remote sensing

MDPI

Article

Orientation- and Scale-Invariant Multi-Vehicle Detection and Tracking from Unmanned Aerial Videos

Jie Wang, Sandra Simeonova and Mozhdeh Shahbazi *

Department of Geomatics Engineering, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada

* Correspondence: mozhdeh.shahbazi@ucalgary.ca

Received: 29 July 2019; Accepted: 11 September 2019; Published: 16 September 2019

Abstract: Along with the advancement of light-weight sensing and processing technologies, unmanned aerial vehicles (UAVs) have recently become popular platforms for intelligent traffic monitoring and control. UAV-mounted cameras can capture traffic-flow videos from various perspectives providing a comprehensive insight into road conditions. To analyze the traffic flow from remotely captured videos, a reliable and accurate vehicle detection-and-tracking approach is required. In this paper, we propose a deep-learning framework for vehicle detection and tracking from UAV videos for monitoring traffic flow in complex road structures. This approach is designed to be invariant to significant orientation and scale variations in the videos. The detection procedure is performed by fine-tuning a state-of-the-art object detector, You Only Look Once (YOLOv3), using several custom-labeled traffic datasets. Vehicle tracking is conducted following a tracking-by-detection paradigm, where deep appearance features are used for vehicle re-identification, and Kalman filtering is used for motion estimation. The proposed methodology is tested on a variety of real videos collected by UAVs under various conditions, e.g., in late afternoons with long vehicle shadows, in dawn with vehicles lights being on, over roundabouts and interchange roads where vehicle directions change considerably, and from various viewpoints where vehicles' appearance undergo substantial perspective distortions. The proposed tracking-by-detection approach performs efficiently at 11 frames per second on color videos of 2720p resolution. Experiments demonstrated that high detection accuracy could be achieved with an average F1-score of 92.1%. Besides, the tracking technique performs accurately, with an average multiple-object tracking accuracy (MOTA) of 81.3%. The proposed approach also addressed the shortcomings of the state-of-the-art in multi-object tracking regarding frequent identity switching, resulting in a total of only one identity switch over every 305 tracked vehicles.

Keywords: traffic monitoring; vehicle detection; multi-vehicle tracking; vehicle re-identification; unmanned aerial vehicles; deep convolutional neural network.

1. Introduction

Unmanned aerial vehicles (UAVs) are popularly applied in a large variety of remote sensing and monitoring applications in both civil and military contexts [1–3]. Intelligent transportation and traffic monitoring are found to be among the applications, where UAVs are receiving increasing interest [4–7]. Data for traffic monitoring and analysis can be collected by various sensors, such as lidar, radar, and video cameras. Among these, video-based traffic analysis is receiving more attention due to the significant advancement of machine learning and computer vision techniques [8–10]. Multi-vehicle detection and tracking is a fundamental task in the video-based analysis of traffic that can be used

for various purposes, e.g., speed control [11], vehicle counting [12], automatic plate recognition [13], and incident analysis [14]. Traditionally, traffic videos are collected by ground surveillance cameras at fixed positions and orientations [15–17]. However, in complex road scenarios, when the traffic moves through multiple roundabout or interchange roads, vehicles' maneuvers cannot be tracked completely using these traditional camera setups [18]. In these cases, UAVs can be complementary platforms to collect traffic videos from various altitudes and viewpoints. Vehicle detection-and-tracking in UAV videos is both rewarding and challenging. For instance, at a higher altitude and with an oblique view direction, a larger field of view can be offered that increases the area under surveillance. However, this advantage comes with the challenge of causing the vehicles to appear in significantly varying resolutions. An instance of this problem is shown in Figure 1, where vehicle #16 is observed in an area of 7125 pixel^2 at the bottom left corner of the video frame (Figure 1b) compared to the same car seen a few seconds later in a bounding box of 1925 pixel^2 at the top left corner of the video frame (Figure 1e). Additionally, in multi-object tracking (MOT) tasks, one of the most significant challenges is to alleviate incorrect identity switches. That is, a vehicle cannot be tracked efficiently unless it is identified as the same object for the whole or most of the video frames, in which it appears. The state-of-the-art in MOT is conventionally focused on tracking pedestrians and sport players [19]. Comparatively, tracking vehicles while correctly re-identifying (Re-Id) them is more challenging since appearance similarities among vehicles are much more than humans as vehicle models and colors are limited [20]. In the field of vehicle tracking, Re-Id approaches focus on re-identifying a vehicle across multiple videos captured by different cameras as opposed to re-identifying a vehicle in the same video where it undergoes severe appearance changes [21–23]. This problem becomes even more pronounced in UAV videos. Since vehicles appear smaller in such videos, their appearance features also become more limited and less distinctive. Moreover, using ground surveillance cameras, the vehicles are observed as they drive straight in parallel directions. They may also turn a few degrees, e.g., 90 degrees to take a left or right turn [18]. Therefore, the appearance features of the same vehicle do not vary much in a ground surveillance video [18,20,24]. In UAV traffic videos, however, when vehicles drive through roundabouts or interchange roads, their appearance features vary considerably depending on their relative orientation to the camera. An instance of this challenge is presented in Figure 1. This sample video is captured from approximately 70 m above the roundabout level with close to 40 degrees tilt. The appearance features of vehicle #16 change aggressively as the vehicle completes its maneuver from the exit ramp through two roundabouts to the highway entrance ramp. In this case, it is exceptionally challenging to robustly track the vehicle and consistently re-identify it through its whole trajectory. Vehicle occlusion by either other vehicles or road assets, such as traffic signs, is also a notable problem in traffic flows captured by UAVs. Finally, since the UAV moves freely, no assumptions can be made about the camera being stationary. That is, the optical flow happening in the video is the result of both the scene flow and the camera motion; this makes the vehicle motion-estimation task less trivial.

In this paper, we propose a vehicle tracking-by-detection approach applicable to UAV videos, which is robust against orientation and scale variations. A state-of-the-art object detector, the most recent version of You Only Look Once (YOLOv3) [25,26], is fine-tuned by custom-labeled datasets for vehicle detection. Multi-vehicle tracking is achieved by integrating Kalman filtering (for motion estimation) and deep appearance feature matching (for vehicle re-identification). The contributions of this research article are three-fold: (1) adapting a deep-learning-based vehicle re-identification (Re-Id) approach for tracking multiple vehicles using UAVs in near-real-time (11 frames per second (fps) in tracking and 30 fps in detection) from high-resolution videos (2720p); (2) achieving an unprecedented performance in terms of maximizing detection precision and recall, minimizing identity switches, and maximizing multi-object tracking accuracy; (3) maintaining high detection and tracking accuracy under complex traffic-monitoring scenarios that rarely are tested in the literature, e.g. in late afternoons with long vehicle shadows, in dawn with vehicles lights being on, over roundabouts and interchange roads where vehicles directions change considerably, and from various dynamic viewpoints where vehicles' appearance undergo substantial perspective distortions including considerable orientation

and scale changes. The test videos applied for performance analysis in this study are made publicly available at the following link: https://github.com/jwangjie/UAV-vehicle-tracking.

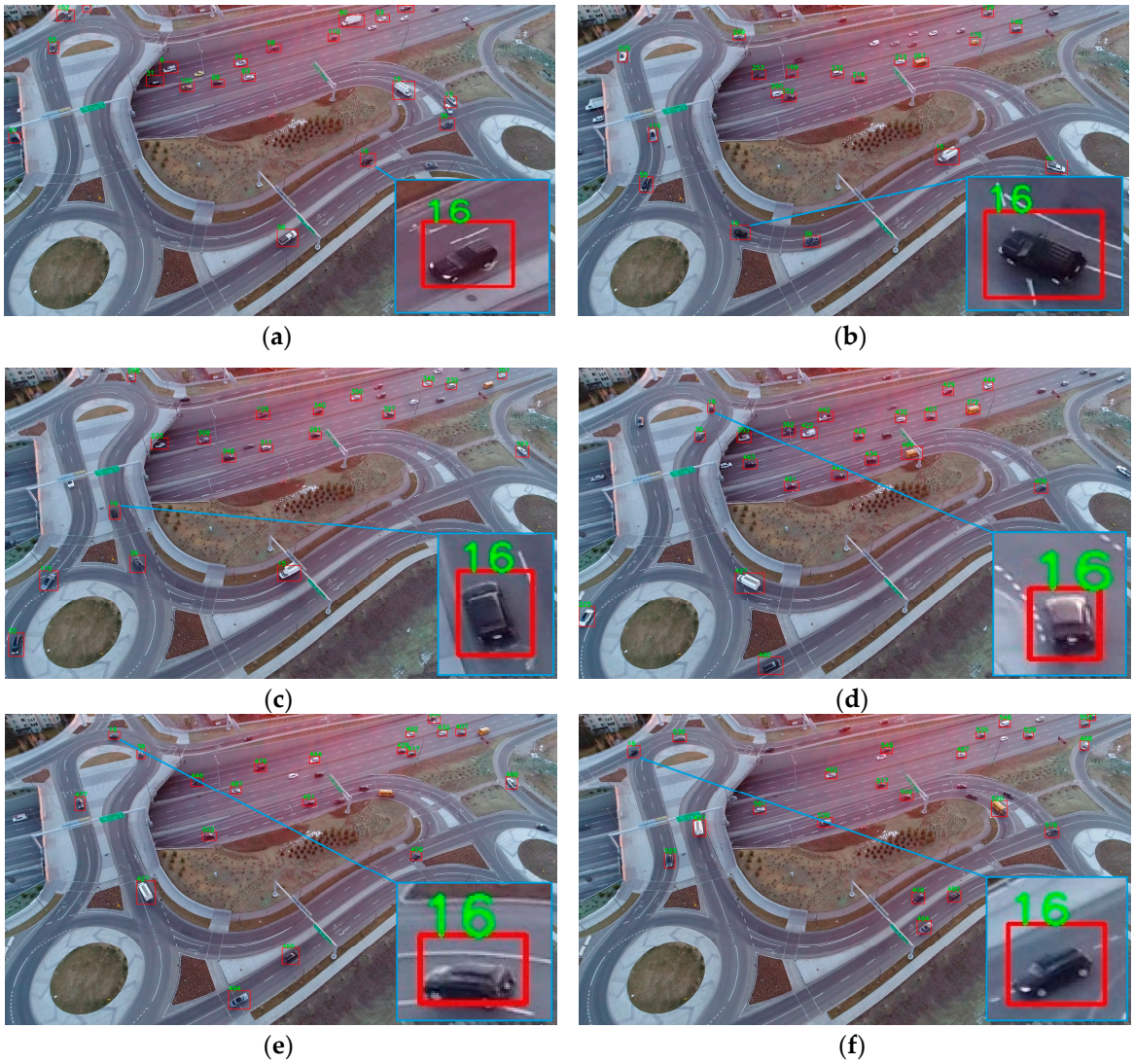


Figure 1. Vehicle detection and tracking at roundabouts from unmanned aerial vehicle (UAV) videos. As vehicle #16 is driving through roundabouts (shown sequentially from (**a**) to (**f**)), its appearance and resolution change considerably.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work on vehicle detection and tracking. Section 3 presents the applied methodology. Sections 4 and 5 describe the experiments and present the obtained results. Section 6 discusses the outcomes of the experiments. Concluding remarks and perspectives around future work are provided in Section 7.

2. Related Work

In general, MOT is a solution to detect and predict the trajectories of multiple objects simultaneously while maintaining their identities through a video sequence [27]. As such, the main two components of MOT are detection and tracking. MOT techniques are generally categorized as online and offline (batch) approaches [28]. Offline MOT methods estimate object trajectories using detection results from the past and future frames, while the online MOT methods utilize the past and current frames only. Online MOT is the preferred technique for traffic monitoring since it allows real-time analysis of the traffic flow. In online methods, the tracker first receives the detection results in the current frame. Then,

data association should be performed to relate the detection results in the current frame to those of the previous frames.

The most recent techniques of multi-object detection are based on CNNs, which are either based on a single-network or two-stage networks. Two-stage detectors, such as fast region-based convolutional network (Fast R-CNN) [29], Mask R-CNN [30] and Faster R-CNN [31], first use region proposal networks (RPN) to predict the location of potential objects. Then, they refine bounding-box prediction and classification in the second network by average-pooling features over the proposal regions. In the literature, there are many car detection studies that use two-stage network detectors. Examples include R-CNN with multi-scale feature upsampling by deconvolution [32], Faster R-CNN with hyper region proposal and a cascade of boosted classifiers [33], and scale-invariant Faster R-CNN with context-aware region-of-interest pooling [34]. Single-network detectors, such as YOLO [35], single-shot multi-box detector (SSD) [36] and RetinaNet [37], perform bounding-box prediction and classification through a single network. SSD has been used for vehicle detection in several studies; e.g. in its original form [38], with a modified architecture combined with the Slim ResNet-34 [39], aided with temporally identified regions of interest [40]. YOLO and RetinaNet architectures are also used in the literature for vehicle detection. They generally report higher accuracy in terms of detecting smaller objects in the large image space. Examples include YOLOv2 [41] with improved loss normalization, anchor-box clustering and multi-layer feature fusion [42], YOLOv3 with additional prediction layers [43], Tiny YOLOv3 with repeated up-sampling and additional passthrough layers [44], and focal loss-based RetinaNet [45]. These approaches are performed on relatively low-resolution videos in order to achieve high detection speed and proper accuracy. Examples include resolution of 960 × 540 pixels at 34 fps with average precision of 43%–79% [38], resolution of 300 × 300 pixels at 20 fps with mean average precision (mAP) of 77% [39], varying resolutions up to 1920 × 1080 pixels with processing time of 0.09 sec per frame (~ 11 fps) with F1-score of 39% [40], varying resolutions up to 608 × 608 pixels with processing time of 0.038 sec per frame (~26 fps) with mAP of 68% [42], resolution of 960 × 540 pixels at 9 fps with mAP of 85% [43], resolution of 512 × 512 pixels at 75 fps with mAP of 80%–89% [44], resolution of 960 × 540 pixels at 21 fps with mAP of 74% [45]. A comprehensive report on the performance of single-network detectors against two-stage networks can be found in [46]. In general, comparative studies show that single-network detectors have gained more popularity due to their acceptable precision-time tradeoff. Although two-stage networks provide comparable detection accuracy to single-network ones, their running speed is much slower for practical applications. Single-network (also known as region-free) detectors can offer real-time detection in high-resolution videos [47]. This is an essential characteristic in order to analyze the traffic flow in real-time. Faster R-CNN and YOLOv3 are compared for vehicle detection in reference [48]; it was concluded that YOLOv3 outperforms Faster R-CNN in terms of both recall and speed even though they result in similar precision. It is also known that, in two-stage detectors, the classifier network following the RPN faces difficulties in correctly distinguishing vehicles from complex backgrounds [33]. There are also other studies in vehicle detection that use conventional machine learning and computer vision techniques. In these techniques, traditional features such as histogram of oriented gradient (HOG), Haar-like wavelets, local binary patterns, color probability maps, and semantic features are used with classifiers, such as linear support vector machine (SVM), multi-kernel SVM, and Adaboost, for vehicle detection [49–53]. Object detectors based on discriminately trained deformable part models have also been reported for vehicle detection [54–56]. A comprehensive review of traditional vehicle detection techniques can be found at [10]. These methods provide comparable accuracy to CNN-based approaches. However, they are based on features, that are hand-crafted to represent the details of real data, and classifiers, whose run-time grows considerably with the volume of available features. Therefore, they are not sufficient to provide a balance among the accuracy of vehicle detection, time efficiency, and completeness without requiring human involvement [57].

Traditional multi-object tracking approaches rely only on motion indications; e.g., multiple hypothesis tracking [58], joint probabilistic data association filter [59], Kalman filter [60], Hungarian

algorithm with Kalman filter [61], and probability hypothesis density filter [62]. The complexity of these filters increases considerably as the number of tracked objects increases. This renders them unacceptable for real-time tracking of vehicles in dense urban traffics. Besides, solely relying on motion clues makes these approaches unsuitable for complex scenarios where maneuvers are not predictable using simple kinematic models. Therefore, in recent MOT approaches motion features are integrated with those of appearance to improve vehicle re-identification. Such appearance features can be extracted using CNNs [63–67]. For instance, in reference [64], a deep association metric is trained based on a large-scale human dataset [68]. This metric was successfully applied to track humans through long periods of partial occlusions. Recently, Li et al. [69] proposed a speed estimation method from traffic videos captured by UAVs. They first follow a tracking-by-detection framework for vehicle tracking and then conduct vehicle speed estimation. The vehicle detector is YOLOv3 trained by a dataset collected from a ground surveillance camera [70] as well as a custom-labeled airborne video. Intensity-based similarity measures between bounding boxes along with their intersection-over-union (IOU) scores are applied for tracking. This approach was tested in simulated scenes at AirSim [71] as well as real UAV traffic videos. While speed-estimation results were metrically evaluated, no quantitative measures were provided for detection and tracking. The tracking method of [69] has originally been applied for human tracking [61], which was shown to suffer from severe identity-switching issues. There are also traditional approaches based on computer vision techniques that are still popular [22,72,73]. In such approaches, discriminative methods using hand-crafted features, such as scale-invariant feature transforms (SIFT), speeded up robust features (SURF), region-based features or edge-based features, are applied for re-identifying vehicles [74–78]. Optical-flow estimation using variational methods [73,79,80], e.g., the Lucas–Kanade method, and correlation-based filters [40], e.g., background-aware correlation filter, are also used for vehicle tracking. Traditional computer vision-based methods, however, generally cannot provide an efficient and reliable feature detector/descriptor for large scale videos of dense urban traffic flows and rarely are tested for multi-vehicle tracking in real-time. These approaches are frequently applied in simple traffic scenarios, e.g., vertical bird's eye view of vehicles driving on straight roads. Leading-edge techniques for vehicle detection, tracking, and multi-camera vehicle re-identification using ground surveillance cameras were presented through the last AI City Challenge [18]. Tang et al. [22] proposed an award-winning tracking-by-detection paradigm to detect and track multiple vehicles from closed-circuit television (CCTV) cameras. In their study, vehicle detection is performed using YOLOv2, where a manually labeled dataset is applied to train the detector via fine-tuning. Then, a complex loss-function, fusing visual and semantic features, is used for data association. Visual features extracted from RGB, HSV, Lab, LBP, and HOG spaces for each vehicle target are learned by an adaptive histogram-based appearance model. Semantic features, including trajectory smoothness, velocity changes, and temporal information, are incorporated into a bottom-up clustering strategy to address the re-identification problem. In their test videos, although vehicle tracking was performed within a short distance from the camera (maximum 50 m), the identity-switching issue could still be noticed. The winners of the second place of the traffic speed estimation challenge [23] also followed a tracking-by-detection paradigm. They use a model called DenseNet as the object detector, generate one bounding box for each detected vehicle, and then apply the minimum-cost-maximum-flow optimization method for the inter-frame association of the bounding boxes. They also apply Kalman filtering to obtain a smoother trajectory. No direct experimental results or videos were provided to show how well their methodology handled the identity-switching issue.

It can be concluded that developing an efficient and reliable multi-vehicle tracking approach for large-scale videos in complex road structures is still a challenging task and a topic of ongoing research. This is especially the case for UAV videos since, at higher altitudes and oblique moving viewpoints, minimal appearance features are available for the targets.

3. Methodology

The most significant challenge for vehicle detection and tracking from UAV videos is that a limited amount of distinctive information, such as plate number, is available for each vehicle. Thus, in this paper, we propose a deep learning-based tracking-by-detection approach. This approach is reliably applicable to long-term tracking (more than 6 min) in high-resolution videos (2720-by-1530 pixels) captured by UAVs from complex road structures with dense traffics. The details of this approach are presented in the following sections.

3.1. Detection

The third version of You Only Look Once, YOLOv3, is currently leading the state-of-the-art in real-time object detection [26] and, accordingly, is used in this study too. The architecture of YOLOv3 is shown in Figure 2.

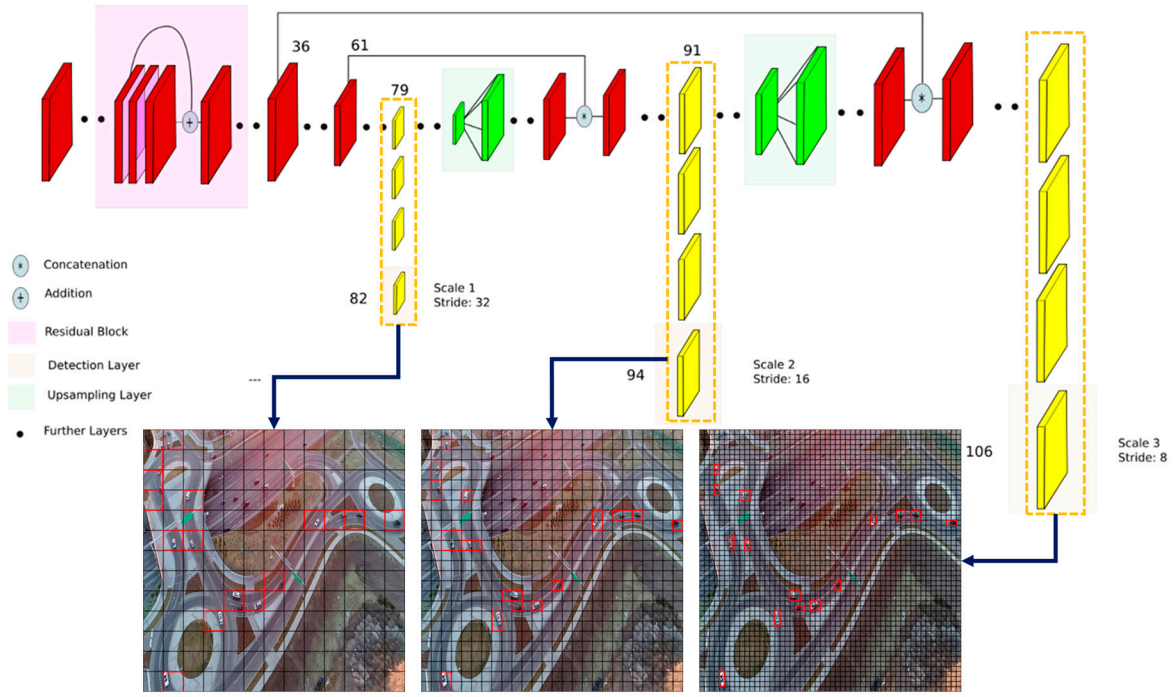
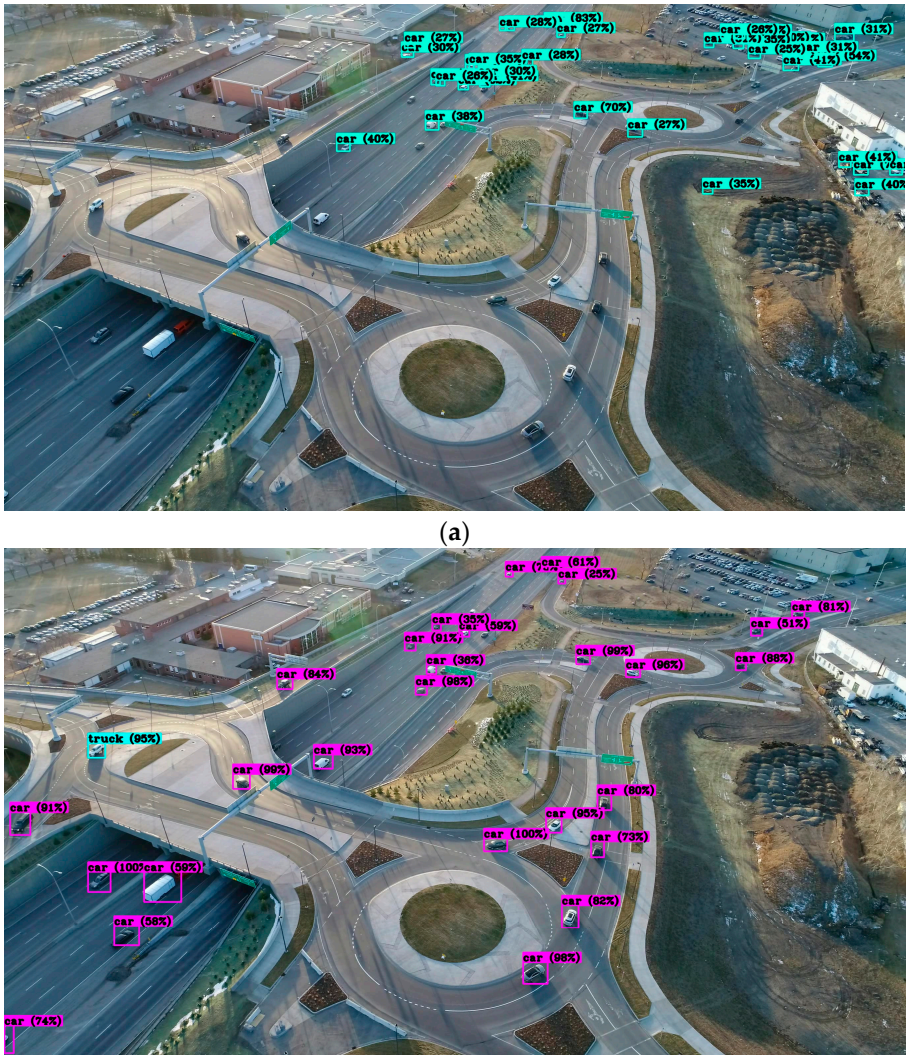


Figure 2. The You Only Look Once (YOLOv3) network architecture, modified from [25].

In YOLOv3, detection is performed by applying 1×1 kernels on feature maps in three different resolutions. As shown in Figure 2, the first detection happens at layer 82, resulting in a detection feature map of size $13 \times 13 \times 255$. The feature map of layer 79 is first connected to several convolutional layers and then is up-sampled to size 26×26. The feature map is then concatenated with the feature map of layer 61. The second detection is made at layer 94, resulting in a detection feature map of size $26 \times 26 \times 255$. The third detection happens at layer 106, yielding a feature map of size $52 \times 52 \times 255$. A similar procedure is repeated, where the feature map of layer 91 is connected to few convolutional layers and then is concatenated with the feature map of layer 36. The 13×13 layer is responsible for detecting large objects, the 52×52 layer can detect small objects, and the 26×26 layer succeeds in detecting medium-size objects [25]. This is the most interesting feature of YOLOv3 regarding multi-vehicle detection in UAV videos; that is, it generates detection probabilities at three different scales. This feature makes the detector scale-invariant to some extent and improves the accuracy of detecting smaller vehicles [81].

There are several existing datasets and pre-trained networks for vehicle detection [26,81,82]. However, most of these datasets are captured by ground vehicles or surveillance cameras. The vehicle detectors trained by these datasets perform poorly on UAV videos, as shown in Figure 3.



(**a**)

(**b**)

Figure 3. Vehicle detection in UAV videos: (**a**) vehicle detection by YOLOv3 trained via ground-surveillance videos; (**b**) vehicle detection by fine-tuning YOLOv3 using custom-labeled airborne videos.

Several videos were captured (by a professional drone service company) for this work by a DJI Phantom 4 Pro with a resolution of 2720*1530 pixels at 30 fps from different positions, angles, and lighting conditions in four different days (please see the source of the videos in the Acknowledgments section). Two of the DJI videos along with two public UAV datasets, aerial-cars-dataset [48] and UAV Detection and Tracking (UAVDT) Benchmark [47] were used for fine-tuning YOLOv3. The vehicle detector, YOLOv3, was fine-tuned from a pre-trained set of weights (i.e., darknet53.conv.74 on ImageNet). The three datasets used for fine-tuning our vehicle detector included: 155 images from aerial-cars-dataset (Figure 4a), 1374 images from the M0606 folder UAVDT-Benchmark (Figure 4b), and our custom-labeled 157 images (Figure 4c).

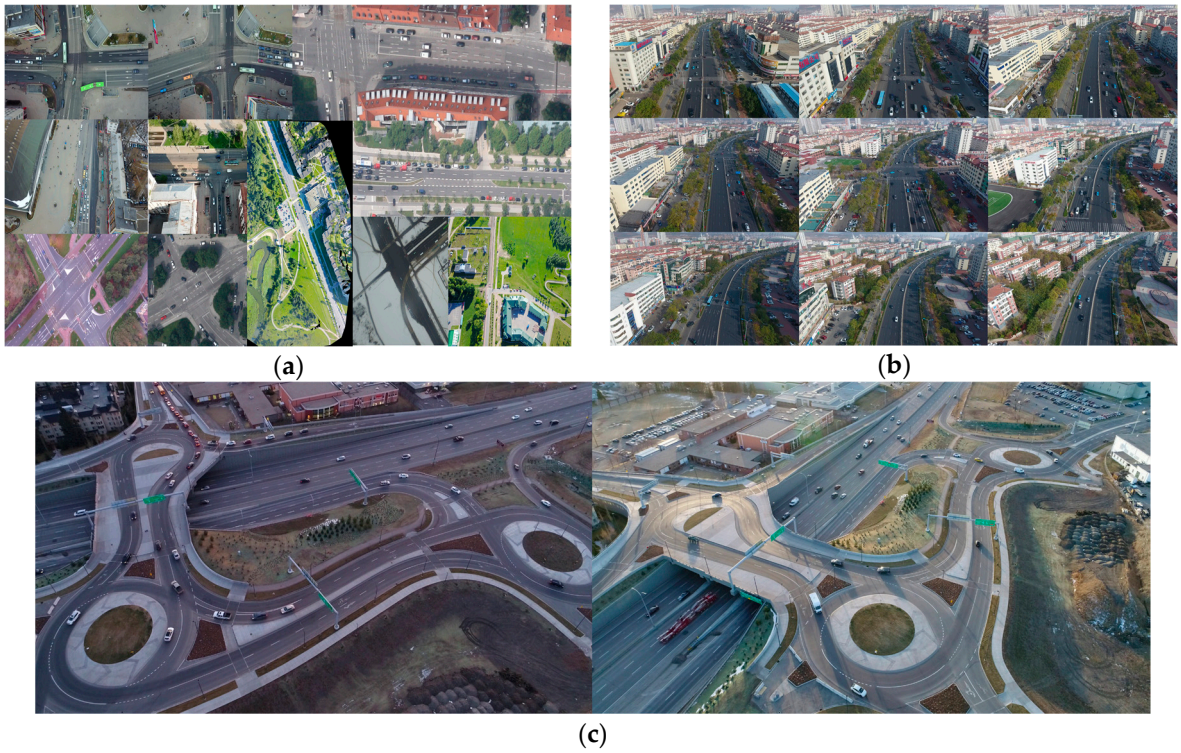


Figure 4. Datasets used for fine-tuning YOLOv3 vehicle detector: (**a**) 155 images from aerial-cars-dataset; (**b**) 1374 images from the UAVDT-Benchmark (M0606); (**c**) 157 images from our UAV videos (DJI_0004 and DJI_0016).

3.2. Tracking

To achieve reliable vehicle tracking, data-association methods are required to relate the detected vehicles in the current frame to those of the previous frames. In this work, motion estimation based on Kalman filtering is integrated with deep appearance features to robustly track multiple vehicles across video frames while maintaining their identities regardless of excessive viewpoint (scale and orientation) changes in UAV videos. The proposed framework for multi-vehicle tracking-by-detection is shown in Figure 5.

Table 1. The architecture of the wide residual network [63] trained for vehicle re-identification.

Layers	Patch Size/Stride	Output Size
Conv 1	$3 \times 3/1$	$32 \times 128 \times 64$
Conv 2	$3 \times 3/1$	$32 \times 128 \times 64$
MaxPool 3	$3 \times 3/2$	$32 \times 64 \times 32$
Residual 4	$3 \times 3/1$	$32 \times 64 \times 32$
Residual 5	$3 \times 3/1$	$32 \times 64 \times 32$
Residual 6	$3 \times 3/2$	$64 \times 32 \times 16$
Residual 7	$3 \times 3/1$	$64 \times 32 \times 16$
Residual 8	$3 \times 3/2$	$128 \times 16 \times 8$
Residual 9	$3 \times 3/1$	$128 \times 16 \times 8$
Dense 10		128
l_2-norm		128

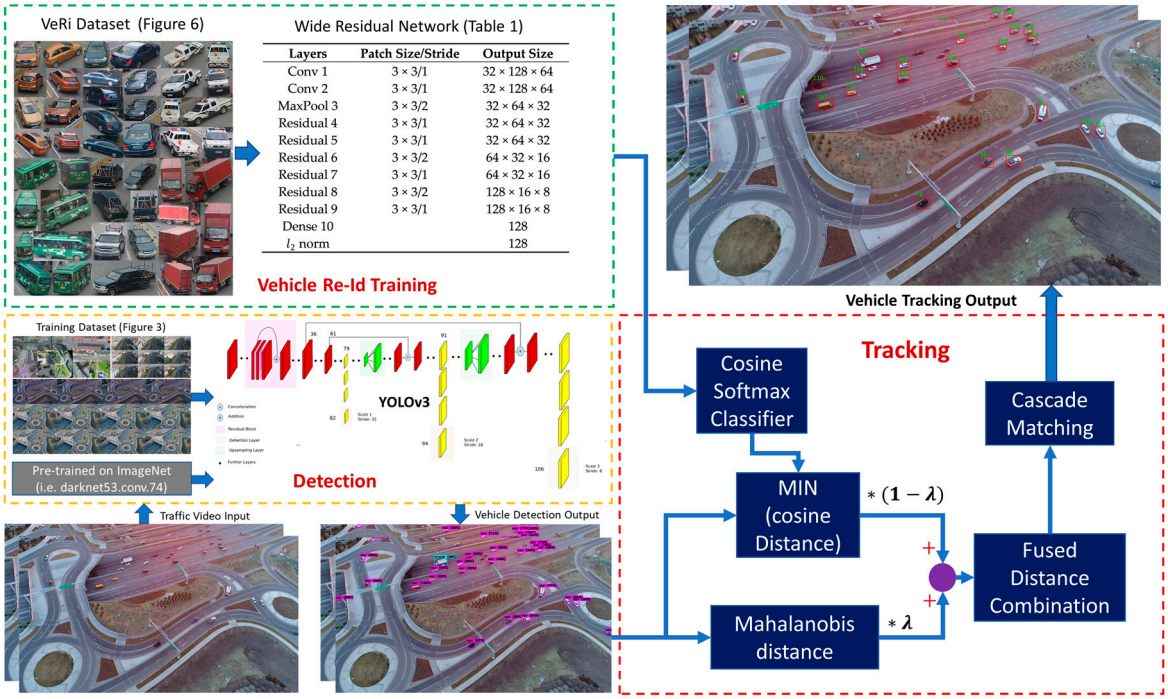


Figure 5. The proposed framework for multiple vehicle tracking-by-detection. Detection: the vehicle detector was fine-tuned on YOLOv3 by a custom labeled dataset together with two other datasets (Figure 4). Re-Id training: deep vehicle Re-Id appearance features were extracted by training a wide residual network (Table 1) with the VeRi dataset (Figure 6). Tracking: the detected bounding boxes were used to calculate the Mahalanobis distance (Equation (1)) as the motion metric (Equation (2)), and the pixels inside the bounding boxes were used to calculate the minimum cosine distance (Equation (3)) as the deep appearance similarity metric (Equation (4)); the two metrics were then integrated in a weighted form to conduct data association using cascade matching.

Figure 6. Sample images of VeRi dataset [20].

3.2.1. Motion and Deep Appearance Features

The inter-frame movement of each vehicle was described with a constant-velocity motion model. For each vehicle, the tracking state vector contains eight elements $(x,\ y,\ r,\ h,\ \dot{x},\ \dot{y},\ \dot{r},\ \dot{h})$, where $(x,\ y)$ denotes the location of the center of the bounding box in the image, r is the aspect ratio, h represents the height of the bounding box, and their corresponding first-order derivatives in consecutive frames are denoted as $(\dot{x},\ \dot{y},\ \dot{r},\ \dot{h})$. The coordinates of the bounding boxes detected in the new frame as well as the appearance features extracted from the pixels inside the bounding boxes form the observations required to update the vehicles state. Appearance features are important aspects of this tracking

technique. To this end, in this paper, a deep CNN architecture (wide residual network) designed initially for human identification by [63] was trained with a Vehicle Re-id (VeRi) dataset available in [20]. This dataset was created from real-world traffic videos captured by 20 surveillance cameras. The cameras were installed along several roads and recorded videos at 25 fps and 1920×1080 resolution. This dataset includes 49325 images from 776 vehicles, in which each vehicle is captured by at least two cameras from different viewpoints, as shown in Figure 6.

The architecture of the wide residual network applied for vehicle Re-Id is shown in Table 1. This wide residual network is very similar to that originally designed for learning human appearance features [63,83]. This network is composed of two convolutional layers and six residual layers, followed by a dense layer to extract feature descriptors of length 128. The final l_2-norm layer projects the features to a unit hypersphere.

3.2.2. Data Association

A data association metric was designed to integrate both motion and deep appearance features. Detected and identified bounding boxes from the last frame are transitioned to (predicted in) the next frame using a constant-velocity motion model. A squared Mahalanobis distances between the predicted bounding boxes and the newly detected ones are calculated as the motion metric as follows,

$$d_1(i,\ j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i) \tag{1}$$

where d_j denotes the location of j-th newly detected bounding box, y_i and S_i represents the mean and covariance matrix of the i-th predicted bounding box. A threshold can be tuned to control the minimum confidence to conduct the data association between objects i and j. The decision can be made by an indicator $b_{i,j}^{(1)}$ defined as follows,

$$b_{i,j}^{(1)} = \begin{cases} 1, & d_1(i,\ j) \le t^{(1)} \\ 0, & d_1(i,\ j) > t^{(1)} \end{cases} \tag{2}$$

The indicator will be equal to 1 if the Mahalanobis distance is smaller or equal to a threshold $t^{(1)}$, *which is* tuned by trial-and-error. In this case, the association between objects i and j will happen, and state update will be formed. Sever identity switches occur when using motion as the only metric for data association. Therefore, a second metric, called hereafter as the deep appearance-similarity metric is also applied. This metric measures the minimum cosine distance of the appearance features between objects i and j as follows,

$$d_2(i,\ j) = \min_k (1 - r_j^T r_k^{(i)}) \tag{3}$$

where r_j is the appearance feature vector of the recently detected object j, and $r_k^{(i)}$ represents the most recent k feature vectors of the tracked object i. In his study, parameter k is empirically set to a maximum number of 100 available vectors. Similar to the motion metric, to determine whether an association is acceptable or not, a binary indicator of the deep appearance-similarity metric was defined as follows,

$$b_{i,j}^{(2)} = \begin{cases} 1, & d_2(i,\ j) \le t^{(2)} \\ 0, & d_2(i,\ j) > t^{(2)} \end{cases} \tag{4}$$

A suitable threshold for this indicator was empirically found using the VeRi dataset shown in Figure 6 by comparing appearance-similarity metrics corresponding to correct and false association-hypotheses.

The Mahalanobis distance (d_1) between the locations of the predicted and the newly detected bounding boxes along with the cosine distance (d_2) between the deep appearance features of the tracked objects and the newly detected objects can be integrated as follows,

$$c_{i,\,j} = \lambda d_1(i,\ j) + (1-\lambda) d_2(i,\ j) \tag{5}$$

By tuning the weight factor, λ, two metrics can contribute complementarily to the re-identification and tracking tasks. For example, when there is substantial camera or object motion, constant-velocity motion models are not valid anymore. Thus, the appearance metric becomes more significant by setting $\lambda = 0$; and on the contrary, when there are limited vehicles on the road without long-term partial occlusions, setting $\lambda = 1$ can accelerate tracking computation by shutting down the calculation of the deep appearance-similarity metric. An improved Hungarian algorithm was then applied as the matching approach [64] for data association using the metric of Equation (5). In our implementation, a maximum loss time of 30 frames (since in our videos, the frame rate is 30 fps) is considered. In order to avoid redundant computations, if a tracked object (object *i*) is not re-identified in the most recent 30 frames passed since its last instantiation, it will be assumed that it has left the scene. If the object is seen again later, a new ID will be assigned to it.

4. Experiments

In the experiments, a DJI Phantom 4 Pro quadrotor with a gimbal-mounted camera was used to capture various traffic videos. The camera was adjusted to a resolution of 2720*1530 pixels and speed of 30 fps. The computer used in performing these experiments was equipped with a six-core Intel Core i7-8700 CPU with 12 threading, operating at 3.20 GHz, an NVIDIA GeForce GTX 1080 Ti and Intel UHD Graphics 630 GPU, and 32.0 GB of RAM. The main software packages installed on this desktop included Ubuntu 16.04, Python 3.5, TensorFlow 1.4.0, OpenCV 3.4.0, and CUDA 9.0.

The proposed vehicle tracking-by-detection approach was evaluated based on various test videos, as shown in Figure 7, including three videos collected by the DJI Phantom 4 Pro, one video of UAVDT-Benchmark dataset (M0101), and two videos of FHY-XD-UAV-DATA (scene 2 and scene 5) [69]. Three DJI videos including video 1 (DJI_0006) collected in a typical lighting condition (Figure 7a), video 2 (DJI_0013) collected in a late afternoon with long vehicle shadows (Figure 7b), and video 3 (DJI_0001) recorded in dawn with vehicle lights (Figure 7c) were kept aside for testing the proposed approach. Although the test videos were captured from the same scene as that of the training videos (Figure 4c), these three videos had never been seen by the vehicle tracker. UAVDT-Benchmark dataset (M0101) shown in Figure 7d is also different from the training video (M0606) in Figure 4d. Scene 2 (rural street) and Scene 5 (intersection) of dataset FHY-XD-UAV-DATA, shown in Figure 7e,f, were also completely new videos and never used for training.

For DJI videos, the flight altitude above the roundabout layer is from 60 and 150 m, with tilt angles of 30–60 degrees. We have calculated this information by estimating the exterior orientation parameters of the camera at some sample frames via photogrammetric bundle adjustment with a scale constraint added by considering the length of middle road dash-lines (3.15 m). This is only an approximation since the UAV constantly moves either by purpose or due to wind forces. Speed limits in these datasets vary between 100 km/h (in the underlying highway) and 0 km/h (at yield points at roundabouts and ramps). For the UAVDT-Benchmark dataset (M0101), the altitude is described as higher than 70 m [47]. For the FHY-XD-UAV-DATA dataset (Scene 2 and Scene 5), the flight altitude is generally between 50 to 80 m [69]. No information about the title angles of these videos is available.

Figure 7. Testing videos: (**a**) DJI video 1 (DJI_0006) collected with a standard lighting condition; (**b**) DJI video 2 (DJI_0013) collected in late afternoons with long vehicle shadows; (**c**) DJI video 3 (DJI_0001) recorded in dawn with vehicle lights being on; (**d**) UAVDT-Benchmark dataset intersection (M0101); (**e**) Scene 2 of FHY-XD-UAV-DATA; (**f**) Scene 5 of FHY-XD-UAV-DATA.

To evaluate the proposed tracking method in the DJI videos, only ramps and roundabouts, bounded by the red polygon in Figure 8, comprised our region of interest (ROI). Even though our method performs equally well on the highways, the proposed tracking technique was only evaluated for the vehicles in the ROI. There were two reasons for selecting this ROI. First, the primary motivation of this work is to address multi-vehicle tracking in complex road structures that are not feasible by ground surveillance videos. Second, manually counting the trajectory of the dense traffic underneath the highway for creating the tracking ground-truth was too time-consuming. Please note that this ROI only applies for tracking evaluation; that is, the detection accuracy was tested on the whole area of the frame. UAVDT-Benchmark and FHY-XD-UAV-DATA videos were evaluated at the whole video frames both for detection and tracking.

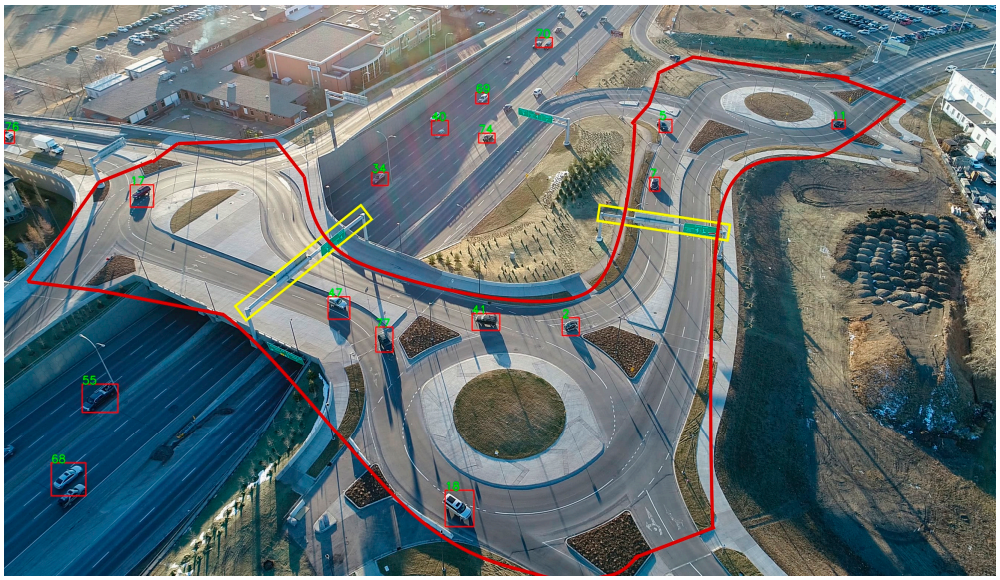


Figure 8. The red polygon denotes the region of interest for evaluating the tracking technique; it comprises the roundabouts and ramps, which are examples of complex road structures.

The performance of the proposed multi-vehicle tracking-by-detection approach was quantitively analyzed via various metrics [27,84,85]. These metrics include: true positive (TP), false positive (FP), true negative (TN), false negative (FN), identification precision (IDP), identification recall (IDR), F1 score, multiple-object tracking accuracy (MOTA), mostly tracked (MT), mostly lost (ML), and identity switching (IDSW). Among these metrics, TP, FP, TN, FN are applicable to both detection and tracking, IDP, IDR, and F1 score are used for detection evaluation, and MOTA, MT, ML, and identity switching (IDSW) are applied for tracking evaluation. The definitions of the above metrics are provided below.

- TP: True Positive, Number of positive observations that are correctly predicted as positive.
- FP: False Positive, Number of negative observations that are incorrectly predicted as positive.
- TN: True Negative, Number of negative observations that are correctly predicted as negative.
- FN: False Negative, Number of positive observations that are incorrectly predicted as negative.
- IDP: Identification Precision, True positive divided by the total number of observations that are predicted as positive.

$$IDP = \frac{TP}{TP + FP} \times 100 \tag{6}$$

- IDR: Identification Recall, True positive divided by the total number of positive observations.

$$IDR = \frac{TP}{TP + FN} \times 100 \tag{7}$$

- F1 score: Harmonic mean used to fuse IDP and IDR.

$$F1 = \frac{2 * IDP * IDR}{IDP + IDR} \times 100 \tag{8}$$

- MT and ML: These are to evaluate what portion of the trajectory of a vehicle is recovered by the tracking method. An object is mostly tracked (MT) if it is successfully tracked for at least 80% of its life span (the time during which it is observable in the video). If a track is recovered for less than 20% of its total length, it is said to be mostly lost (ML). It is irrelevant for MT and ML whether the identity of the object remains the same.
- IDSW: Number of times the identity of the tracked object changes.

- MOTA: It is the most widely used metric since it summarizes the overall tracking accuracy in terms of FN, FP and IDSW as follows,

$$MOTA = (1 - \frac{\sum_i (FN_i + FP_i + IDSW_i)}{\sum_i GT_i}) \times 100 \quad (9)$$

where i is the frame index, and GT is the number of ground truth vehicle objects.

Due to the long duration of the testing videos (usually more than 10 min), high frame rate (30 fps), and a large number of vehicles observable in the videos, benchmarking the ground-truth was considerably labor-intensive. Therefore, the ground-truth data for detection evaluation was collected at a frame rate of 1 fps. We believe that the total number of observable cars do not change considerably within 1 s; thus, selecting this frame rate for ground-truth creation does not impact the fairness of our evaluations. For tracking, however, the ground-truth data need to be generated at the original video frame rate (i.e., 30 fps). Therefore, the middle 30 s of each video was selected for generating the ground-truth data for tracking evaluation. For M0101 video of UAVDT-Benchmark and Scene 5 video of FHY-XD-UAV-DATA, since the whole duration of the videos was less than 1 min, the complete duration of the videos was used for tracking evaluation.

5. Results

Tables 2 and 3 summarize the performance of the proposed tracking-by-detection approach in terms of the evaluation metrics introduced in Section 4. The upward arrow (↑) and downward arrow (↓) beside each evaluation metric should be interpreted as the higher the score, the better performance, and the lower the score, the better performance, respectively. In Tables 2 and 3, the total number of ground-truth vehicles is listed in the second column as Total #.

Table 2. The evaluation metrics of multi-vehicle detection from UAV videos.

Video	Total #	TP↑	FP↓	TN↑	FN↓	IDP↑	IDR↑	F1 Score↑
DJI video 1	4969	4804	0	3	165	100.00	96.68	98.3
DJI video 2	864	647	56	0	217	92.03	74.88	82.6
DJI video 3	6984	6134	22	0	850	99.64	87.83	93.4
M0101	191	170	3	0	21	98.27	89.01	93.4
Scene 2	153	138	16	0	15	89.61	100.00	94.5
Scene 5	5901	4881	23	0	985	99.53	83.21	90.6

Table 3. The evaluation metrics of multi-vehicle tracking from UAV videos.

Video	Total #	TP↑	FP↓	TN↑	FN↓	MT↑	ML↓	IDSW↓	MOTA↑
DJI video 1	14496	11949	0	0	2456	52.63%	0%	24	82.89
DJI video 2	5413	3529	299	0	1298	46.15%	0%	90	68.83
DJI video 3	15120	7319	38	0	2543	44%	4%	45	82.63
M0101	558	4441	8	0	610	76.19%	0%	8	88.74
Scene 2	10027	9385	783	0	597	90.90%	0%	13	86.11
Scene 5	16061	11305	161	0	3227	68.57%	5.7%	22	78.77

To provide a reference for assessing the performance of our solution, the results from a state-of-the-art detection-and-tracking approach [47] are provided in Table 4. In [47], four deep CNN detectors, such as Faster R-CNN and SSD [36], combined with eight MOT methods, such as MDP [86], were tested for multi-vehicle tracking in UAV videos. The best performance metrics among all detector-and-tracker combinations are presented in Table 4.

Table 4. The state-of-the-art performance in detection-and-tracking according to [47].

IDP↑	IDR↑	F1 Score↑	MT↑	ML↓	MOTA↑
74.5	55.0	61.5	47.3	19.5	43.0

The rule that we followed for counting vehicle IDSW is the same as all other MOT methods in the literature; that is, an IDSW occurs whenever the assigned identifier to the vehicle changes from one frame to another one.

6. Discussion

From Table 2, it can be seen that high accuracy for vehicle detection was achieved for all test videos except for DJI video 2 with the lowest F1-score of 82.6%. The main reason for this outcome is that the vehicles in this video appear with long shadows, as shown in Figure 7b. None of our training datasets included such illumination conditions, and thus shadows were sometimes wrongly detected as vehicle objects. While compared to the state-of-the-art accuracy in vehicle detection in Table 4, our vehicle detector provides better detection performance (i.e., higher F1 score). The vehicle tracking results are summarized in Table 3. In general, from the MOTA metric, it can be concluded that our proposed method achieved a robust and accurate multi-vehicle tracking performance in complex traffic scenarios. The DJI video 2 (Figure 7b) has still the lowest MOTA due to too many false positives that correspond to shadows wrongly detected as vehicles. For Scene 5, the main reason for having a lower MOTA is that there were some vehicles which were not detected accurately (high FN). We believe that this is mainly related to the fact that the cars in this dataset have considerable shape/model differences with the cars available in our training datasets. For example, buses and trucks in Scene 5 (collected in China) are very different from the training ones (collected in Canada and the United States). The traffic-control booth in the middle of Scene 5, as shown in Figure 7f, was also continuously detected as a stationary vehicle. This is another reason to achieve a relatively lower MOTA in Scene 5.

Vehicles were mostly tracked (MT) with an average of 48% on DJI videos and 79% on other test videos. In the DJI test videos, it was noticed that (1) vehicles were lost when they approached the corners and the edges of the frames probably since the lens distortions had more notable impacts on the appearance of the vehicles; (2) white vehicles were tracked better compared to black vehicles, which could be due to the fact that our training datasets contained fewer black vehicles than white ones. Tracking black vehicles was even more challenging under limited light conditions, such as DJI video 3 (Figure 7c). More custom UAV traffic images with different lighting and weather scenarios as well as a wider variety of vehicle models and colors would be helpful in order to have a more robust vehicle tracker.

From Table 3, it can be concluded that the proposed approach addressed the identity switching problem very well, resulting in a total of 202 identity switches over 61675 tracked vehicles. In total, with the proposed tracking and re-identification approach, only one identity switch occurs when tracking 305 vehicles. The worst performance in terms of identity switch still occurs in DJI video 2. We noticed that, in this test video, despite the poor performance of the re-identification module, the identity of the vehicles could be recovered every now and then. For instance, the identity switches of a vehicle are shown in Figure 9. The identification number of this vehicle changes as follows: 786 → 884 → 902 → 909 → 912 → 914 → 902 → 884 → 918 → 974 → 977 → 918 → 1004 → 1085 → 1004 → 1088 → 1004. We believe that the identity could switch back since the appearance-similarity

metric (Equation (3)) considers the past one-hundred descriptors of every track. Therefore, if the appearance of a car becomes similar to one of its past poses even if it is located many frames behind the current time, the car can still be correctly re-identified. When counting the IDSW metric, if these identity recoveries were considered, then there would only be three IDSWs instead of 17 IDSWs. That is, between Figures 9b and 9h, between Figures 9i and 9l, and between Figures 9m and 9p, there are no identity switches. There is one identity switch from Figure 9a to Figure 9b, one from Figure 9h to Figure 9i, and another one from Figure 9l to Figure 9m. If such a rule were used in counting the identity switches in Table 3, then only a total of 138 identity switches would happen over 61675 tracked vehicles.

It was also noticed that our re-identification approach is robust to occlusions, occurring for less than 30 frames. For instance, Figure 10 shows an example of such a situation where vehicle # 3238 drives towards the traffic sign at frame 7587. At frame 7592, it is partially occluded, yet it is identified. Then, it is blocked by the traffic sign from frame 7593 to frame 7612. As it partially re-appears at frame 7613, it is re-identified correctly.

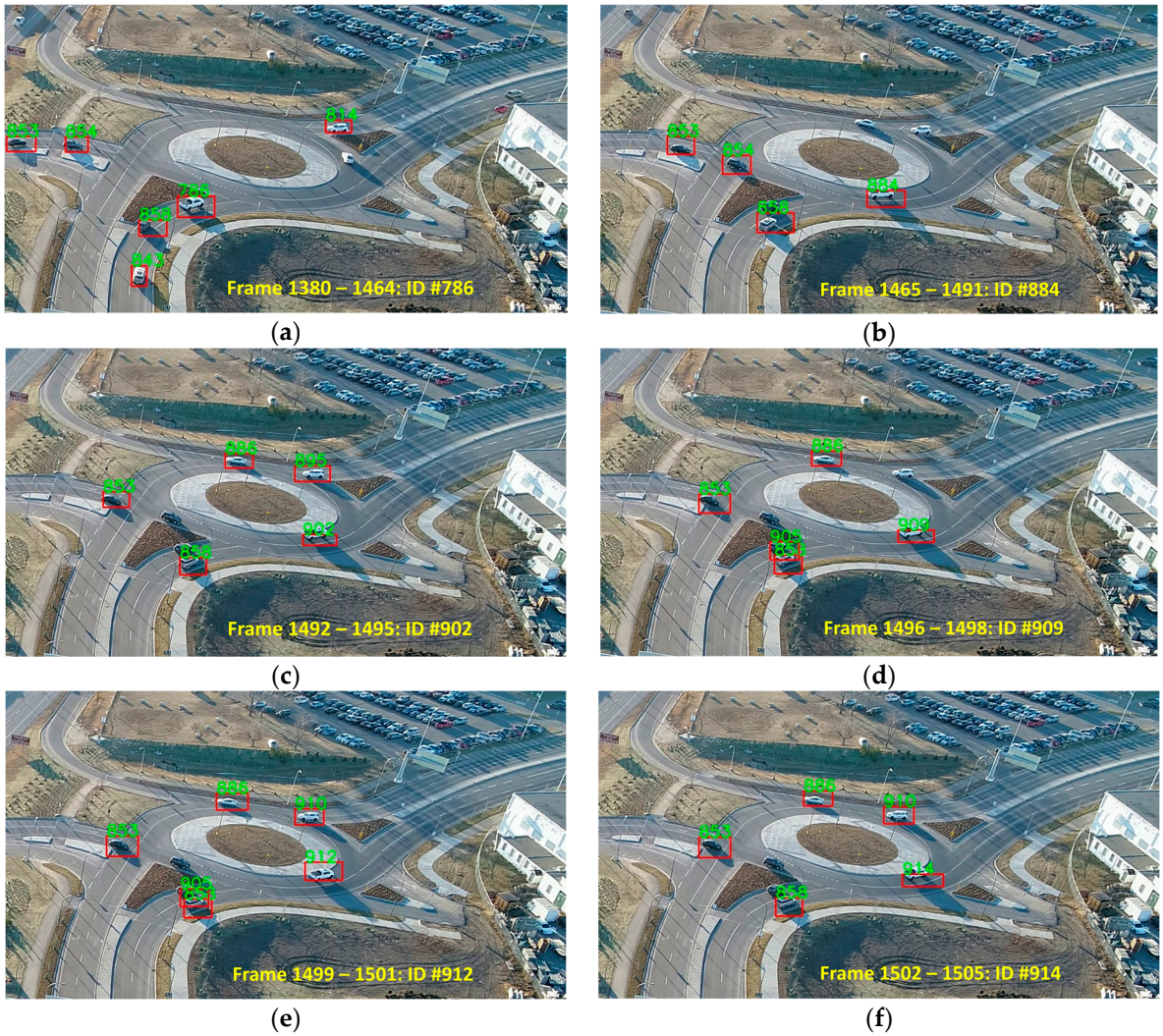


Figure 9. *Cont.*

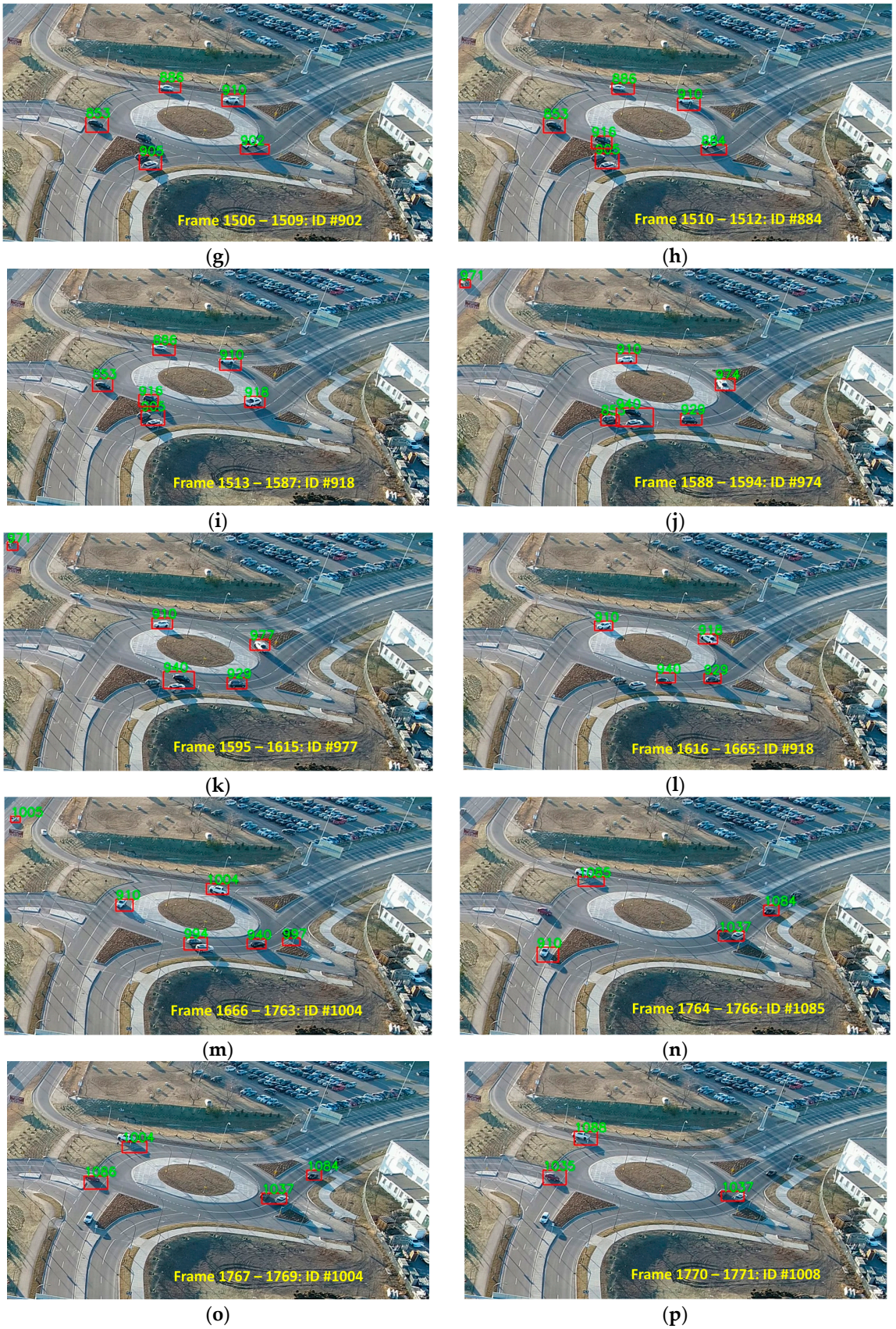


Figure 9. The identity of the vehicles could be recovered every now and then. The identity of the same vehicle changes as: (**a**) 786; (**b**) 884; (**c**) 902; (**d**) 909; (**e**) 912; (**f**) 914; (**g**) 902; (**h**) 884; (**i**) 918; (**j**) 974; (**k**) 977; (**l**) 918; (**m**) 1004; (**n**) 1085; (**o**) 1004; (**p**) 1008. This example is selected from DJI video 2.

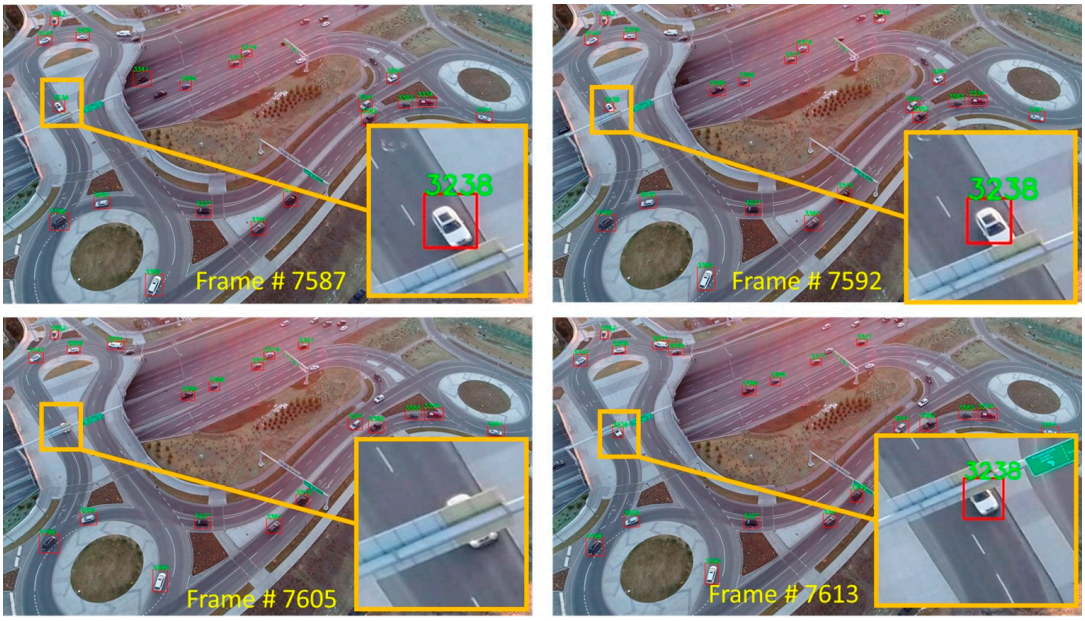


Figure 10. The identity of vehicles is correctly recovered even though they are occluded by other objects for several frames. These examples are selected from DJI video 1.

7. Conclusions and Future Work

In this paper, we proposed a reliable approach for detecting and tracking vehicles from UAV videos. The first objective of this work was to detect vehicles in UAV videos accurately. YOLOv3 was fine-tuned for this purpose. As a result, a high vehicle detection accuracy with an average F1 score of 92.1% was achieved at a speed of 30 fps on 2720p videos. The second objective was to achieve reliable multi-vehicle tracking in complex road structures such as roundabouts, intersections, and ramps from videos taken at various dynamic viewpoints in varying illumination conditions. A tracking-by-detection paradigm was applied where motion and appearance features were integrated to both track the vehicles and re-identify them. The tracking method performed accurately with an average MOTA of 81.3%. The proposed approach also addressed the identity switching issues, resulting in a total of one identity switch over every 305 tracked vehicles.

The weakness of the proposed detection approach was its sensitivity to long vehicle shadows. The main challenge with the proposed tracking approach was its sensitivity to vehicle colors, e.g., black color that strongly correlates with both asphalt color and shadow color. More custom UAV traffic images with different lighting and weather conditions as well as vehicle models and colors will be beneficial to train a more robust vehicle tracker. Traditional data augmentation approaches, as well as generative adversarial networks, can be helpful in improving training data. In addition, in the future, geometric calibration of the camera will be performed, and images will be rectified in order to reduce mis-detection errors near image corners. Of course, this will increase the run time of the approach since the videos are high-resolution and rectifying the frames involves color interpolation. However, in the cases of fisheye and wide-angle lenses with large radial lens distortion coefficients, this step could be important for 3D trajectory extraction and accurate speed estimation.

The other aspect of this line of work, which needs further debate, is regulatory and safety issues involved in the commercial operation of drones. It is mandatory to consider all the impacts of the environment on UAV operation. Each system is capable of functioning well in limited environmental conditions such as air pressure, temperature, humidity, solar radiance and lightning, wind speed, air turbulence, pollution and other atmospherics conditions. The main common reasons that limit public accessibility to unmanned flight certifications are: i) danger of crashing into people and properties, ii) colliding with other aircraft and, iii) disturbing and disordering other civil and military services. In theory, the certification of civilian drone operations is regulated by such authorities as the European

Union Aviation Safety Agency (EASA) in Europe, the Federal Aviation Administration (FAA) in the USA, the Civil Aviation Safety Authority (CASA) in Australia and the Canadian Aviation Regulations (CAR) in Canada. They impose strict rules to prevent unreliable and unnecessary unmanned flight operations. Specifically, in Canada, Transport Canada (TC) is responsible for civil UAV operations. TC enforces a weight limit of 250 g to 25 kg for civil drones. This will not be an issue for traffic-monitoring applications since there are many camera-equipped drones commercially available in this weight range. In addition, TC requires that drones are operated within the pilots' visual-line-of-sight. This limits the application of UAVs for long-term traffic monitoring since a pilot must observe the UAV continuously. Moreover, deploying drones over bystanders and, thus, attended cars, is categorized as an advanced operation. This category obliges the drone pilot to pass an advanced knowledge test and flight review exam to receive the appropriate pilot certificate.

Author Contributions: Conceptualization, J.W. and M.S.; methodology, J.W. and M.S. and S.S.; software, J.W.; validation, S.S.; formal analysis, J.W. and S.S; investigation, J.W. and M.S.; resources, M.S.; data curation, S.S.; writing—original draft preparation, J.W.; writing—review and editing, M.S.; visualization, J.W. and S.S.; supervision, M.S.; project administration, M.S.; funding acquisition, M.S.

Funding: This study was supported through NSERC Discovery.

Acknowledgments: The authors would like to acknowledge that the videos captured by DJI Phantom Pro 4 used in the Experiments section are provided by Professor Dr. Lina Kattan, from the University of Calgary.

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

References

1. Shahbazi, M.; Théau, J.; Ménard, P. Recent applications of unmanned aerial imagery in natural resource management. *GIScience Remote Sens.* **2014**, *51*, 339–365. [CrossRef]
2. Pajares, G. Overview and current status of remote sensing applications based on unmanned aerial vehicles (UAVs). *Photogramm. Eng. Remote Sens.* **2015**, *81*, 281–330. [CrossRef]
3. Shakhatreh, H.; Sawalmeh, A.H.; Al-Fuqaha, A.; Dou, Z.; Almaita, E.; Khalil, I.; Othman, N.S.; Khreishah, A.; Guizani, M. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research Challenges. *IEEE Access* **2019**, *7*, 48572–48634. [CrossRef]
4. Menouar, H.; Guvenc, I.; Akkaya, K.; Uluagac, A.S.; Kadri, A.; Tuncer, A. UAV-enabled intelligent transportation systems for the smart city: Applications and challenges. *IEEE Commun. Mag.* **2017**, *55*, 22–28. [CrossRef]
5. Kanistras, K.; Martins, G.; Rutherford, M.J.; Valavanis, K.P. Survey of unmanned aerial vehicles (uavs) for traffic monitoring. In *Handbook of Unmanned Aerial Vehicles*; Springer: Dordrecht, The Netherlands, 2015; ISBN 9789048197071.
6. Barmpounakis, E.N.; Vlahogianni, E.I.; Golias, J.C. Unmanned aerial aircraft systems for transportation engineering: Current practice and future challenges. *Int. J. Transp. Sci. Technol.* **2016**, *5*, 111–122. [CrossRef]
7. Khan, M.A.; Ectors, W.; Bellemans, T.; Janssens, D.; Wets, G. UAV-Based Traffic Analysis: A Universal Guiding Framework Based on Literature Survey. In *Transportation Research Procedia*; Elsevier: Amsterdam, The Netherlands, 2017.
8. Buch, N.; Velastin, S.A.; Orwell, J. A review of computer vision techniques for the analysis of urban traffic. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 920–939. [CrossRef]
9. Al-Smadi, M.; Abdulrahim, K.; Salam, R.A. Traffic surveillance: A review of vision based vehicle detection, recognition and tracking. *Int. J. Appl. Eng. Res.* **2016**, *11*, 713–726.
10. Sivaraman, S.; Trivedi, M.M. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 1773–1795. [CrossRef]
11. Ke, R.; Li, Z.; Kim, S.; Ash, J.; Cui, Z.; Wang, Y. Real-time bidirectional traffic flow parameter estimation from aerial videos. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 890–901. [CrossRef]
12. Dai, Z.; Song, H.; Wang, X.; Fang, Y.; Yun, X.; Zhang, Z.; Li, H. Video-based vehicle counting framework. *IEEE Access* **2019**, *7*, 64460–64470. [CrossRef]
13. Indira, K.; Mohan, K.V.; Nikhilashwary, T. Automatic license plate recognition. In *Advances in Intelligent Systems and Computing*; Springer: Berlin/Heidelberg, Germany, 2019.

14. Ren, J.; Chen, Y.; Xin, L.; Shi, J.; Li, B.; Liu, Y. Detecting and positioning of traffic incidents via video-based analysis of traffic states in a road segment. *IET Intell. Transp. Syst.* **2016**, *10*, 428–437. [CrossRef]
15. Zhang, S.; Wu, G.; Costeira, J.P.; Moura, J.M.F. FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras. In Proceedings of the IEEE International Conference on Computer Vision, Venezia, Italy, 22–29 October 2017.
16. Peppa, M.V.; Bell, D.; Komar, T.; Xiao, W. Urban traffic flow analysis based on deep learning car detection from CCTV image series. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences—ISPRS Archives, Delft, The Netherlands, 1–5 October 2018.
17. Sochor, J.; Spanhel, J.; Herout, A. BoxCars: Improving fine-grained recognition of vehicles using 3-D bounding boxes in traffic surveillance. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 97–108. [CrossRef]
18. Naphade, M.; Chang, M.C.; Sharma, A.; Anastasiu, D.C.; Jagarlamudi, V.; Chakraborty, P.; Huang, T.; Wang, S.; Liu, M.Y.; Chellappa, R.; et al. The 2018 NVIDIA AI city challenge. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018.
19. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Zhao, X.; Kim, T.-K. Multiple object tracking: A literature review. *arXiv* **2014**, arXiv:1409.7618.
20. Liu, X.; Liu, W.; Ma, H.; Fu, H. Large-scale vehicle re-identification in urban surveillance videos. In Proceedings of the IEEE International Conference on Multimedia and Expo, Seattle, WA, USA, 11–15 July 2016.
21. Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; Huang, T. Deep relative distance learning: Tell the difference between similar vehicles. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2167–2175.
22. Tang, Z.; Wang, G.; Xiao, H.; Zheng, A.; Hwang, J.N. Single-camera and inter-camera vehicle tracking and 3d speed estimation based on fusion of visual and semantic features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018.
23. Feng, W.; Ji, D.; Wang, Y.; Chang, S.; Ren, H.; Gan, W. Challenges on large scale surveillance video analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 69–76.
24. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.Y. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 865–873. [CrossRef]
25. Kathuria, A. What's new in YOLO v3. *Towar. Data Sci.* **2018**. Available online: https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b (accessed on 1 July 2019).
26. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
27. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
28. Yoon, J.H.; Lee, C.-R.; Yang, M.-H.; Yoon, K.-J. Structural constraint data association for online multi-object tracking. *Int. J. Comput. Vis.* **2019**, *127*, 1–21. [CrossRef]
29. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015.
30. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venezia, Italy, 22–29 October 2017.
31. 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] [PubMed]
32. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2016.
33. Tang, T.; Zhou, S.; Deng, Z.; Zou, H.; Lei, L. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. *Sensors* **2017**, *17*, 336. [CrossRef]
34. Hu, X.; Xu, X.; Xiao, Y.; Chen, H.; He, S.; Qin, J.; Heng, P.A. SINet: A scale-insensitive convolutional neural network for fast vehicle detection. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 1010–1019. [CrossRef]
35. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

36. 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*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 21–37.
37. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
38. Anisimov, D.; Khanova, T. Towards lightweight convolutional neural networks for object detection. In Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, 29 August–1 September 2017.
39. Chen, L.; Zhang, Z.; Peng, L. Fast single shot multibox detector and its application on vehicle counting system. *IET Intell. Transp. Syst.* **2018**, *12*, 1406–1413. [CrossRef]
40. Zhao, D.; Fu, H.; Xiao, L.; Wu, T.; Dai, B. Multi-object tracking with correlation filter for autonomous vehicle. *Sensors* **2018**, *18*, 2004. [CrossRef]
41. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
42. 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]
43. Kim, K.-J.; Kim, P.-K.; Chung, Y.-S.; Choi, D.-H. Multi-scale detector for accurate vehicle detection in traffic surveillance data. *IEEE Access* **2019**, *7*, 78311–78319. [CrossRef]
44. Ju, M.; Luo, J.; Zhang, P.; He, M.; Luo, H. A simple and efficient network for small target detection. *IEEE Access* **2019**, *7*, 85771–85781. [CrossRef]
45. Wang, X.; Cheng, P.; Liu, X.; Uzochukwu, B. Focal loss dense detector for vehicle surveillance. In Proceedings of the 2018 International Conference on Intelligent Systems and Computer Vision, ISCV 2018, Fez, Morocco, 2–4 April 2018.
46. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
47. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
48. Benjdira, B.; Khursheed, T.; Koubaa, A.; Ammar, A.; Ouni, K. Car Detection using unmanned aerial vehicles: Comparison between faster R-CNN and YOLOv3. In Proceedings of the 2019 1st International Conference on Unmanned Vehicle Systems-Oman, UVS, Muscat, Oman, 5–7 February 2019.
49. Cao, X.; Wu, C.; Yan, P.; Li, X. Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos. In Proceedings of the International Conference on Image Processing, ICIP, Brussels, Belgium, 11–14 September 2011.
50. Xu, Y.; Yu, G.; Wang, Y.; Wu, X.; Ma, Y. A hybrid vehicle detection method based on viola-jones and HOG + SVM from UAV images. *Sensors* **2016**, *16*, 1325. [CrossRef] [PubMed]
51. Liang, P.; Teodoro, G.; Ling, H.; Blasch, E.; Chen, G.; Bai, L. Multiple kernel learning for vehicle detection in wide area motion imagery. In Proceedings of the 15th International Conference on Information Fusion, FUSION, Singapore, 9–12 July 2012.
52. Grabner, H.; Nguyen, T.T.; Gruber, B.; Bischof, H. On-line boosting-based car detection from aerial images. *ISPRS J. Photogramm. Remote Sens.* **2008**, *63*, 382–396. [CrossRef]
53. Sun, X.; Wang, H.; Fu, K. Automatic detection of geospatial objects using taxonomic semantics. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 23–27. [CrossRef]
54. Niknejad, H.T.; Mita, S.; McAllester, D.; Naito, T. Vision-based vehicle detection for nighttime with discriminately trained mixture of weighted deformable part models. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, ITSC, Washington, DC, USA, 5–7 October 2011.
55. Leon, L.C.; Hirata, R. Vehicle detection using mixture of deformable parts models: Static and dynamic camera. In Proceedings of the Brazilian Symposium of Computer Graphic and Image Processing, Ouro Preto, Brazil, 22–25 August 2012.
56. Pan, C.; Sun, M.; Yan, Z. The study on vehicle detection based on DPM in traffic scenes. In Proceedings of the International Conference on Frontier Computing, Tokyo, Japan, 13–15 July 2016; pp. 19–27.

57. Chen, X.; Xiang, S.; Liu, C.L.; Pan, C.H. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1797–1801. [CrossRef]
58. Kim, C.; Li, F.; Ciptadi, A.; Rehg, J.M. Multiple hypothesis tracking revisited. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015.
59. Rezatofighi, S.H.; Milan, A.; Zhang, Z.; Shi, Q.; Dick, A.; Reid, I. Joint probabilistic data association revisited. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015.
60. Lee, M.-H.; Yeom, S. Tracking of moving vehicles with a UAV. In Proceedings of the 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), Toyama, Japan, 5–8 December 2018; pp. 928–931.
61. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the International Conference on Image Processing, ICIP, Phoenix, AZ, USA, 25–28 September 2016.
62. Goli, S.A.; Far, B.H.; Fapojuwo, A.O. An accurate multi-sensor multi-target localization method for cooperating vehicles. In *Theoretical Information Reuse and Integration*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 197–217.
63. Wojke, N.; Bewley, A. Deep cosine metric learning for person re-identification. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, Lake Tahoe, NV, USA, 12–15 March 2018.
64. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the International Conference on Image Processing, ICIP, Beijing, China, 17–20 September 2018.
65. Liu, X.; Liu, W.; Mei, T.; Ma, H. PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Trans. Multimed.* **2018**, *20*, 645–658. [CrossRef]
66. Zhu, J.; Zeng, H.; Huang, J.; Liao, S.; Lei, Z.; Cai, C.; Zheng, L. Vehicle re-identification using quadruple directional deep learning features. *IEEE Trans. Intell. Transp. Syst.* **2019**, 1–11. [CrossRef]
67. Luo, W.; Yang, B.; Urtasun, R. Fast and furious: Real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
68. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In *Computer Vision–ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 868–884.
69. Li, J.; Chen, S.; Zhang, F.; Li, E.; Yang, T.; Lu, Z. An adaptive framework for multi-vehicle ground speed estimation in airborne videos. *Remote Sens.* **2019**, *11*, 1241. [CrossRef]
70. Lyu, S.; Chang, M.-C.; Du, D.; Wen, L.; Qi, H.; Li, Y.; Wei, Y.; Ke, L.; Hu, T.; Del Coco, M.; et al. UA-DETRAC 2017: Report of AVSS2017 & IWT4S challenge on advanced traffic monitoring. In Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–7.
71. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. *AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles*; Springer: Cham, Switzerland, 2018.
72. Niu, H.; Gonzalez-Prelcic, N.; Heath, R.W. A UAV-based traffic monitoring system-invited paper. In Proceedings of the 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, Portugal, 3–6 June 2018; pp. 1–5.
73. Wang, L.; Chen, F.; Yin, H. Detecting and tracking vehicles in traffic by unmanned aerial vehicles. *Autom. Constr.* **2016**, *72*, 294–308. [CrossRef]
74. Leitloff, J.; Rosenbaum, D.; Kurz, F.; Meynberg, O.; Reinartz, P. An operational system for estimating road traffic information from aerial images. *Remote Sens.* **2014**, *6*, 11315–11341. [CrossRef]
75. Heintz, F.; Rudol, P.; Doherty, P. From images to traffic behavior—A UAV tracking and monitoring application. In Proceedings of the FUSION 2007–2007 10th International Conference on Information Fusion, Quebec, QC, Canada, 9–12 July 2007.
76. Liu, F.; Liu, X.; Luo, P.; Yang, Y.; Shi, D. A new method used in moving vehicle information acquisition from aerial surveillance with a UAV. In *Advances on Digital Television and Wireless Multimedia Communications*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 67–72.
77. Cao, X.; Wu, C.; Lan, J.; Yan, P.; Li, X. Vehicle detection and motion analysis in low-altitude airborne video under urban environment. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1522–1533. [CrossRef]

78. Ren, W.; Beard, R.W. Trajectory tracking for unmanned air vehicles with velocity and heading rate constraints. *IEEE Trans. Control Syst. Technol.* **2004**, *12*, 706–716. [CrossRef]
79. Cao, X.; Gao, C.; Lan, J.; Yuan, Y.; Yan, P. Ego motion guided particle filter for vehicle tracking in airborne videos. *Neurocomputing* **2014**, *124*, 168–177. [CrossRef]
80. Cao, X.; Lan, J.; Yan, P.; Li, X. Vehicle detection and tracking in airborne videos by multi-motion layer analysis. *Mach. Vis. Appl.* **2012**, *23*, 921–935. [CrossRef]
81. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3D Object representations for fine-grained categorization. In Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 1–8 December 2013; pp. 554–561.
82. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
83. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. *arXiv* **2016**, arXiv:1605.07146.
84. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
85. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision—ECCV 2016 Workshops*; Hua, G., Jégou, H., Eds.; Springer: Cham, Switzerland, 2016; pp. 17–35.
86. Xiang, Y.; Alahi, A.; Savarese, S. Learning to track: Online multi-object tracking by decision making. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) IEEE, Washington, DC, USA, 7–13 December 2015; pp. 4705–4713.