



Article

Novel Deep Learning Domain Adaptation Approach for Object Detection Using Semi-Self Building Dataset and Modified YOLOv4

Ahmed Gomaa ^{1,*} and Ahmad Abdalrazik ²

¹ National Research Institute of Astronomy and Geophysics (NRIAG), Helwan 11731, Egypt

² Electrical Engineering Department, Faculty of Engineering, Port Said University, Port Said 42524, Egypt; ahmadabdalrazik@eng.psu.edu.eg

* Correspondence: ahmed.gomaa@ejust.edu.eg or ahmed.gomaa@nriag.sci.eg

Abstract: Moving object detection is a vital research area that plays an essential role in intelligent transportation systems (ITSs) and various applications in computer vision. Recently, researchers have utilized convolutional neural networks (CNNs) to develop new techniques in object detection and recognition. However, with the increasing number of machine learning strategies used for object detection, there has been a growing need for large datasets with accurate ground truth used for the training, usually demanding their manual labeling. Moreover, most of these deep strategies are supervised and only applicable for specific scenes with large computational resources needed. Alternatively, other object detection techniques such as classical background subtraction need low computational resources and can be used with general scenes. In this paper, we propose a new a reliable semi-automatic method that combines a modified version of the detection-based CNN You Only Look Once V4 (YOLOv4) technique and background subtraction technique to perform an unsupervised object detection for surveillance videos. In this proposed strategy, background subtraction-based low-rank decomposition is applied firstly to extract the moving objects. Then, a clustering method is adopted to refine the background subtraction (BS) result. Finally, the refined results are used to fine-tune the modified YOLO v4 before using it in the detection and classification of objects. The main contribution of this work is a new detection framework that overcomes manual labeling and creates an automatic labeler that can replace manual labeling using motion information to supply labeled training data (background and foreground) directly from the detection video. Extensive experiments using real-world object monitoring benchmarks indicate that the suggested framework obtains a considerable increase in mAP compared to state-of-the-art results on both the CDnet 2014 and UA-DETRAC datasets.

Keywords: background subtraction; YOLO v4; deep learning; object detection; domain adaptation



Citation: Gomaa, A.; Abdalrazik, A. Novel Deep Learning Domain Adaptation Approach for Object Detection Using Semi-Self Building Dataset and Modified YOLOv4. *World Electr. Veh. J.* **2024**, *15*, 255. <https://doi.org/10.3390/wevj15060255>

Academic Editors: Yuan-Wei Tseng, En-Chih Chang and Chun-An Cheng

Received: 18 May 2024

Revised: 5 June 2024

Accepted: 9 June 2024

Published: 12 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Traffic moving object detection in real-world recorded videos is used to recognize the changing or moving areas in the camera view [1–3]. It is an essential task in computer vision and plays a vital role in various applications, including intelligent transportation systems (ITSs), fault detection, area monitoring, motion capture, disaster management, scene analysis, and visual surveillance [4–8]. Several image processing algorithms have been developed, in the literature, to detect a particular object or event in the video scenes. The use of machine learning in image processing has recently experienced considerable growth, and these methods have been exploited in video surveillance applications. Traditional object detection strategies can be divided into three different groups: the frame difference strategy [9], background subtraction (BS) approach [10], and optical flow. The optical flow-based strategy requires measuring the entire frame's optical information, which often makes the detection process in real-time applications hard to satisfy. On the other hand,

the detection-based frame difference approach is simple and fast, but only parts of the moving objects are identified. This is because the aforementioned approach only compares the difference in successive frames between the moving objects and the background. Background subtraction is one of the most common object detection techniques. It first implements the modeling of the background and then identifies the moving objects by measuring the difference between the input frame and the background. One big downside of background subtraction is that it is not straightforward to model the background over the entire sequence in real scenarios. However, object detection/recognition based on traditional methods depends on handcrafted background features and predefined assumptions that fit only specific applications. Therefore, research has been attracted lately to deep learning systems that outperform traditional approaches.

The utilization of deep learning methods removes the need for handcrafted feature-based and predefined assumption-based object detection/recognition, and the raw image data can be fed directly into a deep learning network for achieving detection/recognition. Background subtraction with CNNs has recently achieved noticeable advance. This incorporates low /mid /high-level functionality across an end-to-end structure, resulting in good representational flexibility as well as substantial current positions. Many CNN-based object detection methods have attracted much attention, such as SSD [11], YOLO v3 [12], R-CNN [13], DPM [14], FG-BR-NET [15], and Center-Net [16]. Although the aforementioned methods can effectively overcome many challenges of real-world applications, these methods usually require a large amount of densely labeled video training data, which are hard to collect in real-world scenarios. Moreover, when a detection model is trained using generic datasets (like MS COCO and PASCAL VOC) and applied to a specific scene, the distribution gap between the specific scene data and general data limits the model's precision. Due to the lack of labeled data from a specific scene, it is difficult to train the model for that scene. Even if we can obtain some labeled data from a specific scene, a pre-trained model would perform poorly due to the changing environmental variables.

In the context of intelligent transportation systems, clustering techniques play a pivotal role in analyzing and managing traffic data effectively. It is essential to recognize that various clustering techniques, such as fuzzy clustering and DBSCAN, offer distinct advantages in handling diverse traffic scenarios. Fuzzy clustering, allows for the classification of data points with degrees of belonging, facilitating more nuanced analysis in cases where data may not be distinctly separable. On the other hand, DBSCAN is particularly effective in identifying clusters of arbitrary shape and dealing with noise, making it suitable for dynamic and complex traffic environments.

The application of these clustering methods in intelligent transportation systems has been explored in several significant works. For instance, in [17], the authors demonstrated how fuzzy c-means clustering could be used to effectively segment and manage traffic data, leading to improved traffic flow and reduced congestion. Their study highlighted the ability of fuzzy c-means to handle overlapping data points, which is common in urban traffic scenarios where the distinction between different traffic states is not always clear-cut. Similarly, in [18], the use of DBSCAN was shown to be highly effective in detecting traffic patterns and anomalies in real-time traffic data. The study illustrated the robustness of DBSCAN in managing large datasets with noise and its capability to identify clusters of various shapes and sizes, which is crucial for real-time traffic monitoring and incident detection. Additionally, hierarchical clustering has also shown significant potential in the domain of intelligent transportation systems. In [19], the authors explored the use of hierarchical clustering to analyze traffic data, demonstrating its effectiveness in identifying nested clusters that represent different levels of traffic congestion and flow. This approach allows for a more detailed analysis of traffic patterns, providing insights that can be used to optimize traffic management strategies at both macro and micro levels. In creating a hierarchy of clusters, hierarchical clustering facilitates a deeper understanding of the underlying structure of traffic data, making it a valuable tool for comprehensive traffic

analysis. However, we employed the hierarchical clustering as an important step in our model to minimize the manual annotation process.

In this paper, we propose a new detection framework that adopts a scene-specifying technique with a semi-automatic labeler. The semi-automatic labeler can replace manual labeling by adopting the motion information extracted directly from the detection video to provide labeled training data (background and foreground). We initially used the low-rank decomposition to discriminate between the background and foreground and then used the results for the training. Moreover, we present a modified version of the YOLO V4 architecture, and instead of training from scratch, we fine-tuned the model on a larger dataset and used YOLO V4 weights to initialize the model. In summary, this study provides the following four main contributions:

- An efficient object detection and classification framework is presented. The developed strategy exploits the traditional background subtraction features and CNN benefits to obtain a semi-automatic learning strategy with more reliable detection and classification results.
- A modified architecture of YOLO V4 is suggested in this work with a new continuous, smooth, and self-regularized activation function that achieves a better network nonlinear fitting ability.
- A semi-automatic CNN-based learning strategy is presented, employing hierarchical structuring and object feature information analysis to refine the images used for fine-tuning.
- Extensive experiments on two benchmark datasets demonstrate the effectiveness of the proposed strategy and its ability to achieve better detection and classification results with low annotation costs.

The main steps of the proposed object detection framework are described in the following subsections. The rest of this work is organized as follows. Section 2 summarizes the related work. Section 3 presents the proposed strategy. Section 4 reports the experimental results and discussion. Sections 5 and 6 present the future work and conclusion.

2. Related Work

It has been shown that efficient object detection can be obtained by adopting CNNs [20], such as regions with convolutional neural networks (R-CNN), single-shot detectors (SSD), background subtraction-based CNNs, and You Only Look Once (YOLO) models. The R-CNN strategy [21] divides object detection problems into three stages: generating bounding box proposals, measuring each region's features, and classifying those regions. The authors of [22] presented a Fast RCNN with a significantly improved accuracy and limited speed performance. To solve this problem, the authors of [23] proposed an end-to-end detection strategy, with a single-shot multibox detector (SSD) that obtains proposal regions through uniform extraction, which improves the detection speed. Recently, the authors of [12] utilized multi-scale prediction to provide a modified YOLO with an enhanced detection speed, which improves the basic classification network. The strategy was able to achieve a large number of object detections by performing joint training for both object detection and classification. On the 156-class version of COCO, YOLO9000 achieved a 16% mean average precision (MAP), while YOLO can detect 9000 separate classes, with limited accuracy. A tedious and time-consuming foreground annotation process can be manually conducted, taking about 60 second for each frame. Background subtraction approaches based on CNNs have attracted more attention because they effectively solve real-world issues (such as dynamic backgrounds, shadows, and lighting variations). However, these approaches often need a huge quantity of highly labeled training data images, which is rarely available in real-life applications.

Recently, CNNs have been applied in background subtraction (change detection) and achieved notable improvements. Two typical methods have been proposed, namely, fully automatic and user interaction-based methods. The first one learns a specific CNN model for each video but requires the manual labeling of training frames on the fly. The other

one learns a universal model offline, but has a limited performance in handling various surveillance scenarios. The CNN-based background subtraction methods have attracted attention lately, as they are effective in addressing real-world challenges, such as shadows, dynamic backgrounds, and illumination changes. However, these methods usually require a large amount of densely labeled video training data, which are rarely collected in real-world situations. The authors of [24–26] trained neural network models using half of the video frames. The authors of [27] tried to achieve better experimental results using a lot of viewer training samples. The authors of [28] introduced a cascade-CNN method, which achieved the best performance in the largest change detection dataset (i.e., CDnet2014). However, for every video sequence, it needs to manually select the training frames and label the corresponding ground truth. Obviously, such a semi-automatic manner is labor consuming, especially when the changing circumstances of video surveillance scenes are imposed. Alternatively, the new strategy in [24] suggests to learn a universal model offline, avoiding the manual labeling of each video. However, the universal model also has a limit in its performance in handling various surveillance scenarios. In [29,30], the authors introduced a guided learning strategy that learns a specific CNN for each video to ensure accuracy but manages to avoid manual labeling.

The authors of [24] introduced a new strategy to generate a background model by combining a segmentation mask from the SuBSENSE algorithm and the output of the Flux Tensor algorithm, which can dynamically change the parameters used in the background model based on the motion changes in the video frames. Then, they presented a novel background subtraction strategy based on a CNN. They adopted 5 per cent of random video frames for the training, and they employed pixel-level feedback instead of utilizing the manually training images to reduce the human involvement. However, the CNN yielded poor results in some CDnet categories since the background images provided by the proposed algorithm were insufficient for good background subtraction. Therefore, their method exhibited poor segmentation for these categories, and as a consequence, the average FM dropped significantly.

In [28], the authors suggested an accurate semi-automatic technique for extracting foreground objects in surveillance sequences. The system aims to deliver accurate findings to be considered ground truth by providing minimal human interactions. The model does not require many instances for data fitting because the moving foreground and background are highly redundant between images. The model is based on a multi-resolution CNN with a cascaded architecture. Given a specific video, the user first outlines foreground objects from a small set of frames. These manually annotated frames are then used as training data by the algorithm. After training, the model automatically labels the remaining video images. The authors pre-trained their model using a larger dataset, Pre-training was performed only once with the Motorway dataset, with pixel-accurate ground truth of video surveillance images. After transferring the weights to the models, they fine-tuned the CNN parameters for each video. However, the algorithm yielded poor results on some datasets of images with different visual appearances from those in the Motorway dataset.

To address the problems associated with background subtraction based-CNNs, the authors of [29] presented a new deep background subtraction method by introducing a guided learning strategy. The main idea was to learn a specific CNN model for each video to ensure accuracy but manage to avoid manual labeling. Firstly, the authors implemented the SubSENSE strategy for obtaining initial segmentation results before employing their designed adaptive method that selects robust pixels to guide the training process. In addition, they also designed a simple strategy to select informative frames for guided learning automatically. They designed two selection strategies: a simple strategy to automatically select informative frames and an adaptive strategy to select reliable pixels for guided learning. Moreover, they directly used the model pre-trained on a large dataset, ImageNet, and then fine-tuned the CNN model for each video. In [30], modern CNN-based detectors needed large data inputs, usually requiring manual labeling. In the presented research approach, the weakly supervised learning paradigm was used to train a CNN-based detector

employing labels obtained automatically through an application of a video background subtraction algorithm (GMM). Background subtraction strategies combined with object detection were firstly presented in [31] by introducing a semi-supervised background model-based training strategy to discriminate between an object or clutter and retraining the model with the best images. In [32], traditional background subtraction was used with a CNN for anomaly detection. The authors used the SuBSENSE as an unsupervised background subtraction method to extract moving objects in scenes and a pre-trained GoogLeNet to classify these objects.

Recent advancements in semi-supervised learning and object detection have enhanced the performances and capabilities of detection models like YOLO. The authors of [33] presented a significant contribution called the SSVOD (Semi-Supervised Video Object Detection with Sparse Annotations) framework, which introduces a novel approach to video object detection using semi-supervised learning. The authors proposed a method that leverages sparse annotations in video frames, significantly reducing the labeling effort while maintaining a high detection accuracy. This method combines temporal information and consistency regularization to improve the robustness of video object detection models, making it a valuable contribution to the field. Moreover, a survey on semi-supervised graph clustering [34] provided an in-depth analysis of graph-based clustering methods under semi-supervised settings, highlighting the advancements and applications in various domains. This survey underscores the versatility of semi-supervised learning in different contexts and its impact on improving clustering performance. Recent strategies include new activation functions such as the SwiGLU (Switchable Gated Linear Units), which have shown promise in improving both convergence and performance in various neural network architectures [35]. Additionally, contemporary studies on semi-supervised learning techniques have proposed methods like MixMatch [36] and FixMatch [37], which combine data augmentation with consistency regularization, demonstrating substantial improvements in image classification and object detection tasks.

Unlike the other approaches, our framework uses background subtraction (low-rank decomposition) to detect the moving areas and construct objects from them while keeping the CNN-based object detector result intact. As a result, the system can recognize all moving object types in the sequence, whether it is moving or not, independent of the CNN's trained classes. The main contribution is a detection framework that overcomes manual labeling and creates a semi-automatic labeler that can replace manual labeling using motion information to supply the labeled training data (background and foreground) directly from the detection video. Low-rank decomposition is firstly used to differentiate the foreground and background regions successfully. Then, we refine the results by employing hierarchical clustering and object feature analysis with less human involvement to utilize the refined images for the training. Moreover, we present a modified architecture of YOLO v4, and instead of training from scratch, we fine-tune the CNN parameters.

3. Methodology

Despite the advancements in object detection techniques [38], the accuracy of a detection model trained on generic datasets when applied to a specific scenario is constrained by the distribution distance between the generic and specific scene details. Training a model for a specific scene is difficult because of the lack of labeled data for that scene. Moreover, if we can extract some labeled data for this scene, the pre-trained model will fail because of the change in environmental conditions. In view of these concerns, we suggest semi-automatic scene-specific object detection based on parallel vision between a generic object detection model and novel refined background subtraction result for a specific scene, as summarized in Figure 1. The following section describes our strategy for improving the generic ground truth-trained object detection methods by presenting a semi-automatic manner for reducing the labor involved. The strategy contains three main components in each frameset—a candidate extractor, refining with clustering, and retraining the modified

YOLO V4 architecture—as shown in Figure 1. The main steps of the proposed scheme are described in the following subsections.

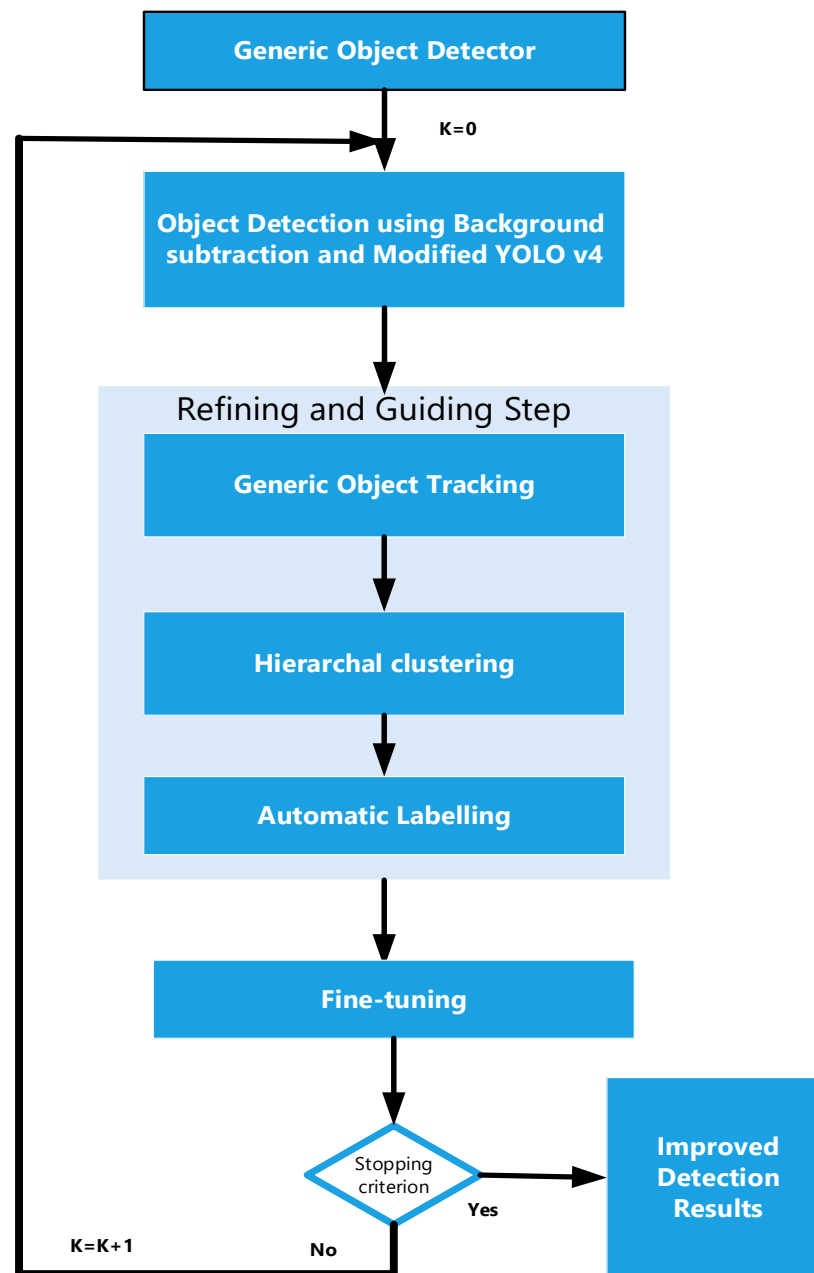


Figure 1. Workflow chart of the proposed system for a specific frameset K.

3.1. Initial Object Detection

This step initially extracts and classifies the objects in the first frame, every fixed number of frames, as shown in Figure 1. We utilized two strategies that worked in parallel in this stage, in which the classification performance was improved by retraining the convolutional network with new annotated data, and a background subtraction method that extracts the foreground objects as a basis for forming new classes or producing more data of an already recognized object class [39]. Object classifiers have recently demonstrated their ability to recognize objects in images; however, if they are not trained for a specific class or if their training samples do not include a special case of an object, the network will fail to classify the object or will classify it incorrectly.

This stage aims to extract initial object detection results so that a user can check them and, if necessary, correct or add classifications. We can categorize these initial results into three categories:

- Correctly classified objects; we will not lose such information.
- Incorrectly classified objects; the labels of these objects must be corrected.
- Unclassified objects; the user will classify such objects.

YOLO v4 is considered the state-of-the-art object detection strategy, so we modified this architecture. Moreover, we selected BS-based low-rank decomposition as a background subtraction methodology because it includes a group sparsity constraint and the background model can be updated online, improving the object detection accuracy in complex dynamic scenes. YOLO can classify and detect objects by forming bounding boxes around them. Meanwhile, the background subtraction method with low-rank and sparse decomposition also detects moving objects by assigning bounding boxes around them that can be compared to YOLO's output. Figure 2 shows an example of using YOLOv4 and the background subtraction method to detect objects in a sofa scene. In this image, it is worth noting the importance of using the background subtraction along with state-of-the-art object detection schemes to improve the object detection results. As shown in Figure 2, the person is detected by both YOLOv4 and BS methods, while the background subtraction cannot detect the dining table, chair, or sofa, because these objects are part of the background, and the BS only detects the box. Moreover, YOLOv4 can detect the dining table, chair, and sofa and cannot detect the box because YOLO was not trained on the box class. Such different scenarios are taken into consideration for designing the proposed framework, ensuring that the network retains its previous knowledge beside being able to incorporate new scenarios.



Figure 2. A test result of YOLOv4 and background subtraction for the sofa scene example. The dining table, chair, and sofa are only detected by YOLOv4, while the box is only recognized by the background subtraction. (a) Background subtraction with sparse + low-rank decomposition output. (b) YOLOv4 output.

3.2. Object Clustering and Labeling

In the training stage, a huge number of objects are extracted using background subtraction and an object classifier. So, a clustering method is used to automatically gather the objects with the same appearances and class in one group, which facilitates humans in adding and correcting the object labels. There are various types of clustering methods that can be applied in data structure grouping [40–42].

The clustering process in the proposed strategy is based on two stages to improve the clustering performance; the first stage is called temporal grouping, while the second stage is hierarchical clustering. In the temporal grouping stage, both time and position information are combined, and objects with similar sizes in the same or close positions in subsequent frames are grouped to be in the same cluster. Hence, if objects have the same

label, they will be grouped together, while groups with no labels are clustered based on the hierarchical clustering strategy.

If an object does not remain in the scene for at least a few consequent frames, it is considered a noisy detection and discarded. Figure 3 indicates the temporal group forming process using the optical flow-based object tracking method with the size and position analysis of the detected bounding boxes. The bounding box is related to the same object when it is detected in subsequent frames with a position and size difference within a threshold, or it is regarded as a noise when these values have a large difference. Because BGS's bounding boxes are more sensitive to changes in light and shadows, the proposed system prioritizes the YOLO's bounding boxes if both strategies detect an object. The groupings are then packaged into even bigger clusters, requiring just one label for all of the images included inside them, thus minimizing human work. These bigger bundles represent the clusters. The final output of the temporal grouping consists of two types of clusters:

- Labelled clusters: groups with classification based on YOLO classifier (e.g., person, vehicle, and boat).
- Unlabelled clusters: groups without classification and with similar hierarchical characteristics.



Figure 3. Example of results extracted from streetCornerAtNight and blizzard scenes. **First row:** YOLO-based detection results; **second row:** detection results using BGS with their bounding boxes.

The hierarchical clustering strategy divides foreground objects into several subcategories, including person, car, van, bus, and others, and the user is responsible for reviewing them. The final cluster output is then used for transfer learning, which is used as the training data of the modified YOLO v4. Hierarchical clustering is usually divided into two types: agglomerative and divisive. The first considers all data points as the smallest cluster. It combines the two closest ones from bottom to top according to specific distance metric and linkage criteria, while the other works in the opposite direction. The clustering process is implemented in three main steps, as follows:

1. Assume that the foreground objects $O = \{o_1, o_2, o_3, \dots, o_k\}$.
2. A bottom-up approach is used to aggregate the samples, with the Euclidean distance being the measure of similarity between categories:

$$d(o_i, o_j) = \|o_i - o_j\|_2 \quad (1)$$

where $i, j = 1, 2, 3, \dots, k$. The average distance between all pairs of objects in any two clusters is used as a linkage criterion:

$$D(a, b) = \frac{1}{n_a n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} d(o_{ai}, o_{bj}), \quad (2)$$

where a and b are clusters, and n_a and n_b are the object numbers in clusters a and b , respectively. Similarly, in clusters a and b , o_{ai} and o_{bj} represent the i th and j th objects, respectively.

3. The Car, van, and bus categories in the hierarchical clustering are selected separately, while the remaining categories are classified as other categories.

Figure 4 shows the hierarchical clustering process of the detected foreground objects.



Figure 4. Hierarchical clustering and labeling.

3.3. Manual Classification Refinement and Incremental Learning

In this work, human involvement was used to improve the effectiveness of the original object detection by providing new and correct label information. Furthermore, fine-tuning was used to adapt the network with different occurrences of known object classes, while incremental learning was also used to add a new object classes to the network. Manual involvement was used to refine the object labels by removing the falsely labeled groups from a specific cluster and labeling the unlabeled clusters. Furthermore, the user removed the noise groups from the unclustered groups and moved the individual groups to their correct clusters. After the labeling refinement and review performed by the user, the system gathers the final clusters with the correct labels and then automatically annotates the associated frames of the detected objects. As a result, an output dataset is generated that is ready for the CNN retraining process. The retraining process depends on the object types in the clusters, and if they are known classes, the new data will be added, and the original training will continue with the same network configuration. On the other hand, when introducing a new class to the network, the last layers will be modified with new neurons to classify the new class in the last layers. Figure 5 shows examples of unlabeled clusters, misplaced groups, and noise points, and it indicates the importance of the manual intervention process.

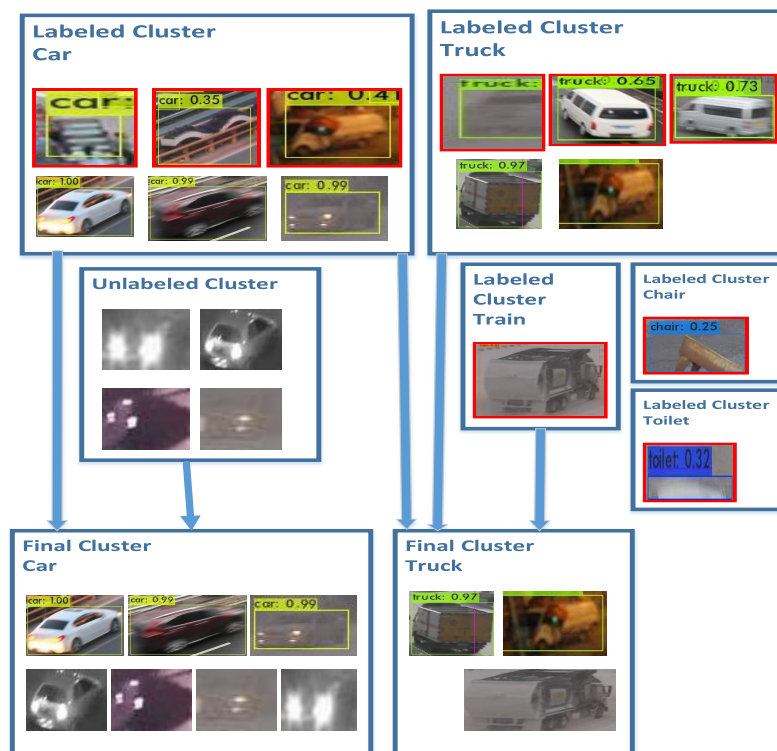


Figure 5. Selection process example. Highlighted objects do not belong to the right class and must be manually rectified. First and second rows: clustering output including unlabeled and labeled clusters. Third row: final clusters after the user reviews the truck and car labels.

3.4. YOLOv4 Modification

For this stage, we modified the YOLOv4 architecture by presenting a new model to ensure better accuracy in automatic model learning. The suggested model introduces a new activation function, a non-monotone, continuous, self-regularized, and smooth activation function, since a better activation function ensures a better network nonlinear fitting ability. The suggested activation function is $f(x) = \arctan(\ln(1 + e^{\alpha x}))$, which will make the gradient propagate effectively with fewer side effects. It is unbounded as $x \rightarrow +\infty$ and bounded as $x \rightarrow -\infty$. The first derivative of the new activation function can be obtained by differentiating $f'(x) = w(x) + \frac{\alpha x e^{\alpha x}}{\Phi(x)}$. When $w(x) = \arctan(\ln(1 + e^{\alpha x}))$, $\phi(x) = ((e^{\alpha x} + 1) * (2 * \ln(e^{\alpha x} + 1) + 1))$.

The new activation function added to YOLOv4 was designed to be smooth, continuous, self-regularized, and non-monotonic, which are desirable properties for an activation function in a neural network. The suggested activation function is also shown to have a faster convergence speed and lower loss compared to Mish. The suggested activation function is more effective in improving the model's performance due to its ability to alleviate the problem of gradient disappearance during training. This activation function was designed to have a larger gradient change when the input value approaches the extremes, which helps alleviate the problem of gradient disappearance during training. This property is particularly important in deep neural networks where the gradients can become very small, making it difficult for the model to converge. This activation function also has a more moderate gradient change compared to Mish, which helps the model generalize better to unseen data. So, this activation function is more robust and less prone to overfitting compared to Mish, which is beneficial for improving the model's performance. This activation function was compared with Mish, another popular activation function used in YOLOv4, and it has a higher mean average precision (mAP) of 80% compared to Mish's 77%. A preliminary result was recorded for the new object detection model. It can be seen from Figure 6 that the total loss was reduced by 4.32%, and the Val loss by 6.37%,

when using the new activation function compared with the Mish activation function in a state-of-the-art object detection model, and the loss function convergence of our model is faster during the iterations.

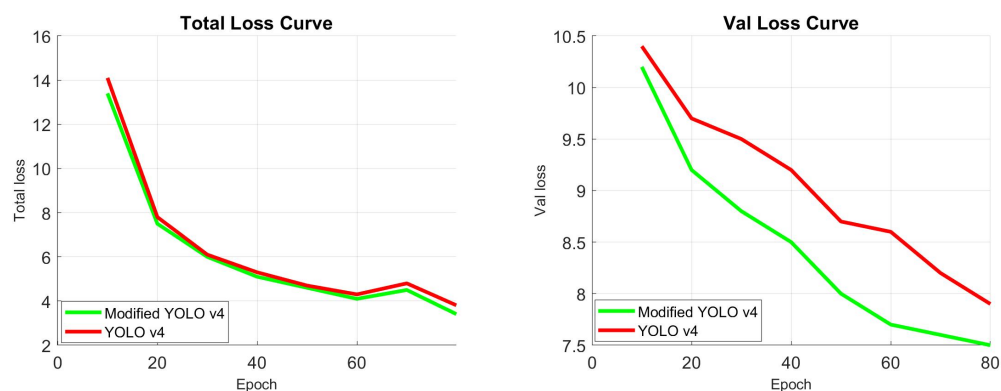


Figure 6. Comparison between the YOLO v4 object detection model and the proposed object detection model.

4. Experimental Results and Discussion

This study tested the proposed strategy's efficacy on different datasets, including the CD.net 2014 [43] and UA-DETRAC [44] datasets, to investigate its robustness and generalization. Although recent deep learning-based object detection approaches, such as YOLO v4, YOLO v3, DPM, FG-BR-NET, and Center-Net, have achieved good accuracy, they still need improvements to achieve an excellent performance in a specific scene. Such detectors are not considered domain-specific detectors; they are simply universal detectors. For example, YOLO V4 can obtain outstanding results in a normal scene, but its capabilities may be restricted when applied to a specific field. Moreover, based on the initial experiment results, YOLO v4 did not perform as well as expected in some scenes. These types of scenes had a turbulent environment with wrong classification or no detection results and a bad weather environment with low recall and misclassification results. Five videos were selected from the CD.net 2014 dataset, including blizzard scene, street corner at night, turbulence0, turbulence2, and busy-boulevard. The experiments were conducted by employing MATLAB on an Intel i7-4810MQ CPU 2.80 GHz, 16 GB of RAM, and a Quadro K1100M GPU with 2 GB of video RAM. We configured the epoch value to 200, the batch size to 8, the learning rate to 0.001, the momentum to 0.97, and the decay rate to 0.0005. During the training phase, 10% of the training set was randomly selected to fine-tune the training parameters. In order to speed up the training of the model, we first iteratively trained 100 epochs of the YOLOv4 with a pre-training weight and replaced the Mish of YOLOv4 with our suggested activation function. In iterating 100 epochs with a pre-training weight again, it can be seen that the loss function convergence of YOLOv4 using our suggested activation function was faster during the iterations.

The manual involvement process is responsible for refining the resulting clusters and looking for the ones we want to utilize (unlabeled or labeled). Subsequently, the groups that did not belong to the correct cluster were reviewed and filtered out, as shown in Figure 5; for example, we removed the truck from both the car and train class. Finally, we assigned the correct labels to the refined final clusters. We evaluated orphan groups in certain situations, such as those filtered out of a cluster and noise points. We allocated a label to them, as shown in Figure 5, after filtering the group that contains a truck and labeled it as truck. In all experiments, we utilized two steps to train our network for a better performance. We used the YOLOv4 original weight as a starting point, which was trained on the COCO dataset. Also, we adopted transfer learning with the refinement cluster to create a strong structure for object detection. The detection accuracy in the first experiment was evaluated using quantitative performance criteria that were utilized as a standard

evaluation: the recall, precision, and F-measure, where R = recall, P = precision, FN = false negatives, FP = false positives, and F = F-measure.

In all the experiments, the network efficiency was improved for the specific scene. Improvements in the recall, precision, and F-measure can be seen in all scenes, as shown in Table 1. The most obvious example is the turbulence0 scenario, in which the original YOLOv4 and YOLOv3 were nearly, completely incapable of accurately detecting anything. In evaluating the performances of the object detection models, particularly in real-time applications such as surveillance or intelligent transportation systems, the frames per second (FPSs) metric holds paramount importance. The FPSs directly correlate with the speed at which the model can process video frames, impacting its ability to detect objects swiftly and respond to dynamic scenarios in real time. Comparing the FPSs across different versions of YOLO models reveals valuable insights into their efficiency and computational speed, as shown in Table 1. Generally, YOLOv3 tends to offer the highest FPSs (19 FPSs) among the considered versions, making it well suited for applications where real-time processing is crucial, and sacrificing a slight degree of accuracy is acceptable. On the other hand, YOLOv4 achieved 15 FPSs, and the modified version introduced in this work achieved 17.5 FPSs, which showcases notable improvements in both accuracy and speed compared to its predecessors. Despite a slight reduction in FPSs compared to YOLOv3, the modified YOLOv4 strikes a favorable balance between speed and accuracy, making it a compelling choice for applications demanding an enhanced detection performance without significant compromises in the processing speed.

Table 1. Object detection result comparison for CDnet-2014 sequences.

Video Sequence	Compared Algorithm	TP	FP	FN	Precision %	Recall %	F-Measure %	FPS
Street corner at night	YOLOv3	93	11	396	89	19	31	20
	YOLOv4	339	36	495	90.4	40.6	56	15
	Proposed	680	18	154	97.4	81	85.1	17
turbulence0	YOLOv3	0	655	1287	0	0	0	19
	YOLOv4	1001	145	286	87.3	77.8	82.3	15
	Proposed	1190	70	97	94.4	92.4	93.3	18
Blizzard	YOLOv3	994	151	713	87	58	70	18
	YOLOv4	1350	55	437	96	75	84	15
	Proposed	1590	40	197	97.5	88.9	93	17

Figure 7 shows the results of our proposed method compared with five state-of-the-art object detection approaches on the UA-DETRAC dataset. As clearly shown in Figure 7, the overall mean average precision of our proposed method is 89.44% on the UA-DETRAC test dataset, which is 69%, 42.74%, 13.35%, 10%, and 6.5% higher than those of DPM, R-CNN, YOLO v3, FG-BR-NET, and CenterNet, respectively. It is asserted that the R-CNN strategy performs better than the DPM strategy with an AP score of 51.21. The proposed state-of-the-art detection methods achieve AP scores higher than 70% AP, such as YOLOv3 with a 77.55% AP score, FG-BR-NET with a 80.53 % AP score, and CenterNet with a 83.57% AP score. As shown in Figure 7a–g, in including various subset environments, the presented method reaches the highest AP scores; for instance, when testing on a hard subset, an 80.13%, AP score is achieved. Most detectors achieved AP scores of less than 80% in the poor illumination environment such as night and rainy scenes, except our proposed strategy, which achieved 90.25%, and 84.82% AP in the night and rainy scenes, respectively. Moreover, the proposed strategy performs relatively well in scenes with better lighting conditions; for example, we achieved 94.70% and 92.04% AP scores in sunny and cloudy day scenes, respectively.

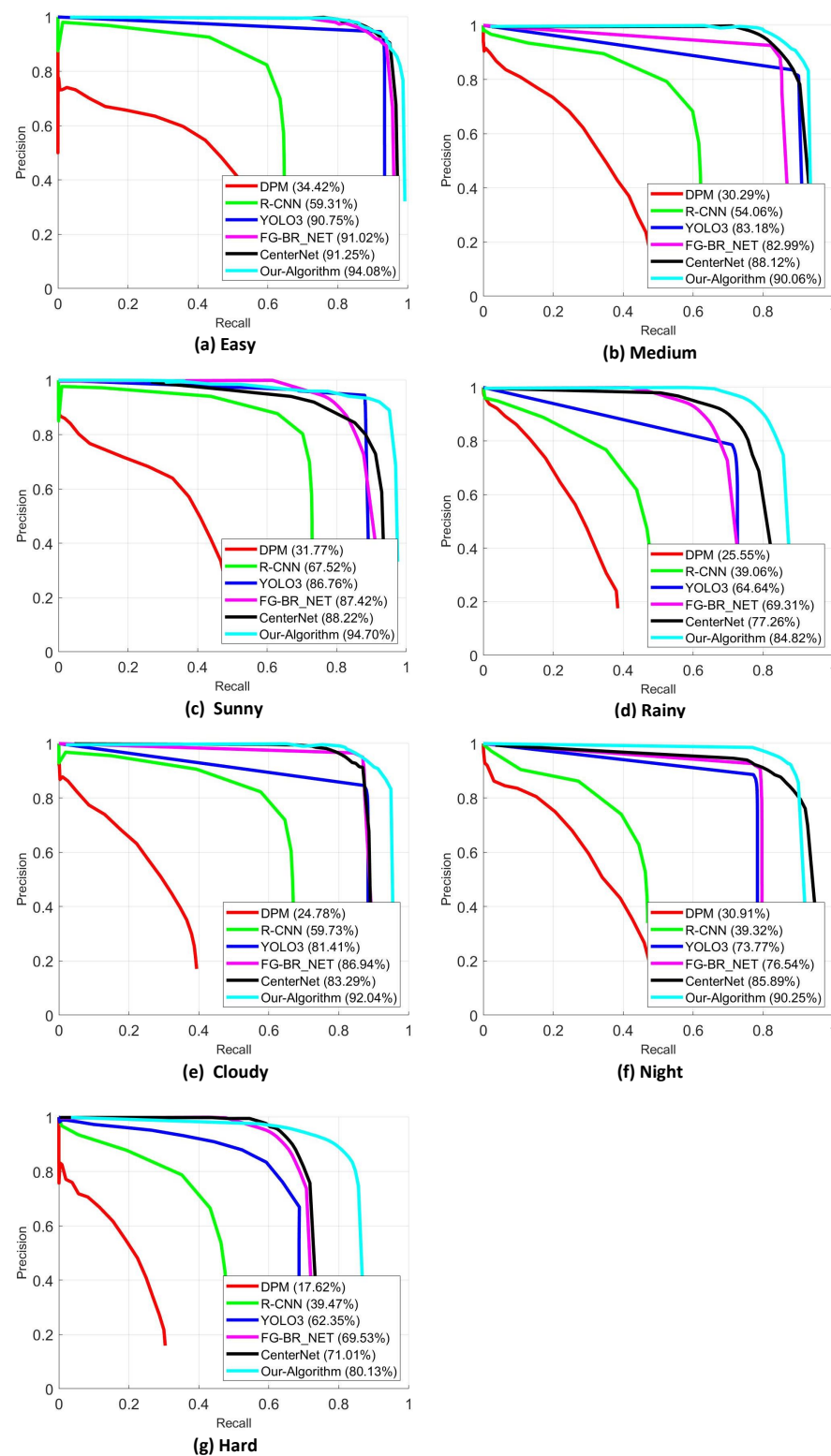


Figure 7. Precision–recall figures for different detection approaches on the UA-DETRAC dataset benchmark, including medium, easy, sunny, cloudy, rainy, hard, and night subsets. The legend scores show the AP scores used to compare the performance of the proposed strategy with the state-of-the-art object detection algorithms.

As shown in Figure 8, the detection methods can only perform relatively well on cars among all types of vehicles, and the proposed method outperformed all the compared strategies with a 95.6% AP score. It is worth mentioning that the proposed algorithm

outperformed the compared algorithms in terms of the AP score, especially for the other category because the refined background subtraction results added more diversity of images in the training. In other words, the more vehicle data used for the training, the more diversity added to the training images. For example, if we train a CNN model to detect a specific vehicle size on a sunny day, and the model has only seen images of a specific type of vehicle in a certain environment, then in practice, if it sees a vehicle with a different size in a night environment, it may not do so well at detecting and recognizing this vehicle. If we add more data to the CNN model to encompass more types of vehicles, then our training data will become more diverse. Moreover, in the proposed strategy, we can add new object classes to the training images, which will result in improving the AP score of the other category. Hence, the large variations in aspect ratio and scale for different vehicles are handled, in addition to the limited amount of training images.

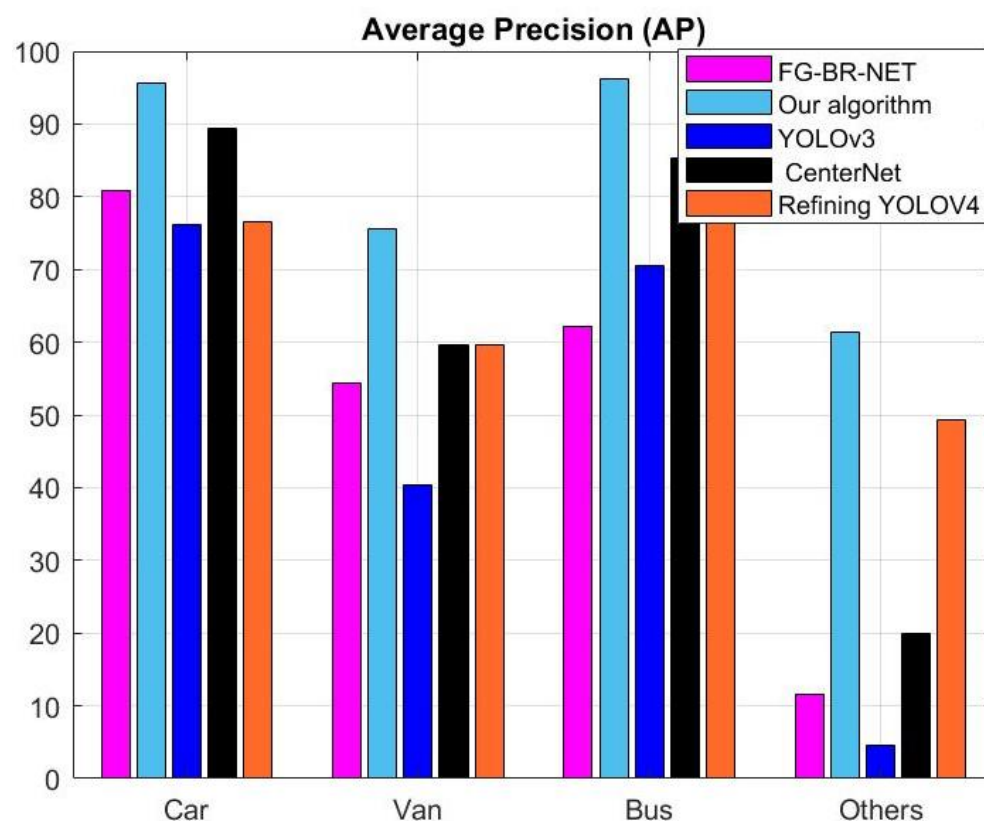


Figure 8. Influence of different object detection strategies on AP for various classes.

The authors of [15] proposed a new method to detect and classify objects utilizing the background subtraction strategies to generate a foreground image that will be given to the network as an additional input. They also utilized a feedback strategy between the background subtraction and detection results. The authors of [45] fine-tuned a state-of-the-art object detection strategy, YOLOv4, to precisely fit the needs of vehicle detection using various pre-existing changes and methodologies. The refined YoloV4 model boosted the performance in terms of each individual aforementioned metric. In [16], a new framework was presented based on a one-stage keypoint-based detector called CenterNet. They create two specialized modules, center pooling and cascade corner pooling, which enrich data acquired by both the top-left and bottom-right corners and offer more recognized data from the core regions. Each object is detected as a triplet of keypoints rather than a pair, which enhances the precision and recall. The proposed strategy is simple, elegant, and efficient when compared to these models. Furthermore, background subtraction is only required during the training phase, not during inference. CDnet2014 and UA-DETRAC sequences were used to evaluate the proposed method. It can be shown from Figure 9 that

the proposed method can be used in different traffic conditions and angles, demonstrating its durability and generality across a variety of traffic scenarios. Moreover, the method clearly achieves a high detection accuracy even with different scales of vehicles and varied monitor angles. In other words, the method can be adapted for a specific scene, which benefits from the background results and fine-tuning. Meanwhile, it is scale-insensitive due to the presented modified YOLOv4.

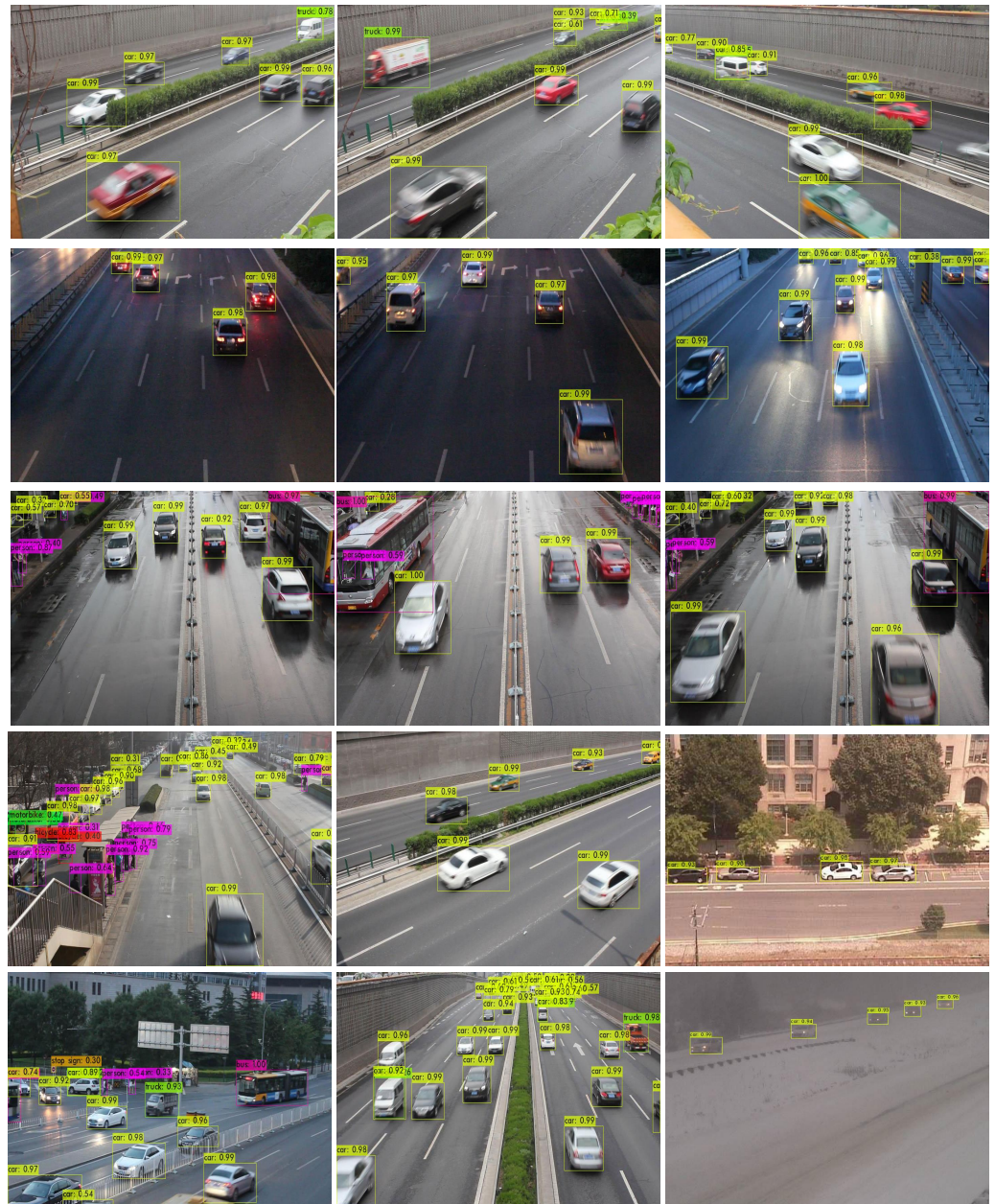


Figure 9. Sample results on CDnet2014 dataset and UA-DETRAC sequences. First row: crowded and cloudy. Second row: night scenes with varied light. Third row: rainy scenes. Fourth row: sunny sequences. Fifth row: hard scenes.

5. Future Work

The deployment of object detection systems in surveillance contexts necessitates the careful consideration of ethical implications, particularly regarding potential biases and fairness across different demographic groups. In our work, which integrates a modified YOLOv4 model with background subtraction techniques for unsupervised object detection,

it is critical to ensure that these systems do not perpetuate or exacerbate existing societal biases. These biases can arise from imbalanced training datasets that under-represent certain demographic groups, leading to disparities in detection accuracy and reliability. One of the primary sources of bias in deep learning models is the training dataset. If the dataset is not diverse, the model may not perform well for under-represented groups. To mitigate this risk, it is essential to collect data from various geographic locations and ensure the representation of different ages, genders, ethnicities, and environmental conditions. This diversity in the training data helps the model learn features that are generalizable across different demographic groups, thereby enhancing its fairness and reliability. For our model, we suggested using semi-automatic scene-specific object detection based on parallel vision between generic object detection models and using a novel refined background subtraction result for each specific scene, where our dataset was collected from various geographic locations to ensure the representation of different environmental conditions. Additionally, implementing fairness metrics and conducting regular bias audits can help identify and correct any biases in our model. The continuous monitoring and updating of the training data and model parameters in our strategy maintained fairness over time.

In our future work, we will focus more on new feedback mechanisms from real-world deployments in different applications to provide valuable insights on the model's performance across diverse settings. These feedback loops allow for ongoing improvements and adjustments to ensure that the model remains general, fair, and effective. Scaling the proposed method to handle larger datasets and more complex scenes presents an exciting avenue for exploration in future work. Scalability is a critical consideration for the practical deployment of object detection systems, especially in dynamic surveillance contexts. To address this challenge, we will further investigate advanced deep learning strategies aimed at improving model efficiency and scalability. One promising direction is the exploration of advanced model compression techniques, such as quantization, pruning, and knowledge distillation, augmented through attention mechanisms and transformer architectures. In incorporating attention mechanisms, the model can focus on relevant spatial and temporal features, enhancing its ability to detect objects in complex scenes with cluttered backgrounds or occlusions. Additionally, leveraging transfer learning and domain adaptation techniques, such as adversarial training, domain alignment, and self-training, can further improve scalability by enabling the model to generalize across diverse environments and datasets. Furthermore, to enhance the real-time performance, we plan to explore hardware acceleration solutions, including the utilization of specialized hardware accelerators such as GPUs or TPUs, as well as optimized software implementations for parallel processing. By harnessing the computational power of these hardware platforms and integrating attention mechanisms and transformer architectures, we can achieve faster inference speeds and an enhanced responsiveness, enabling real-time object detection in high-throughput surveillance scenarios. We aim to develop a scalable and efficient object detection framework capable of handling large-scale surveillance datasets and complex real-world scenes. The framework will not only be technically advanced but also socially responsible, ensuring a fair and unbiased performance across different demographic groups by enhancing safety and security while maintaining ethical standards.

6. Conclusions

In this work, we proposed a novel detection framework that combines a modified version of the YOLOv4 architecture and background subtraction techniques to perform unsupervised object detection for surveillance videos. Our approach addresses the limitations of existing object detection methods, which often require large, manually labeled datasets and are computationally intensive. By leveraging a semi-automatic labeling strategy, we reduce the need for manual labeling and improve the efficiency of the detection process. Our framework employs a low-rank decomposition-based background subtraction technique to extract moving objects followed by a clustering method to refine the results. The refined results are then used to fine-tune a modified YOLOv4 model, enhancing its performance

for specific scenes. The key contributions of this study include an efficient object detection and classification framework that combines traditional background subtraction with CNN benefits, resulting in a semi-automatic learning strategy with more reliable detection and classification outcomes.

Additionally, we introduced a modified YOLOv4 architecture featuring a new continuous, smooth, and self-regularized activation function that enhances the network's nonlinear fitting ability. Our semi-automatic CNN-based learning strategy utilizes hierarchical structuring and object feature analysis to refine the training images. Extensive experiments on benchmark datasets demonstrated the framework's effectiveness in achieving superior detection and classification results with reduced annotation costs. Extensive experiments on benchmark datasets demonstrated the framework's effectiveness in achieving superior detection and classification results with reduced annotation costs. On the CDnet 2014 dataset, our method achieved an average precision of 94.4%, which is a significant improvement over the state-of-the-art methods. Similarly, on the UA-DETRAC dataset, our framework achieved a mAP of 89.44%, outperforming other contemporary methods. These results indicate that our approach not only enhances the detection accuracy but also significantly reduces the manual effort required for dataset annotation. The proposed framework strikes a balance between computational efficiency and detection accuracy, holding promise for applications in self-driving vehicles, smart traffic management, and security surveillance for enhancing safety, efficiency, and reliability in dynamic environments.

To enhance the scientific value of this work, we recommend several directions for future research. First, exploring the integration of advanced machine learning techniques, such as reinforcement learning, could further improve the model's adaptability to diverse surveillance environments. Second, incorporating additional data augmentation strategies and synthetic data generation could enhance the robustness of the detection framework. Third, expanding the application of the framework to other domains, such as traffic monitoring and anomaly effectiveness detection in different contexts. Furthermore, addressing challenges such as occlusion and varying illumination conditions could lead to a more robust detection performance. Investigating the use of multi-sensor data fusion, combining video data with other sensory inputs like LiDAR or thermal imaging, could also provide comprehensive situational imaging, as well as awareness in surveillance applications.

Author Contributions: Conceptualization, A.G.; methodology, A.G.; software, A.G.; validation, A.G. and A.A.; formal analysis, A.A. and A.G.; investigation, A.A. and A.G.; resources, A.A. and A.G.; data curation, A.G.; writing—original draft preparation, A.G.; writing—review and editing, A.G. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Gomaa, A.; Abdelwahab, M.M.; Abo-Zahhad, M.; Minematsu, T.; Taniguchi, R.I. Robust vehicle detection and counting algorithm employing a convolution neural network and optical flow. *Sensors* **2019**, *19*, 4588. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Gomaa, A.; Abdelwahab, M.M.; Abo-Zahhad, M. Efficient vehicle detection and tracking strategy in aerial videos by employing morphological operations and feature points motion analysis. *Multimed. Tools Appl.* **2020**, *79*, 26023–26043. [\[CrossRef\]](#)
3. Gomaa, A.; Minematsu, T.; Abdelwahab, M.M.; Abo-Zahhad, M.; Taniguchi, R.I. Faster CNN-based vehicle detection and counting strategy for fixed camera scenes. *Multimed. Tools Appl.* **2022**, *81*, 25443–25471. [\[CrossRef\]](#)
4. Gomaa, A.; Abdelwahab, M.M.; Abo-Zahhad, M. Real-time algorithm for simultaneous vehicle detection and tracking in aerial view videos. In Proceedings of the 2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), Windsor, ON, Canada, 5–8 August 2018; pp. 222–225.
5. Abdalrazik, A.; Gomaa, A.; Kishk, A.A. A wide axial-ratio beamwidth circularly-polarized oval patch antenna with sunlight-shaped slots for gnss and wimax applications. *Wirel. Netw.* **2022**, *28*, 3779–3786. [\[CrossRef\]](#)
6. Elbshbeshi, A.; Gomaa, A.; Mohamed, A.; Othman, A.; Ghazala, H. Seismic hazard evaluation by employing microtremor measurements for Abu Simbel area, Aswan, Egypt. *J. Afr. Earth Sci.* **2022**, *196*, 104734. [\[CrossRef\]](#)

7. Abdalrazik, A.; Gomaa, A.; Afifi, A. Multiband circularly-polarized stacked elliptical patch antenna with eye-shaped slot for GNSS applications. *Int. J. Microw. Wirel. Technol.* **2024**, 1–7. [\[CrossRef\]](#)
8. Salem, M.; Gomaa, A.; Tsurusaki, N. Detection of Earthquake-Induced Building Damages Using Remote Sensing Data and Deep Learning: A Case Study of Mashiki Town, Japan. In Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 2350–2353.
9. Ren, J.; Chen, Y. Multiple objects parameter detection in urban mixed traffic scene. *J. Transp. Inf. Safe* **2009**, 27, 47–54.
10. Unzueta, L.; Nieto, M.; Cortés, A.; Barandiaran, J.; Otaegui, O.; Sánchez, P. Adaptive multicue background subtraction for robust vehicle counting and classification. *IEEE Trans. Intell. Transp. Syst.* **2011**, 13, 527–540. [\[CrossRef\]](#)
11. Ren, J.; Chen, X.; Liu, J.; Sun, W.; Pang, J.; Yan, Q.; Tai, Y.W.; Xu, L. Accurate single stage detector using recurrent rolling convolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5420–5428.
12. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
14. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, 32, 1627–1645. [\[CrossRef\]](#)
15. Fu, Z.; Chen, Y.; Yong, H.; Jiang, R.; Zhang, L.; Hua, X.S. Foreground gating and background refining network for surveillance object detection. *IEEE Trans. Image Process.* **2019**, 28, 6077–6090. [\[CrossRef\]](#)
16. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
17. Akopov, A.S.; Beklaryan, L.A. Traffic Improvement in Manhattan Road Networks With the Use of Parallel Hybrid Biobjective Genetic Algorithm. *IEEE Access* **2024**, 12, 19532–19552. [\[CrossRef\]](#)
18. Javadi, S.; Rameez, M.; Dahl, M.; Pettersson, M.I. Vehicle classification based on multiple fuzzy c-means clustering using dimensions and speed features. *Procedia Comput. Sci.* **2018**, 126, 1344–1350. [\[CrossRef\]](#)
19. Jia, Y.; Zhao, L.; Ma, B. A hierarchical clustering-based routing protocol for wireless sensor networks supporting multiple data aggregation qualities. *Int. J. Sens. Netw.* **2008**, 4, 79–91. [\[CrossRef\]](#)
20. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, 128, 261–318. [\[CrossRef\]](#)
21. Shokri, D.; Larouche, C.; Homayouni, S. A Comparative Analysis of Multi-Label Deep Learning Classifiers for Real-Time Vehicle Detection to Support Intelligent Transportation Systems. *Smart Cities* **2023**, 6, 2982–3004. [\[CrossRef\]](#)
22. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
23. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016, Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
24. Babae, M.; Dinh, D.T.; Rigoll, G. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognit.* **2018**, 76, 635–649. [\[CrossRef\]](#)
25. Zhao, X.; Chen, Y.; Tang, M.; Wang, J. Joint background reconstruction and foreground segmentation via a two-stage convolutional neural network. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 343–348.
26. Chen, Y.; Wang, J.; Zhu, B.; Tang, M.; Lu, H. Pixelwise deep sequence learning for moving object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, 29, 2567–2579. [\[CrossRef\]](#)
27. Liao, J.; Guo, G.; Yan, Y.; Wang, H. Multiscale cascaded scene-specific convolutional neural networks for background subtraction. In *Advances in Multimedia Information Processing, Proceedings of the Pacific Rim Conference on Multimedia, Hefei, China, 21–22 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 524–533.
28. Wang, Y.; Luo, Z.; Jodoin, P.M. Interactive deep learning method for segmenting moving objects. *Pattern Recognit. Lett.* **2017**, 96, 66–75. [\[CrossRef\]](#)
29. Liang, X.; Liao, S.; Wang, X.; Liu, W.; Chen, Y.; Li, S.Z. Deep background subtraction with guided learning. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
30. Cygert, S.; Czyżewski, A. Vehicle detector training with labels derived from background subtraction algorithms in video surveillance. In Proceedings of the 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 19–21 September 2018; pp. 98–103.
31. Rosenberg, C.; Hebert, M.; Schneiderman, H. Semi-supervised self-training of object detection models. In Proceedings of the Seventh IEEE Workshops on Applications of Computer Vision/IEEE Workshop on Motion and Video Computing (WACV/MOTION'05), Breckenridge, CO, USA, 5–7 January 2005; Volume 1, pp. 29–36. [\[CrossRef\]](#)
32. Yu, T.; Yang, J.; Lu, W. Combining Background Subtraction and Convolutional Neural Network for Anomaly Detection in Pumping-Unit Surveillance. *Algorithms* **2019**, 12, 115. [\[CrossRef\]](#)

33. Mahmud, T.; Liu, C.H.; Yaman, B.; Marculescu, D. SSVOD: Semi-Supervised Video Object Detection with Sparse Annotations. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–6 January 2024; pp. 6773–6782.
34. Daneshfar, F.; Soleymanbaigi, S.; Yamini, P.; Amini, M.S. A survey on semi-supervised graph clustering. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108215. [[CrossRef](#)]
35. Shazeer, N. Glu variants improve transformer. *arXiv* **2020**, arXiv:2002.05202.
36. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A holistic approach to semi-supervised learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [[CrossRef](#)]
37. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
38. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
39. Yang, H.; Qu, S. Real-time vehicle detection and counting in complex traffic scenes using background subtraction model with low-rank decomposition. *IET Intell. Transp. Syst.* **2017**, *12*, 75–85. [[CrossRef](#)]
40. Xu, D.; Tian, Y. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [[CrossRef](#)]
41. Zhu, X.; Zhang, S.; Li, Y.; Zhang, J.; Yang, L.; Fang, Y. Low-rank sparse subspace for spectral clustering. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1532–1543. [[CrossRef](#)]
42. Zhu, X.; Zhang, S.; He, W.; Hu, R.; Lei, C.; Zhu, P. One-step multi-view spectral clustering. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 2022–2034. [[CrossRef](#)]
43. Wang, Y.; Jodoin, P.M.; Porikli, F.; Konrad, J.; Benezeth, Y.; Ishwar, P. CDnet 2014: An expanded change detection benchmark dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 387–394.
44. Wen, L.; Du, D.; Cai, Z.; Lei, Z.; Chang, M.C.; Qi, H.; Lim, J.; Yang, M.H.; Lyu, S. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.* **2020**, *193*, 102907. [[CrossRef](#)]
45. Mahto, P.; Garg, P.; Seth, P.; Panda, J. Refining YOLOv4 for vehicle detection. *Int. J. Adv. Res. Eng. Technol. (IJARET)* **2020**, *11*, 409–419.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.