

Article

CNN with New Spatial Pyramid Pooling and Advanced Filter-Based Techniques: Revolutionizing Traffic Monitoring via Aerial Images

Irfan Javid ¹, Rozaida Ghazali ², Waddah Saeed ^{3,*} , Tuba Batool ² and Ebrahim Al-Wajih ⁴ 

¹ Department of Computer Science and IT, University of Poonch, Rawalakot 12350, Pakistan; irfanjavid@upr.edu.pk

² Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat 86400, Malaysia; rozaida@uthm.edu.my (R.G.); hi210012@student.uthm.edu.my (T.B.)

³ School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, UK

⁴ Society Development & Continuing Education Center, Hodeidah University, Hodeidah P.O. Box 3114, Yemen; ebrahim.q.alwajih@gmail.com

* Correspondence: waddah.saeed@dmu.ac.uk

Abstract: The escalation in vehicular traffic, in conjunction with global population growth, has precipitated heightened road congestion, augmented air pollution, and a rise in vehicular accidents. Over the past decade, the global vehicular count has witnessed a substantial surge. In this context, traffic monitoring emerges as a paramount challenge, especially within developing nations. This research introduces an innovative system for vehicle detection and categorization aimed at intelligent traffic monitoring. The system utilizes a convolutional neural network-based U-Net model for the segmentation of aerial images. After segmentation, the outputs are further analyzed for vehicle identification. This vehicle detection utilizes an advanced spatial pyramid pooling (ASPP) mechanism which refines the spatial partitions of the image and captures intricate details, enhancing the accuracy and precision of the detection process. Detected vehicles are then categorized into distinct subcategories. For the effective management and control of high-density traffic flow, the extended Kalman filter (EKF) technique is employed, thereby reducing the reliance on human oversight. In experimental evaluations, our proposed model exhibits exemplary vehicle detection capabilities across the German Aerospace Center (DLR3K) and the Vehicle Detection in Aerial Imagery (VEDAI) datasets. Potential applications of the system encompass vehicle identification in traffic streams, traffic congestion assessment, intersection traffic density analysis, differentiation of vehicle types, and pedestrian pathway determination.

Keywords: spatial pyramid pooling; U-Net; semantic segmentation; extended Kalman filter; vehicle categorization and recognition



Citation: Javid, I.; Ghazali, R.; Saeed, W.; Batool, T.; Al-Wajih, E. CNN with New Spatial Pyramid Pooling and Advanced Filter-Based Techniques: Revolutionizing Traffic Monitoring via Aerial Images. *Sustainability* **2024**, *16*, 117. <https://doi.org/10.3390/su16010117>

Academic Editors: Kuo-Yi Lin, Kuang-Cheng Chai and Ke-Chiun Chang

Received: 30 November 2023

Revised: 15 December 2023

Accepted: 19 December 2023

Published: 22 December 2023



Copyright: © 2023 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

Vehicle recognition can be accomplished using different techniques and approaches, such as computer vision algorithms, machine learning models, radars, and sensor-based technologies [1–3]. Technologies such as radar and vision systems are widely employed for this objective. These solutions offer the advantage of furnishing precise and varied information, encompassing details such as speed and heading direction. However, the drawbacks include the typically substantial installation size and the computational power required [4–6]. In recent times, advancements in computing hardware performance, coupled with the development of sophisticated machine learning techniques, have facilitated the reliable recognition of vehicles in images or video streams through machine vision approaches. The utilization of vision-based traffic systems has gained popularity in recent years, particularly in the domains of traffic monitoring and the control of autonomous cars [7–9].

Implementing a traffic monitoring system presents a practical approach to alleviate traffic congestion. The key purpose of this system is to oversee and record essential traffic data, encompassing metrics like vehicle counts, vehicle types, and their respective speeds. To enhance the efficiency of the roadway system, forecast forthcoming transport needs, and enhance the well-being of travelers, a comprehensive traffic analysis is undertaken based on the collected data. However, the development, deployment, and maintenance of traffic monitoring systems typically incur substantial costs in many nations.

Incorporating cutting-edge surveillance technologies and aerial data is positioned to elevate the efficiency and effectiveness of traffic oversight and administration. This article delves into the dynamic realm of vehicle detection for traffic monitoring systems, capitalizing on aerial imagery acquired through a fusion of modern surveillance technologies. The primary aim of our investigation is to introduce a novel approach, involving initial image segmentation, followed by vehicle detection and classification into diverse categories, to enhance the efficiency of traffic management. The initial step involves capturing aerial images, which subsequently serve as input data for the process of semantic segmentation. After obtaining aerial images, the subsequent phase entails the implementation of an advanced spatial pyramid pooling (ASPP) mechanism within the system architecture. This sophisticated mechanism is meticulously designed to optimize the accuracy and efficiency of vehicle detection within the segmented image, ensuring a more robust and precise identification process. Detected vehicles are then categorized into distinct subcategories. The vehicles that have been detected are then tracked using a specific approach in the final stage of the procedure. In order to provide an extensive and accurate vehicle monitoring method, this entails applying tracking mechanisms, with a focus on the extended Kalman filter-based technique. Moreover, the model presented in this study undergoes validation through a series of experiments conducted on two distinct datasets: the German Aerospace Center (DLR3K) dataset and the Vehicle Detection in Aerial Imagery (VEDAI) dataset. The outcomes of these experiments showcase a notable level of accuracy in both vehicle detection and classification. Importantly, the achieved results surpass the performance of other contemporary advanced methodologies in the field. The major contributions of this work can be delineated as follows:

- Our study presents a cutting-edge hybrid framework employing spatial pyramid pooling within a convolutional neural network (CNN) for streamlined detection, classification, and tracking of vehicles on roads, enhancing transportation system management in diverse settings.
- Vehicle tracking is achieved through the application of extended Kalman filter (EKF)-based methods, a sophisticated variant of the traditional Kalman filter.
- Compared to existing techniques, our approach demonstrates substantial improvements in the detection and classification of vehicles.

2. Related Work

One of the main problems in the fields of science, engineering, and technology study is object detection. It is the accurate localization of items in two dimensions, space and time, along with the use of computer vision technologies to classify these objects. Significant advancements in computer vision and machine learning, especially when it comes to the identification and classification of vehicles, have prompted a number of academic studies.

Within the domain of transportation engineering, the assessment of vehicle characteristics often necessitates classification tasks based on collected vehicle data. Over the last few decades, machine learning (ML) techniques have emerged as powerful tools for data classification [10–12]. Typically, the gathered vehicle data require initial stratification into various classes, followed by the gradual development of the ML model through optimization algorithms. By introducing a new vehicle sample, the model can undergo training to predict its class. Research endeavors have adhered to the outlined framework below.

The decision tree underwent cross-validation for the identification of seven radio-controlled cars within the laboratory, considering the trade-off between accuracy and computational

cost [13]. The experiment yielded an accuracy rate of approximately 100%. Furthermore, a vehicle classification system was devised for field road trials, revealing the reliable capability of the support vector machine (SVM) algorithm in detecting small, medium, and combined vehicles [14]. Additionally, an artificial neural network (ANN) was formulated for axle number-based categorization, employing a dataset comprising 9000 records [15]. Leveraging the principle components analysis method for key properties extraction, the model successfully classified five distinct types of specified automobiles. Furthermore, a study encompassing the k-nearest neighbors (KNN), SVM, and ANN algorithms was conducted to discern various vehicle classifications [13], aligning with the Federal Highway Administration (FHWA) vehicle classification guidelines. There are various challenges faced in vehicle classification with traditional machine learning: (1) Designing features that accurately describe vehicle characteristics can be difficult in typical machine learning and may require domain expertise. Feature engineering is an essential phase, and it can be difficult to manually craft relevant characteristics for precise categorization. (2) Vehicle images and data may contain complex and non-linear patterns that are difficult for traditional machine learning methods to identify. (3) Conventional machine learning models might not be able to adjust adequately to changes in environment, illumination, and weather. (4) Conventional machine learning models could find it difficult to scale effectively when the categorization task becomes more complex.

Contemporary deep learning (DL) techniques have seen widespread application in categorization tasks in recent times [16]. Among these, convolutional neural networks (CNNs) have garnered substantial research focus. In contrast to traditional methods, CNNs demonstrate notably enhanced precision in classification tasks, owing to their ability to extract both local and multi-level features [17–19]. Primarily employed for images—considered as discrete signals with two dimensions—CNNs have been utilized for isolating automobiles from the background through the application of class labels, employing object detection techniques based on CNNs [20,21]. Vehicle spacing data are retrieved when calibrated using real-world coordination systems [22]. By combining the clustering method based on image processing, the axles are identified, leading to the determination of the axle type [23]. Deep learning-based techniques produce better feature representation than hand-crafted features and faster processing times than sliding window-based techniques by utilizing convolutional neural networks (CNNs). One-step and two-step object detectors are the two primary categories of CNN-based detectors. Tables 1 and 2 showcase a collection of one-step and two-step object detection and classification algorithms, respectively.

Table 1. CNN-based one-step object detection and classification algorithms.

Algorithm	Techniques	Improvements	References
RetinaNet	Focal loss for classification balance, ResNet, and FPN.	Solves class imbalance; ResNet as backbone and FPN for feature extraction.	[24]
SSD	Feedforward convolutional architecture and NMS stage.	VGG16 as backbone and multiple convolutional layers for detection.	[25]
YOLO	One-step detection, grid-based, and limited region proposals.	CNN-based; grid cells for bounding box and class predictions.	[25]
YOLOv2	Improved precision and speed; six tasks for detection.	Batch normalization, anchor box convolution, and multi-scale training.	[26]
YOLOv3	DarkNet53 model for feature extraction; multi-label classification.	Multi-size feature maps for Bbox prediction.	[27]
YOLOv4	Combines properties of YOLO, YOLOv2, and YOLOv3; residual system.	SPPNet and PANet for feature fusion; CSPDarkNet53 backbone.	[28]
YOLOv5	CSPDarkNet as backbone and has better capability for small objects.	CSPNet architecture, improved accuracy and speed, and Fcos approach for frame selection area.	[29]

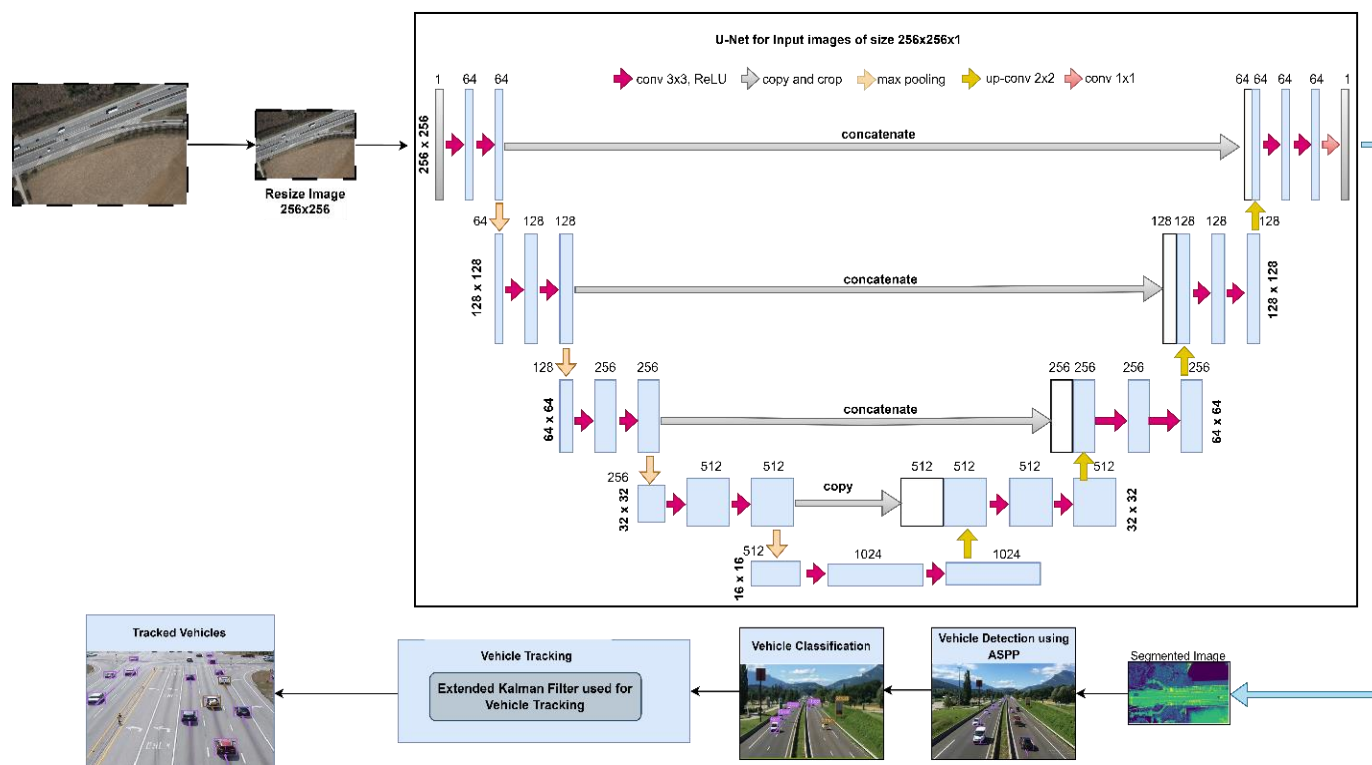
Table 2. CNN-based two-step object detection and classification algorithms.

Algorithm	Techniques	Improvements	References
R-CNN	Region-based ConvNet	AlexNet backbone, Selective search, SVM, and FC.	[30]
Fast R-CNN	Region-based ConvNet	Streamlined training, ROI pooling, and Bbox regressors.	[31]
Faster R-CNN	Region Proposal Network (RPN)	RPN replaces selective search; anchors for proposals.	[32]
R-FCN	Region-based ConvNet	Improved sharing of convolutional parameters; position-sensitive score maps.	[33]

The goal of this research is to advance machine vision technology in the contemporary environment. Detecting vehicles and monitoring traffic are the main goals of the technology that we use in order to manage large-scale transportation. Furthermore, we want to outperform current traffic monitoring and automotive detection systems in terms of both performance and outcomes. To achieve the highest level of vehicle detection accuracy, we want to experiment with deep learning algorithms.

3. Methodology

The initial phase involves the conversion of traffic data videos into a sequential series of frames. Each frame undergoes a segmentation process sequentially until the last frame is reached. Subsequently, the segmented images undergo thorough analysis for vehicle detection, utilizing the advanced spatial pyramid pooling (ASPP) technique. The identified vehicles then undergo classification into distinct vehicle categories. For tracking the detected and classified vehicles, the extended Kalman filter-based vehicle tracking approach is implemented. Figure 1 illustrates the workflow of vehicle detection and classification.

**Figure 1.** The workflow of vehicle detection and classification.

3.1. Preprocessing

In order to improve the efficiency of vehicle identification, tracking, and traffic surveillance, we first started the procedure by transforming the video into a series of separate images, or frames, for further examination. After the traffic video's frames were extracted, different kinds of noise were thoroughly examined. A variety of filtering techniques (Gaussian Filter, Median Filter, and Bilateral Filter) were carefully applied to these noise variations in order to minimize their influence. The kinds of noise that were examined encompassed those that were added during the video acquiring process, ambient interference, and inherent irregularities. To successfully reduce its impact on the image quality, a filtering strategy was required. We were able to determine and choose the best filter for noise using this analysis, which guaranteed the best outcome. This detailed approach was designed to carefully refine the preprocessing steps and establish a strong foundation for effective vehicle detection and classification. The effects of filtration on the extracted images are illustrated in Figure 2.

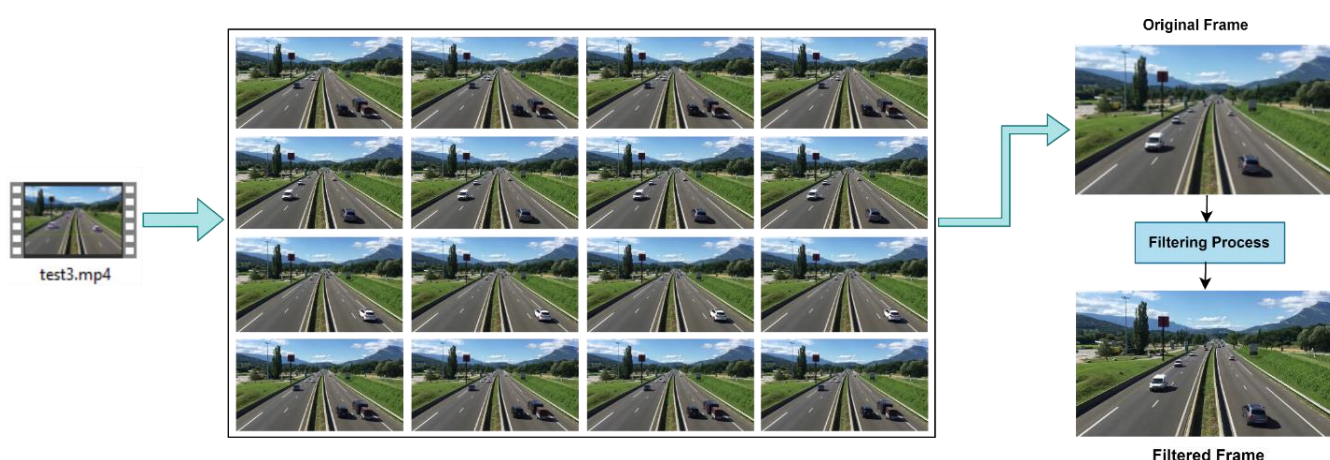


Figure 2. Frame extraction and filtering process.

3.2. U-Net-Based Semantic Segmentation

3.2.1. U-Net Architecture

In our U-Net based semantic segmentation approach, we employed a deep convolutional neural network architecture that is well suited for image segmentation tasks. A symmetric growing path allows for exact localization, while a contracting path captures context. These two paths make up the U-Net design. A sequence of convolutional and pooling layers make up the contracting path, which gradually reduces spatial dimensions while increasing the number of channels. To retrieve spatial information, the expanding path concatenates feature maps from the contracting path with transposed convolutions. By facilitating the merging of low-level and high-level features, skip connections improve the accuracy of segmentation. Throughout the network, rectified linear unit (ReLU) activation functions and batch normalization are used to foster convergence.

3.2.2. Training Process

Using a stochastic gradient descent (SGD) optimizer, the model parameters were optimized during a predetermined number of epochs as part of the training procedure. Effective convergence during training required a proper learning rate. Following extensive testing, a learning rate of 0.001 was found to be the ideal combination since it offered an appropriate balance between avoiding overshooting the minimum loss and allowing for fast convergence. Over a span of 50 epochs, the model was trained, enabling iterative learning and improvement in its segmentation abilities. It was essential to monitor the loss function at every epoch in order to evaluate convergence and avoid over fitting.

Impressive semantic segmentation results were obtained by combining the U-Net architecture with 50 training epochs and a learning rate of 0.001, which allowed for the exact delineation of objects and regions within the images. Figure 3 illustrates the U-Net architecture used for image segmentation.

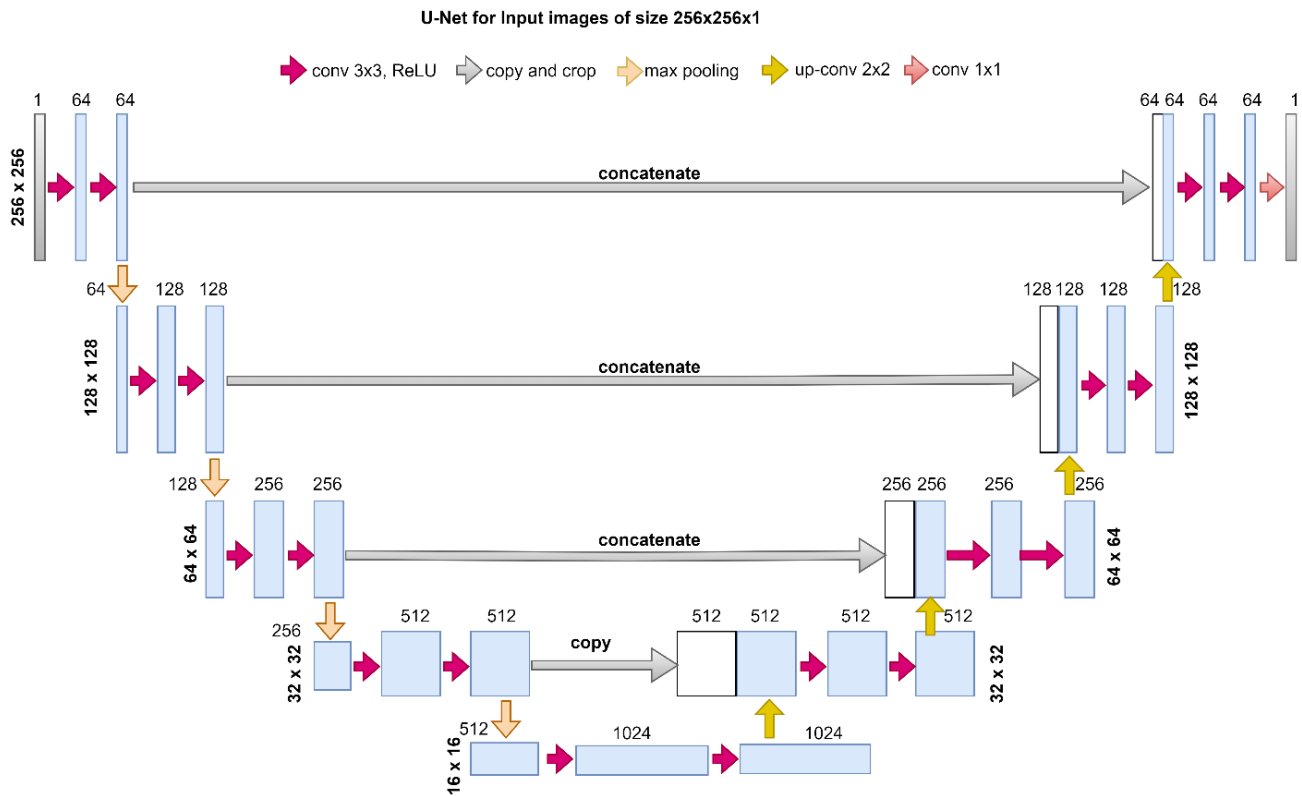


Figure 3. U-Net architecture implemented for image segmentation.

3.3. Advanced Spatial Pyramid Pooling-Based Vehicle Detection

Local ambiguity in image interpretation has been the focus of several investigations, as evidenced by studies carried out on different datasets. Interestingly, the FusionNet architecture [34] efficiently exploited the synergy between local features and global context within datasets such as ImageNet [35] and ADE20K [36]. The highlighted studies showcase an inventive integration that emphasizes the value of utilizing contextual information to improve the feature set. This, in turn, leads to more reliable image analysis and interpretation in different situations. However, in a more complex circumstance, it falls short. Using spatial pyramid pooling, an efficient object recognition technique, the pyramid pooling network (PPN) [37] combines many subregions to provide contextual insights on a large scale. By integrating four subregion pyramid pooling module scales—among which is a crucial global pooling layer—PPN achieves a comprehensive comprehension of contextual subtleties. Simultaneously, the non-overlapping pooling layers take up bins of varying dimensions, ensuring a consistent stride and kernel size throughout. This careful arrangement guarantees a subtle, yet coherent, method for gathering contextual data. The use of non-overlapping pooling reduces the spatial size of the feature map, which is achieved by dividing it by the size of the kernel. However, in order for this module to function correctly, the input feature map must align in terms of being a multiple of the kernel size. The result of pooling and subsequent increase in samples can cause alignment issues. In a given scenario when several kernel sizes are used, establishing alignment compatibility is critical for the non-overlapping pooling module to function properly. To maintain seamless integration, the input feature map must be a multiple of the aggregate sizes of these ker-

nels, avoiding any potential alignment inconsistencies during the pooling and subsequent up-sampling phases.

In contrast to the non-overlapping pooling module's limitations, the advanced spatial pyramid pooling (ASPP) approach is more versatile and efficient. ASPP adds flexibility to its architecture by incorporating variability in both levels and kernels, which are treated as hyperparameters strategically. This design option enables ASPP to adapt to a variety of situations, increasing its versatility in dealing with varying spatial scales and receptive fields. The first layer collects global information by generating a unified output region, while the remaining three layers (overlapping pooling layers) capture local features. In overlapping pooling layers, the input feature map maintains a fixed size with constant stride and padding. Before implementing the ASPP, non-overlapping pooling with a small kernel is used to minimize the feature map. For compatibility, the pooling process uses only the maximum operation, choosing the highest value in the pool to highlight salient characteristics in the area. The ASPP module exemplifies consistency by accepting input feature maps of various sizes through spatial pooling. Figures 4 and 5 show the results of vehicle detection in selected images from the VEDAI and DLR3K datasets.



Figure 4. Detection of vehicles across VEDAI dataset images.

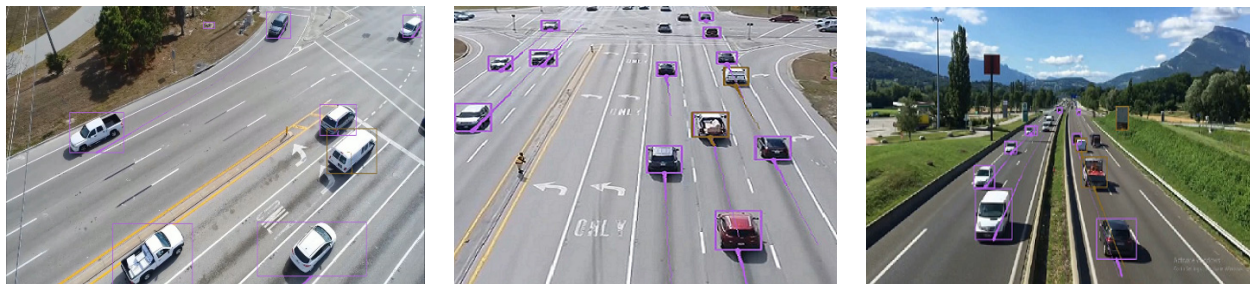


Figure 5. Detection of vehicles across DLR3K dataset images.

3.4. Extended Kalman Filter (EKF)

For some systems with significant non-linearity, the precision of the Kalman filtering process is limited [38]. In order to overcome this constraint, the extended Kalman filter linearizes the non-linear models by linearizing them around the current estimate using a first-order Taylor expansion. By calculating the non-linear functions' Jacobian matrices, the system inside the filter is approximated linearly through the process of linearization. Typically, the following formula can be used to represent both the level of uncertainty or randomness and stochastic discrete-time state space system:

$$m_i = f(m_{i-1}, n_{i-1}) \quad (1)$$

$$z_i = h(m_i) \quad (2)$$

Taylor series can be used to augment the extended Kalman filtering process in the non-linear time state equation and randomness Equations (1) and (2). The specific method for linear approximation of non-linear Equation (1) is to estimate the non-linear system (i.e., the Jacobian matrix) after linearizing it using the state equation and measurement equation. The EKF method's particular steps are listed below.

$$\begin{aligned} m_{i|i-1} &= f(m_{i-1|i-1}, n_{i-1}) \\ Q_{i|i-1} &= X_i Q_{i-1|i-1} X_i^T \end{aligned} \quad (3)$$

The stochastic discrete-time state process of EKF is calculated by Equation (3).

$$\begin{aligned} E_i &= Q_{i|i-1} U_i^T (U_i Q_{i|i-1} U_i^T + W_i S_i M_i^T)^{-1} \\ m_{i|i} &= m_{i|i-1} + E_i (z_i - h(m_{i|i-1})) \\ Q_{i|i} &= (K - E_i U_i) Q_{i|i-1} \end{aligned} \quad (4)$$

The measurement update process of EKF is calculated by Equation (4).

$$X(I|I-1)_{[l,t]} = \frac{\partial f_{[l]}}{\partial x_{[t]}}(m_{i-1|i-1}, n_{i-1}) \quad (5)$$

In the above formula, X is the Jacobian matrix of f to x .

$$U(I)_{[l,t]} = \frac{\partial f_{[l]}}{\partial x_{[t]}}(m_{i|i-1}) \quad (6)$$

In the above formula, U is the Jacobian matrix of f to x .

4. Experimental Setup

4.1. Description of Datasets

We have taken into consideration two intricate aerial imaging datasets—VEDAI and DLR3K—during our research. The following is a description of these datasets.

4.1.1. Vehicle Detection in Aerial Imagery (VEDAI)

The VEDAI (Vehicle Detection in Aerial Imagery) dataset is provided for the specific goal of identifying vehicles in images captured from aerial perspectives. The realistic and unrestricted environments are the source of the images in the VEDAI collection. The VEDAI dataset contains images in different resolutions. The image resolutions are 1024×1024 and 512×512 , respectively [39]. The VEDAI collection includes a variety of backgrounds, including cities, highways, buildings, trees, and so forth. The VEDAI dataset encompasses nine vehicle categories, namely, “bus”, “truck”, “car”, “pick-up”, “camping car”, “tractor”, “plane”, “boat”, and “other”. The predominant focus in VEDAI is on “small land vehicles”, which includes targets like “car”, “pick-up”, “tractor”, “truck”, and “bus”. These specific categories are employed for evaluating the proposed framework discussed in this section.

4.1.2. DLR3K Dataset

Using the DLR 3K dataset, we evaluated the target detection models [34]. Aerial images of cars with a ground sampling distance of roughly 13 cm are included in this dataset. In contrast to the VEDAI dataset, the DLR 3K dataset features a background that is more consistent, with large expanses of urban streets and lawns. The dataset consists of just 20 images at a resolution of 5616×3744 . It is divided into two vehicle types, “car” and “truck”, with the majority of the samples being “car” ones. Every original image is cropped to 256×256 to facilitate model training. This cropping procedure results in 64 cropped

images of the same size for each original image, contributing to a total of 1280 cropped images. We eventually obtained 890 cropped images after eliminating cropped images without targets and keeping just images with the goal “car”. For training, we utilized 620 cropped images, and for testing, we used the remaining 270 images.

4.2. Performance Metrics

In this work, we placed a significant emphasis on employing various metrics to comprehensively evaluate the performance of our model. Among the key metrics utilized, accuracy, recall, and F1 Score stood out as fundamental measures. The deployment of these metrics allowed us to deeply assess the models’ effectiveness in correctly predicting outcomes, identifying true positives, and balancing between precision and recall. These metrics are calculated using the following formulas:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}} \quad (7)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (8)$$

$$\text{F1 - Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

5. Results and Analysis

In this section, we evaluated the recognition and categorization accuracy of the anticipated model on established datasets, ensuring its performance comparison with existing approaches in the field. The proposed model underwent evaluation on two benchmark datasets, namely, DLR-3K and VEDAI. A comprehensive set of metrics, including accuracy, recall, and F1 Score, was computed. Table 3 presents a detailed overview of these measures concerning the VEDAI dataset. To maintain fairness in the assessment, a comparable set of concealed samples from the test data was employed. The results showcased outstanding performance, surpassing current state-of-the-art methods.

Table 3. Comparable statistics on VEDAI dataset.

Vehicle Class	Accuracy	Recall	F1 Score
Car	96.12	88.21	92.71
Bus	97.66	94.33	89.45
Tractor	89.24	92.47	93.24
Pick-up	92.14	89.36	94.83
Truck	98.01	91.57	90.52

Experiments were carried out to demonstrate the efficacy of our anticipated model. The outcomes are illustrated by Figure 6, with the confusion matrix classifying vehicles across the VEDAI dataset. The presented data in Figure 6 distinctly indicate the noteworthy performance of the anticipated model, attaining a mean accuracy of 95.4% in classification across the VEDAI dataset. Additionally, the illustration underscores that the category of cars demonstrated the highest accuracy in classification, while the tractor recorded the lowest accuracy in the classification list.

Our model underwent evaluation, and a comparative analysis was conducted against existing methods in the literature. Table 4 presents a comparison of detection accuracies across the DLR3K dataset, while Table 5 displays a contrast in classification accuracies between the proposed model and state-of-the-art approaches on the VEDAI dataset. As shown in both tables, our proposed model achieved superior performance compared to state-of-the-art approaches.

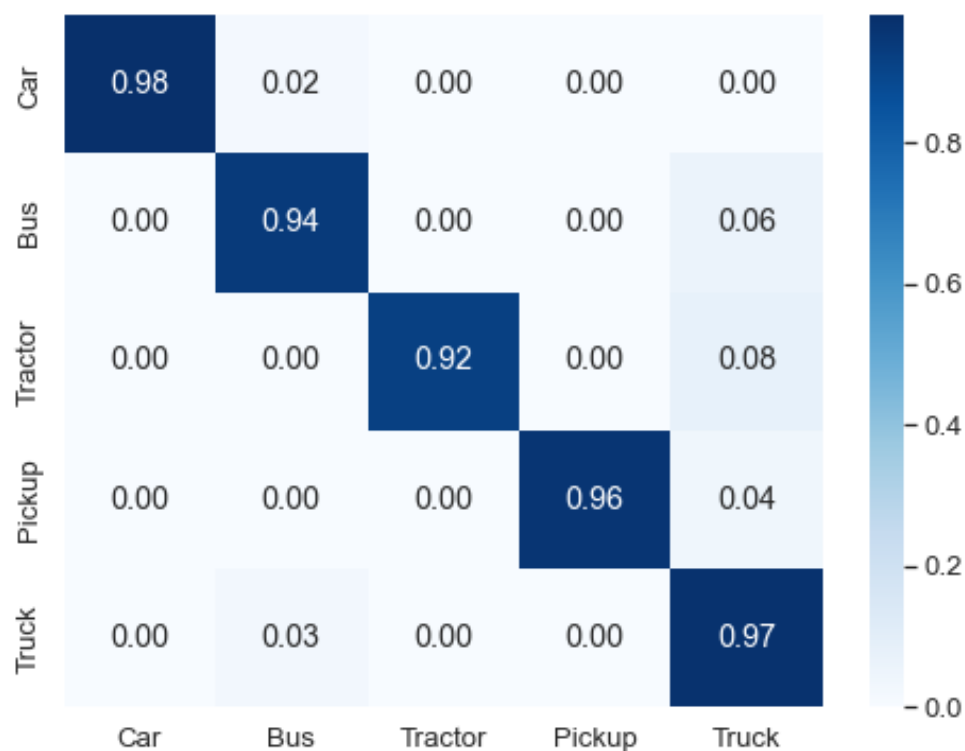


Figure 6. Confusion matrix for vehicle classification across the VEDAI dataset.

Table 4. Comparison of proposed model with other state-of-art approaches on DLR3K dataset.

Technique	Accuracy (%)	Study
Recurrent-Feature Aware Visualization (RFAV)	56.24	[40]
CNN based on Tectofugal–thalamofugal–accessory optic system	84.25	[41]
Customized Pyramid Pooling Method (CPPM)	93.13	[42]
Proposed Model	93.8	

Table 5. Comparison of proposed model with other state-of-art approaches on VEDAI dataset.

Technique	Accuracy (%)	Study
Recurrent-Feature Aware Visualization (RFAV)	51.95	[40]
Customized Pyramid Pooling Method (CPPM)	92.22	[42]
Feature-balanced pyramid network (FBPN)	91.27	[43]
Proposed Model	95.4	

6. Conclusions

This study introduces a framework for recognizing vehicles in aerial images. The proposed paradigm holds promise for applications in smart surveillance systems, intelligent traffic management, and advanced traffic monitoring. The enhanced traffic monitoring system demonstrates improved efficiency in vehicle detection, leveraging the innovative spatial pyramid pooling module. Prior to applying this cutting-edge module for identifying various vehicles in aerial images, the initial module efficiently segments the images. Then, the extended Kalman filter is used to track the classified vehicles.

Our methodology's effectiveness is evaluated not only using the VEDAI and DRL3K datasets but also through experimental results that highlight the relevance of the suggested approach compared to other state-of-the-art methodologies.

The datasets selected for the experiments are dynamic, diverse, and complex, encompassing various types of vehicles. These scenes are captured in both urban and rural settings. Due to the dynamic nature of the scenarios, characterized by disordered vehicle information and crowded backgrounds, our proposed detection module ASPP exhibited varying performance across all datasets. The diminished effectiveness of our proposed system in discerning heavily occluded cars represents a notable limitation. This challenge is particularly pronounced when more than 60% of a vehicle is concealed by external factors, such as buildings and trees. Specifically, the system encounters difficulty in recognizing cars that are partially obscured by buildings and trees, despite its capability to detect vehicles under such conditions. In instances where this occurs, the system may face challenges in accurately identifying and categorizing these substantially obscured vehicles. In our upcoming research endeavors, we aim to improve the efficiency of surveillance-based vehicle tracking. This will involve implementing a comprehensive end-to-end deep learning approach for overall traffic monitoring, covering both vehicle detection and tracking.

Author Contributions: Conceptualization, I.J., R.G., W.S. and T.B.; data curation, I.J.; formal analysis, I.J.; funding acquisition, R.G. and W.S.; investigation, I.J. and T.B.; methodology, I.J., R.G. and W.S.; project administration, R.G.; resources, I.J.; software, I.J.; supervision, R.G.; validation, I.J. and R.G.; visualization, I.J. and E.A.-W.; writing—original draft, I.J. and T.B.; writing—review and editing, I.J., R.G., W.S. and E.A.-W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The VEDAI dataset is publicly available at <https://downloads.greyc.fr/vedai/> (accessed on 1 November 2023). And the DLR3K dataset is publicly available at https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12760/22294_read-52777 (accessed on 1 November 2023).

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper. The research work and its findings were carried out in an unbiased manner, without any financial or personal relationships that could potentially influence the interpretation or presentation of the results. The authors have no affiliations with any organization or entity that might have a direct or indirect interest in the research conducted in this paper. The authors are committed to maintaining the highest standards of scientific integrity and transparency, ensuring that the research outcomes are presented objectively and without any bias or external influence.

References

1. Chen, Y.; Li, Z. An Effective Approach of Vehicle Detection Using Deep Learning. *Comput. Intell. Neurosci.* **2022**, 2022, 2019257. [CrossRef] [PubMed]
2. Punyavathi, G.; Neeladri, M.; Singh, M.K. Vehicle tracking and detection techniques using IoT. *Mater. Today Proc.* **2022**, 51, 909–913. [CrossRef]
3. Liu, Q.; Li, Z.; Yuan, S.; Zhu, Y.; Li, X. Review on Vehicle Detection Technology for Unmanned Ground Vehicles. *Sensors* **2021**, 21, 1354. [CrossRef] [PubMed]
4. Sohail, M.; Khan, A.U.; Sandhu, M.; Shoukat, I.A.; Jafri, M.; Shin, H. Radar sensor based machine learning approach for precise vehicle position estimation. *Sci. Rep.* **2023**, 13, 13837. [CrossRef] [PubMed]
5. Veeraraghavan, H.; Masoud, O.; Papanikolopoulos, N.P. Computer vision algorithms for intersection monitoring. *IEEE Trans. Intell. Transp. Syst.* **2003**, 4, 78–89. [CrossRef]
6. Jain, S.; Jain, S.S. Development of Intelligent Transportation System and Its Applications for an Urban Corridor During COVID-19. *J. Inst. Eng. Ser. B* **2021**, 102, 1191–1200. [CrossRef]
7. Pavel, M.I.; Tan, S.Y.; Abdullah, A. Vision-Based Autonomous Vehicle Systems Based on Deep Learning: A Systematic Literature Review. *Appl. Sci.* **2022**, 12, 6831. [CrossRef]
8. Datondji, S.R.E.; Dupuis, Y.; Subirats, P.; Vasseur, P. A Survey of Vision-Based Traffic Monitoring of Road Intersections. *IEEE Trans. Intell. Transp. Syst.* **2016**, 17, 2681–2698. [CrossRef]
9. Huang, Q.; Liu, J. Practical limitations of lane detection algorithm based on Hough transform in challenging scenarios. *Int. J. Adv. Robot. Syst.* **2021**, 18, 1–13. [CrossRef]
10. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, 2, 420. [CrossRef]

11. Ali, M.; Dewan, A.; Sahu, A.K.; Taye, M.M. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers* **2023**, *12*, 91. [\[CrossRef\]](#)
12. Liu, J.; Huang, Q.; Ulishney, C.; Dumitrescu, C.E. Machine learning assisted prediction of exhaust gas temperature of a heavy-duty natural gas spark ignition engine. *Appl. Energy* **2021**, *300*, 117413. [\[CrossRef\]](#)
13. Vasconcellos, B.R.; Rudek, M.; de Souza, M.A. Machine Learning Method for Vehicle Classification by Inductive Waveform Analysis. *IFAC-Pap.* **2020**, *53*, 13928–13932. [\[CrossRef\]](#)
14. Al-Tarawneh, M.; Huang, Y.; Lu, P.; Tolliver, D. Vehicle Classification System Using In-Pavement Fiber Bragg Grating Sensors. *IEEE Sens. J.* **2018**, *18*, 2807–2815. [\[CrossRef\]](#)
15. Yan, L.; Fraser, M.; Elgamal, A.; Fountain, T.; Oliver, K. Neural Networks and Principal Components Analysis for Strain-Based Vehicle Classification. *J. Comput. Civ. Eng.* **2008**, *22*, 123–132. [\[CrossRef\]](#)
16. Sultana, J.; Rani, M.U.; Farquad, M.A.H. An Extensive Survey on Some Deep Learning Applications. In Proceedings of the 2018 IADS International Conference on Computing, Communications & Data Engineering (CCODE), Shanghai, China, 4–6 May 2018; pp. 1–8. [\[CrossRef\]](#)
17. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [\[CrossRef\]](#)
18. Krichen, M. Convolutional Neural Networks: A Survey. *Computers* **2023**, *12*, 151. [\[CrossRef\]](#)
19. Alahmari, F.; Naim, A.; Alqahtani, H. *E-Learning Modeling Technique and Convolution Neural Networks in Online Education*; River Publishers: Copenhagen, Denmark, 2023. [\[CrossRef\]](#)
20. Tarmizi, I.A.; Aziz, A.A. Vehicle Detection Using Convolutional Neural Network for Autonomous Vehicles. In Proceedings of the 2018 International Conference on Intelligent and Advanced System (ICIAS), Kuala Lumpur, Malaysia, 13–14 August 2018. [\[CrossRef\]](#)
21. Galvez, R.L.; Bandala, A.A.; Dadios, E.P.; Vicerra, R.R.P.; Maningo, J.M.Z. Object Detection Using Convolutional Neural Networks. In Proceedings of the TENCON 2018—2018 IEEE Region 10 Conference, Jeju, Republic of Korea, 28–31 October 2018. [\[CrossRef\]](#)
22. Ojala, R.; Vepsäläinen, J.; Pirhonen, J.; Tammi, K. Infrastructure camera calibration with GNSS for vehicle localisation. *IET Intell. Transp. Syst.* **2023**, *17*, 341–356. [\[CrossRef\]](#)
23. Yao, Z.; Wei, H.; Li, Z.; Corey, J. Fuzzy C-Means Image Segmentation Approach for Axle-Based Vehicle Classification. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, *2595*, 68–77. [\[CrossRef\]](#)
24. Zhou, Z.; Li, Y.; Peng, C.; Wang, H.; Du, S. Image Processing: Facilitating Retinanet for Detecting Small Objects. *J. Phys. Conf. Ser.* **2021**, *1815*, 012016. [\[CrossRef\]](#)
25. Li, Z.; Dong, Y.; Wen, Y.; Xu, H.; Wu, J. A Deep Pedestrian Tracking SSD-Based Model in the Sudden Emergency or Violent Environment. *J. Adv. Transp.* **2021**, *2021*, 2085876. [\[CrossRef\]](#)
26. Sang, J.; Wu, Z.; Guo, P.; Hu, H.; Xiang, H.; Zhang, Q.; Cai, B. An Improved YOLOv2 for Vehicle Detection. *Sensors* **2018**, *18*, 4272. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Zhang, X.; Han, F.; Wang, P.; Jiang, W.; Wang, C. DC-YOLOv3: A novel efficient object detection algorithm. *J. Phys. Conf. Ser.* **2021**, *2082*, 012012. [\[CrossRef\]](#)
28. Liu, Q.; Fan, X.; Xi, Z.; Yin, Z.; Yang, Z. Object detection based on Yolov4-Tiny and Improved Bidirectional feature pyramid network. *J. Phys. Conf. Ser.* **2022**, *2209*, 012023. [\[CrossRef\]](#)
29. Zhang, Y.; Guo, Z.; Wu, J.; Tian, Y.; Tang, H.; Guo, X. Real-Time Vehicle Detection Based on Improved YOLO v5. *Sustainability* **2022**, *14*, 12274. [\[CrossRef\]](#)
30. Fang, S.; Zhang, B.; Hu, J. Improved Mask R-CNN Multi-Target Detection and Segmentation for Autonomous Driving in Complex Scenes. *Sensors* **2023**, *23*, 3853. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Alam, M.K.; Ahmed, A.; Salih, R.; Al Asmari, A.F.S.; Khan, M.A.; Mustafa, N.; Mursaleen, M.; Islam, S. Faster RCNN based robust vehicle detection algorithm for identifying and classifying vehicles. *J. Real-Time Image Process.* **2023**, *20*, 93. [\[CrossRef\]](#)
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [\[CrossRef\]](#)
33. Li, S.; Chen, J.; Peng, W.; Shi, X.; Bu, W. A vehicle detection method based on disparity segmentation. *Multimed. Tools Appl.* **2023**, *82*, 19643–19655. [\[CrossRef\]](#)
34. Quan, T.M.; Hildebrand, D.G.C.; Jeong, W.K. FusionNet: A Deep Fully Residual Convolutional Neural Network for Image Segmentation in Connectomics. *Front. Comput. Sci.* **2021**, *3*, 613981. [\[CrossRef\]](#)
35. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255. [\[CrossRef\]](#)
36. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ADE20K dataset. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5122–5130. [\[CrossRef\]](#)
37. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239. [\[CrossRef\]](#)
38. Huang, Y.; Bao, C.; Wu, J.; Ma, Y. Estimation of Sideslip Angle Based on Extended Kalman Filter. *J. Electr. Comput. Eng.* **2017**, *2017*, 5301602. [\[CrossRef\]](#)

39. Li, K.; Wang, B. DAR-Net: Dense Attentional Residual Network for Vehicle Detection in Aerial Images. *Comput. Intell. Neurosci.* **2021**, *2021*, 6340823. [[CrossRef](#)] [[PubMed](#)]
40. Mandal, M.; Shah, M.; Meena, P.; Devi, S.; Vipparthi, S.K. AVDNet: A Small-Sized Vehicle Detection Network for Aerial Visual Data. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 494–498. [[CrossRef](#)]
41. Wang, Z.; Liu, D.; Lei, Y.; Niu, X.; Wang, S.; Shi, L. Small target detection based on bird's visual information processing mechanism. *Multimed. Tools Appl.* **2020**, *79*, 22083–22105. [[CrossRef](#)]
42. Rafique, A.A.; Al-Rasheed, A.; Ksibi, A.; Ayadi, M.; Jalal, A.; Alnowaiser, K.; Meshref, H.; Shorfuzzaman, M.; Gochoo, M.; Park, J. Smart Traffic Monitoring Through Pyramid Pooling Vehicle Detection and Filter-Based Tracking on Aerial Images. *IEEE Access* **2022**, *11*, 2993–3007. [[CrossRef](#)]
43. Wang, B.; Gu, Y. An Improved FBPN-Based Detection Network for Vehicles in Aerial Images. *Sensors* **2020**, *20*, 4709. [[CrossRef](#)]

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.