance and eliminate the domination of the complex backdrop, the focal loss parameters were changed. Lastly, to improve the capability of detecting small objects, we incorporate a receptive field module. This module expands the receptive field at the detection layers without revealing the intricacy of the model. Our proposal is to merge three open-source datasets in order to train and test the OMOD-RetinNet. On the Nvidia GTX960, the suggested model's efficiency was demonstrated by an examination with a mAP_{50} of 71.18% and a processing speed of 26 FPS. The model's ability to detect objects that are heavily occluded by other objects or environmental elements may be limited. As a future work, the proposed methods will be explored for fine-grained classification of detected objects, for instance, distinguishing between different types of vehicles or identifying specific classes of objects relevant to smart city applications.

Author contributions

Yahia Said: Conceptualization, Methodology, Resources, Writing—original draft preparation, Writing—review and editing, project administration, funding acquisition; Amjad A. Alsuwaylimi: Validation, Formal analysis, Investigation, Data curation, Writing—original draft preparation, Visualization. All authors have read and agreed to the published version of the manuscript.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FPEJ-2024-3030-01".

Conflict of interest

The authors declare no conflict of interest.

References

1. X. Chen, X. Li, S. Yu, Y. Lei, N. Li, B. Yang, Dynamic vision enabled contactless cross-domain machine fault diagnosis with neuromorphic computing, *IEEE/CAA J. Automat. Sinica*, **11** (2024): 788–790. https://doi.org/10.1109/JAS.2023.124107
2. X. Li, S. Yu, Y. Lei, N. Li, B. Yang, Intelligent machinery fault diagnosis with event-based camera, *IEEE Trans. Ind. Inform.*, **20** (2023), 380–389. https://doi.org/10.1109/TII.2023.3262854
3. R. Girshick, D. Jeff, D. Trevor, M. Jitendra, Region-based convolutional networks for accurate object detection and segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2015), 142–158. https://doi.org/10.1109/TPAMI.2015.2437384
4. R. Girshick, Fast R-CNN, In: *2015 IEEE International conference on computer vision (ICCV)*, IEEE, 2015, 1440–1448. https://doi.org/10.1109/ICCV.2015.169
5. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
6. J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks, In: *Proceedings of the 30th international conference on neural information processing systems*, 2016, 379–387.
7. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, In: *2017 IEEE International conference on computer vision (ICCV)*, 2017, 2961–2969. https://doi.org/10.1109/ICCV.2017.322
8. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, *arXiv:1506.02640*, 2015. https://doi.org/10.48550/arXiv.1506.02640

9. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, In: *2017 IEEE Conference on computer vision and pattern recognition (CVPR)*, 2017, 6517–6525. https://doi.org/10.1109/CVPR.2017.690
10. J. Redmon, Ali Farhadi, Yolov3: An incremental improvement, *arXiv:1804.02767*, 2018. https://doi.org/10.48550/arXiv.1804.02767
11. A. Bochkovskiy, C. Wang, H. M. Liao, YOLOv4: Optimal speed and accuracy of object detection, *arXiv:2004.10934*, 2020. https://doi.org/10.48550/arXiv.2004.10934
12. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, et al., SSD: Single shot multibox detector, In: *Computer vision-ECCV 2016*, Springer, **9905** (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
13. T. -Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, In: *2017 IEEE International conference on computer vision (ICCV)*, 2017, 2999–3007. https://doi.org/10.1109/ICCV.2017.324
14. M. Afif, R. Ayachi, Y. Said, E. Pissaloux, M. Atri, An evaluation of RetinaNet on indoor object detection for blind and visually impaired persons assistance navigation, *Neural Process Lett.*, **51** (2020), 2265–2279. https://doi.org/10.1007/s11063-020-10197-9
15. R. Ayachi, M. Afif, Y. Said, M. Atri, Traffic signs detection for real-world application of an advanced driving assisting system using deep learning, *Neural Process Lett.*, **51** (2020), 837–851. https://doi.org/10.1007/s11063-019-10115-8
16. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft coco: Common objects in context, In: *Computer vision-ECCV 2014*, Springer, **8693** (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
17. M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes (voc) challenge, *Int. J. Comput. Vis.*, **88** (2010), 303–338. https://doi.org/10.1007/s11263-009-0275-4
18. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, In: *2016 IEEE Conference on computer vision and pattern rRecognition (CVPR)*, 2016, 770–778. https://doi.org/10.1109/CVPR.2016.90
19. F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, et al., BDD100K: A diverse driving dataset for heterogeneous multitask learning, In: *2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, 2020, 2633–2642. https://doi.org/10.1109/CVPR42600.2020.00271
20. C. Ning, H. Zhou, Y. Song, J. Tang, Inception single shot multibox detector for object detection, In: *2017 IEEE International conference on multimedia & expo workshops (ICMEW)*, 2017, 549–554. https://doi.org/10.1109/ICMEW.2017.8026312
21. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, In: *2016 IEEE Conference on computer vision and pattern recognition (CVPR)*, 2016, 2818–2826. https://doi.org/10.1109/CVPR.2016.308
22. X. Wang, H. -M. Hu, Y. Zhang, Pedestrian detection based on spatial attention module for outdoor video surveillance, In: *2019 IEEE Fifth international conference on multimedia big data (BigMM)*, 2019, 247–251. https://doi.org/10.1109/BigMM.2019.00-17
23. Z. Zhang, J. Wu, X. Zhang, C. Zhang, Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project, *arXiv:1712.0953*, 2017. https://doi.org/10.48550/arXiv.1712.09531
24. X. Wang, Dml dataset. Available from: https://dml-file.dong-liu.com

25. M. Schembri, D. Seychell, Small object detection in highly variable backgrounds, In: *2019 11th International symposium on image and snal pocessing and aalysis (ISPA)*, 2019, 32–37. https://doi.org/10.1109/ISPA.2019.8868719
26. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv:1409.1556*, 2014. https://doi.org/10.48550/arXiv.1409.1556
27. Dataset of images of trash, 2017. Available from: https://github.com/garythung/trashnet
28. W. Wu, J. Lai. Multi camera localization handover based on YOLO object detection algorithm in complex environments, *IEEE Access*, **12** (2024), 15236–15250. https://doi.org/10.1109/ACCESS.2024.3357519
29. A. Yadav, P. K. Chaturvedi, S. Rani, Object detection and tracking using YOLOv8 and DeepSORT, *Adv. Communi. Syst.*, 2024, 81–90. https://doi.org/10.56155/978-81-955020-7-3-7
30. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2017), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

AIMS Press