Traffic monitoring and queue identification with UAV footage in the congested urban networks

Sohyeong Kim, LUTS EPFL
Emmanouil Barmpounakis, LUTS EPFL
Nikolas Geroliminis, LUTS EPFL

Conference Paper STRC 2021

STRC | **21st Swiss Transport Research Conference**
Monte Verità / Ascona, September 12 – 14, 2021

Traffic monitoring and queue identification with UAV footage in the congested urban networks

Sohyeong Kim
LUTS EPFL
Lausanne
T: +41 21 693 6379
E: sohyeong.kim@epfl.ch

Emmanouil Barmpounakis
LUTS EPFL
Lausanne
T: +41 21 693 5397
E: manos.barmpounakis@ethz.ch

Nikolas Geroliminis
LUTS EPFL
Lausanne
T: +41 21 693 2481
E: nikolas.geroliminis@epfl.ch

August 2021

Abstract

Vision-based analysis of traffic has become essential for Intelligent Transportation Systems (ITS) due to the wide variety of applications such as traffic congestion detection, traffic anomaly detection, and travel time prediction. While the most common monitoring traffic methods are based on fixed road monitoring cameras, unmanned aerial vehicles (UAV) have advantages such as broader vision, flexibility, and cost-efficiency compared with fixed road cameras. UAV-based traffic analysis is useful, especially for the macroscopic analysis of urban roads. Recently, pNEUMA dataset is the first of its kind that offers the large-scale urban traffic data collected from multiple UAVs in the form of vehicle's trajectories. In this paper, we present extended version of the pNEUMA dataset called "pNEUMA Vision" which incorporates imagery data and additional vehicle's annotations. We describe the processing procedure of the raw data, available information and annotations of pNEUMA Vision, and the identified noises in this new dataset. We also propose a method of estimating the number of vehicles on the road through image segmentation approach by estimating the density maps of input images. We analyze the result to show our approach can be useful for identifying the traffic queues in the network.

Keywords

UAV, Urban networks, Traffic monitoring, Traffic vision dataset

1 Introduction

Since the expansion of road infrastructure is limited and expensive, the continuous increase in the number of vehicles results in high traffic volumes and congestion levels. Consequently, innovative and effective measures have to be taken to tackle the challenges of ensuring efficient and optimal use of the existing network. For this purpose, the intelligent transportation systems (ITS) are employed to monitor, analyze, and control the traffic.

The development and the improvement of both the microscopic and the macroscopic traffic models are critical for managing the networks using ITS. The task of collecting traffic data and its analysis is a vital element for developing the traffic simulation models. While the literature in microscopic models, such as car-following or lane-changing is vast, most of the existing models have been calibrated with freeway data, as there are not many complete and accurate trajectory datasets for congested urban networks. The most commonly used devices for collecting traffic conditions are induction loops, overhead radar sensors, and fixed video camera systems. Although such traditional devices have provided accurate and useful data, they could only measure the limited area of the traffic networks. This results in several missing points within the networks as the detectors could not cover the whole networkCoifman *et al.* (2006)Barmpounakis *et al.* (2016a). Some other methods use advanced techniques resulting in the big dataset, such as probe vehicles with GPS, vehicle-to-infrastructure (V2I), and smartphone sensor information. These methods usually require manual detection by individuals, and it is not always easy to convert such data to meaningful traffic dataVlahogianni (2015). The use of GPS technology may not be accurate enough for vehicles locations, especially in the metropolitan areas. It is also harder to study the drivers' behavior if the drivers are aware that they are being monitoredSalvo *et al.* (2014b)Barmpounakis *et al.* (2016a).

The recent advancement of the unmanned aerial systems (UAS), the so-called drones, is considered an emerging technology that could benefit transportation domains based on collecting traffic data through cameras mounted on drones. Thanks to their flexibility in their navigation and lower deployment cost compared to the traditional traffic monitoring devices, we could monitor more expansive areas from the top-down view. They can also travel at higher speeds than vehicles on the roads and are not restricted to traveling on the road networkCoifman *et al.* (2006). The influence of local disturbances at the urban scale, such as when vehicles change lanes to turn or overpass a taxi or bus that stopped to serve passengers is still not well understood. To further advance this direction of research a complete monitoring of the surrounding environment both over time and space is necessary, which is impossible with sparse loop detectors or low penetration GPS sensors.

The videos captured from the drone camera can be processed, and the extracted traffic data can be then used for various purposes such as monitoring and surveillance of the traffic streams, recognizing abnormal situations, aid in managing traffic congestion, driving behavior analysisBarmpounakis *et al.* (2016b). Among many UAS-based traffic datasets, pNEUMA is the most extensive urban traffic data presented by Barmpounakis and Geroliminis (2020). It is collected from a field experiment that utilized a swarm of 10 drones. This work was the first of its kind that provides details of congestion propagation in the urban networks covering more than 100km lanes of the roads, including around 100 intersections and almost half a million trajectories in Athenes' central business district.

While there is a lack of large-scale urban network data collected from drones, pNEUMA provides only the trajectory information of the vehicles seen in the collected videos. The imagery traffic data has not yet been used for research using computer vision. However, seeing drones as a future technology that can improve traffic monitoring and control, it is intriguing to model an architecture that outputs useful traffic information from the input videos directly, without extracting the trajectory of individual vehicles. From this point of view, we have been inspired to develop models that can estimate important traffic parameters such as the density of the congested network directly from the videos captured from drones in an urban environment.

This paper is structured as follows: Section 2 reviews how UAS technology has been employed in state-of-the-art transportation research. Section 3 is dedicated to pNEUMA Vision, the new dataset. Section 4 provides experiment details of our proposed density map based vehicle counting method. The experiment result and discussions for identifying the traffic queues in the network are presented in Section 5. Finally, we conclude this paper in Section 6 with future research opportunities with pNEUMA Vision.

2 Background

UAS are considered an emerging technology that can overcome traditional traffic sensors such as induction loops, overhead radar sensors, and fixed video camera systems. These kinds of sensors could only acquire information within the limited area and thus missing a lot of useful information. Compared to traditional traffic sensors, UAS have advantages in terms of operating cost, flexibility of data acquisition. Drones provide high-resolution images with plentiful contextual information by encompassing broad traffic areas and thus providing a larger overview of areas of interests.

The early transportation research using drones was focusing on the methodology on how to extract traffic parameters such as flow and density of the roads Salvo *et al.* (2014b)Apeltauer *et al.* (2015)Khan *et al.* (2017). They stabilized and geo-registered the UAS videos to detect and track the moving vehicles by background subtraction methods and optical flow algorithms. The extracted parameters were then analysed by comparing to simulation data or to data acquired by probe vehiclesSalvo *et al.* (2014a). Since deep neural networks became popular for object detection and recognition, there has been a lot of research starting to train the convolutional neural networks with their drone video data to detect and track vehicles for the various traffic applications. Zhu *et al.* (2018) estimated the density of the area by counting the vehicles through vehicle detection. Zhang *et al.* (2019) focused on tracking and measuring the speed of the vehicles to decide the congestion level of the roads. Ahmed *et al.* (2020) studied the effect of queue-jumping phenomenon of the motorcycle on traffic flow within the green time. Khan *et al.* (2018) identified the flow state and the shockwaves at the signalized intersection through extracting vehicle trajectories from the videos they have collected.

The data collected from drones play a crucial role in the development of deep learning based vehicle detection algorithm for many traffic application research as mentioned before. Many of the research presents their own dataset collected with the purpose for their application. Usually, these datasets are limited in the scene variance and in scale so that this makes it difficult to adopt this dataset for different kind of traffic application algorithm. There also exists a large-scale UAS-based dataset commonly used as a benchmark for deep-learning based vehicle detection algorithm. The SDDRobicquet *et al.* (2016) consists of videos of eight unique scenes over the Stanford University Campus with a resolution of 1400 x 1904 pixels. The dataset comprises various annotations, including 11.2K pedestrians, 6.4K bicyclists, 1.3K cars, 0.3K skateboarders, 0.2K golf carts, and 0.1K buses. The CARPKHsieh *et al.* (2017) is a UAS-based car parking dataset captured from different parking lots which consists of around 90K cars annotated. This dataset is widely used for vehicle detection in the parking areas. The VisDrone 2020Zhu *et al.* (2020) is a UAS-based dataset consists of 400 videos around 265K frames with a resolution of 3840 x 2160 and 10.2K static images with a resolution of 2000 x 1500. The videos are captured both urban and country scenarios from 14 cities in China with different drone platforms under various weather and lighting conditions. This dataset aims to tackle the tasks such as object detection and single and multiple object tracking. The UAVDT datasetDu *et al.* (2018) consists of 80K frames from 10 hours of raw videos captured by drones in various complex scenarios including squares, arterial streets, toll stations, highways, crossing and T-junctions. The frames are annotated manually along with some attributes such as flying altitude, occlusion, vehicle category, weather condition, and camera view. Similar to VisDrone 2020 dataset, UAVDT dataset also target for the object detection, and single and multiple objects tracking tasks. The highD datasetKrajewski *et al.* (2018) is collected from 6 different locations in German highways using drones. It includes more than 110K vehicle annotations and its trajectory including vehicle type, size and manoeuvres. This dataset is created for the safety validation of highly automated vehicles and other tasks such as the analysis of traffic patterns or the parameterization of driver models.

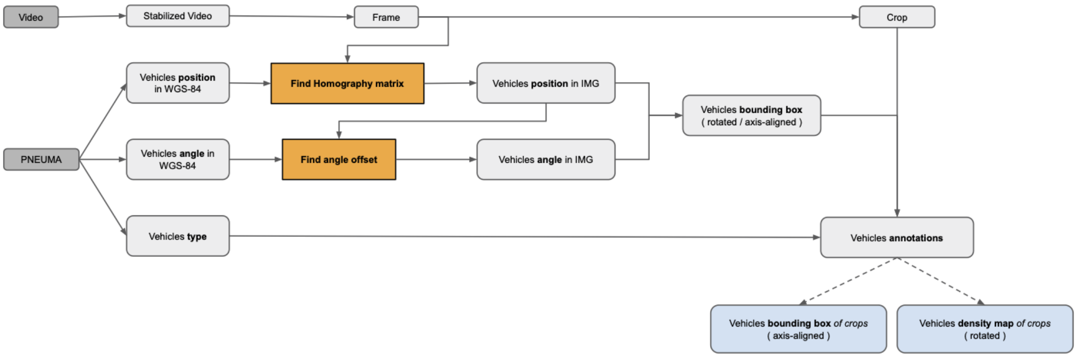
While these datasets are providing abundant data and annotations captured from drones, SSD and CARPK datasets are limited to campus and parking lot scenes so that they are not totally fit for urban traffic network research. Similarly, although VisDrone 2020, UAVDT, and highD dataset captured the actual traffic stream, their scenes are limited to one intersection in urban roads or a specific highway segment. Since one of the important problems in urban transportation research using UAS is to inspect the relationship between multiple sections of the network, it is required to have a dataset which covers the large areas of the traffic networks in the same period. The newly introduced pNUEMA Vision dataset can play the role to tackle this research question and not limited to it as it provides real traffic dataset recorded from 10 drones covering more than hundreds of intersection of congested urban network which is its critical advantage that this dataset has over the other existing drone datasets.

3 pNEUMA Vision

With the emergence of machine learning, data becomes even more important for building transportation models, especially for the traffic research using drones. A significant amount of aerial video footage was collected and analysed in order to create the pNEUMA datasetBarmpounakis and Geroliminis (2020). While pNEUMA includes trajectory data from almost every vehicle (cars, buses, motorcycles, taxis etc.) in a complex urban environment, in order to maximize its applicability to researchers from different disciplines, it is necessary to extract, organise, store and arrange the data into the proper formats. Several steps have been done for data processing as shown in Figure 1; video stabilization, frame extraction, coordinate transformation and adjustment from WGS-84 to image coordinates, and vehicle annotations as bounding boxes. This section will describe more details about dataset processing steps, nature of the new dataset, and possible research direction.

3.1 Data Pre-processing Steps

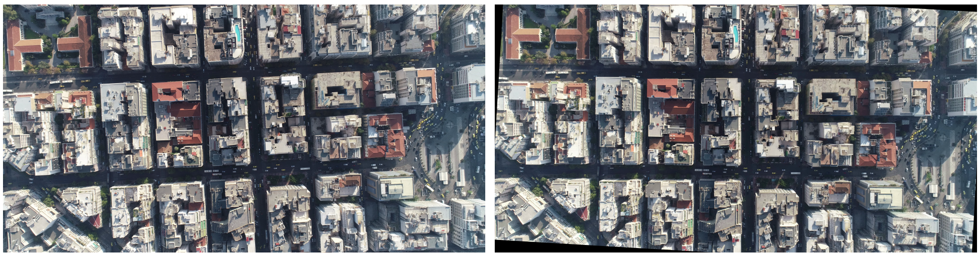
Figure 1: The pipeline for pre-processing pNEUMA for pNEUMA Vision.



3.1.1 Video pre-processing

A certain amount of turbulence occurs when drones are flying, and as a consequence, the videos taken during the flight are unstable. In order to fix this problem, the video stabilization method has been adopted as a pre-processing step. We used the stabilized videos from existing software DataFromSky (2016). The stabilization process matches the central part of the frames throughout the times so that the parts that could not be matched with other frames, mainly border parts of the video, are set to black as shown in Figure 2. Once the videos are stabilized, frames are extracted and saved in a JPEG format.

Figure 2: Comparison between video frames before and after stabilization



(a) Before stabilization (b) After stabilization

3.1.2 Coordinate transformation

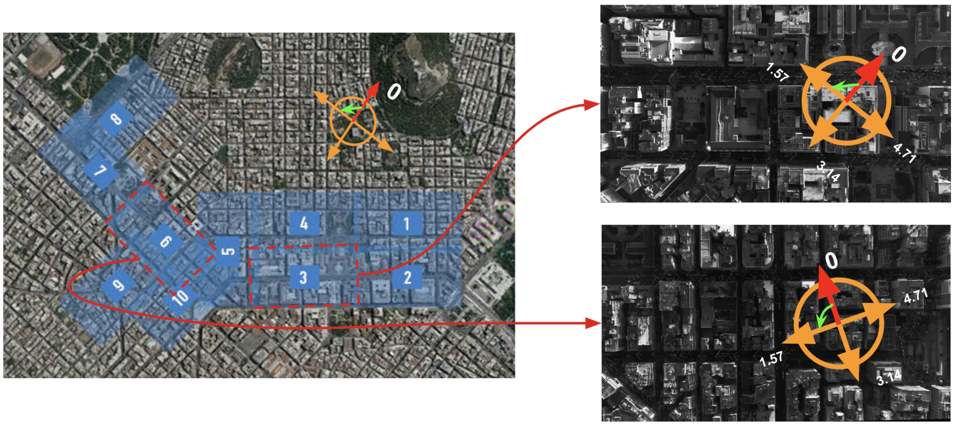
The new dataset, pNEUMA Vision, is basically an imagery dataset so that its annotations have to be in a format in pixels. However, the original pNEUMA dataset only has the annotations in WGS-84 coordinates. Therefore, it is required to conduct a coordinate transformation between real-world coordinate and image coordinate. Assuming that the video is stabilized with high accuracy, we finds the mapping matrix H which maps real-world coordinates (X_{gps}, Y_{gps}) to image coordinates (X_{img}, Y_{img}) using equation 1. Since there were more than hundreds of vehicle per each frames so that it is impossible to find the image coordinates of each vehicles in a video manually. With an assumption that there is a linear transformation between real-world coordinate and image coordinate, knowing exact position in both coordinates of some portion of vehicles in the video will make it possible to derive mapping matrix. We choose more than one-third of vehicles, not limited to a certain area, on a single frame for the better approximation of the mapping matrix. After manually checking their pixel position, the mapping matrix is calculated using least median square method.

$$\begin{bmatrix} X_{img} \\ Y_{img} \\ 1 \end{bmatrix} = H \begin{bmatrix} X_{gps} \\ Y_{gps} \\ 1 \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} X_{gps} \\ Y_{gps} \\ 1 \end{bmatrix} \tag{1}$$

3.1.3 Coordinate angle adjustment

When collecting the dataset, each of the drones was positioned in diverse directions as shown in Figure 3 so that the direction of the vertical axis of each video is not consistent in a real-world coordinate. Since the azimuth of the vehicle is one of the important parameters for the better vehicle annotations, angle adjustment between real-world coordinates and image coordinates for each drone videos is considered.

Figure 3: Diverse direction of each video in pNEUMA in real-world coordinate. (Left) A map of all videos taken. (Right-up) A frame from drone 3 and its direction in real-world coordinate. (Right-bottom) A frame from drone 6 and its direction in real-world coordinate.



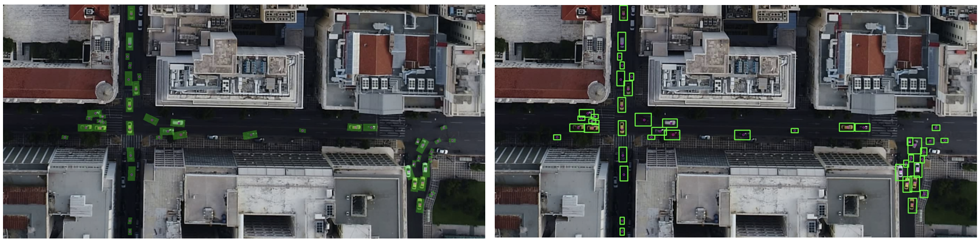
In real-world coordinates, the direction is set in counter clockwise with 0° being the direction to the north in the map whereas in image coordinate, the direction is set clockwise with 0° being the direction of the Y-axis of the image. The equation 2 shows the relationship between azimuth of vehicles in real-world coordinate, θ_{gps}, and the image coordinate, θ_{img}. The α_i is an angle offset for the i-th video and depends on the drone position during data collection this value is different. This angle offset is determined through the comparison between θ_{gps} and θ_{img} derived from a combination of two vehicles in the same frame, considering all the pairs of vehicles within one frame. Once the angle offset is found, we can then calculate the θ_{img} from the real-world azimuth from the pNEUMA dataset.

$$\theta_{img} = f(\theta_{gps}, \alpha_i) = mod_{2\pi}(-\theta_{gps} + \alpha_i) \tag{2}$$

3.1.4 Annotation of vehicles

The final step of the data pre-processing is producing vehicle annotations as of bounding box. The vehicle locations are given as WGS-84 coordinates in the pNEUMA dataset so that putting the tight bounding box around each vehicle requires manual annotation by a human. Even though it is well known that having the data annotation as precise as possible helps in training the deep neural network, we have decided to have less precise bounding-box annotation, yet highly precised as point annotations, on a large scale. Therefore, instead of manually annotating bounding boxes for every vehicle in the videos, the fixed size of bounding boxes is set for each vehicle type. We produced two types of bounding box of the vehicles, one with rotated position and the other with axis-aligned position as shown in Figure 4. To create the rotated bounding boxes, the fixed sized bounding boxes are placed over the vehicles according to the vehicles' positions and they are rotated according to the vehicles' azimuths, which are calculated in the previous step. For bounding boxes of axis-aligned position, we calculated rectangles which encompass the rotated bounding boxes tightly and its width and height are parallel to XY-axis of the frames.

Figure 4: Example of two types of vehicle annotations.



(a) Rotated bounding boxes

(b) Axis-aligned bounding boxes.

3.2 Properties of pNEUMA Vision

3.2.1 Available information and annotations

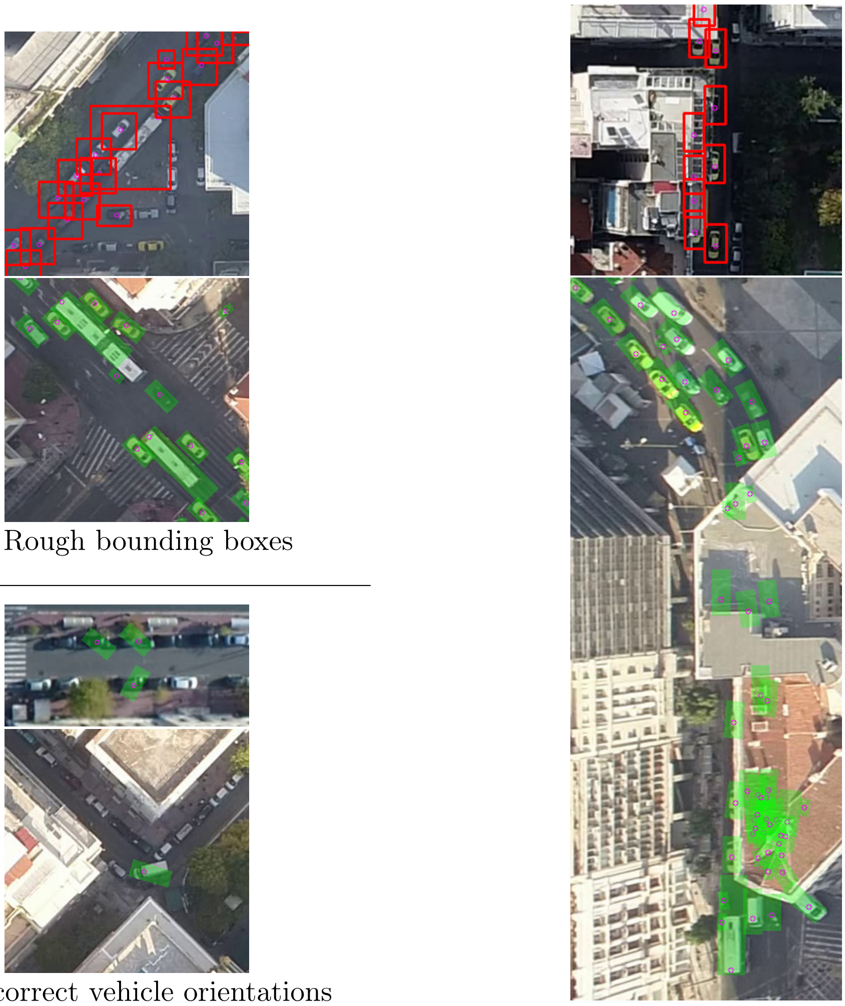
The new dataset pNUEMA Vision can be seen as the expanded version of its predecessor pNEUMA and therefore it inherits much information available in pNEUMA, including vehicle locations, vehicle azimuths, vehicle speeds, vehicle types, and vehicle IDs. Note that vehicle locations, azimuths, and speeds are in a format of pixels in image coordinate. Moreover, pNUEMA Vision contains of 18 sets of 13min length videos with 25 frames per second, and in total 351K frames of size 3840x2100 are available. This dataset is composed of 10 videos from different location above congest urban district of Athens in the morning from 9:00 to 9:30. Also it includes the videos of two regions, taken from drone 2 and drone 3 at 5 different periods in the morning on the same dayBarmpounakis and Geroliminis (2020).

3.2.2 Noises in the bounding box annotations

When producing the bounding box for vehicles, as explained in Section 3.1.4, it was not possible to manually draw bounding boxes for every vehicle. Instead, we rely on the original dataset pNEUMA for their creation, and leaving some noises in the bounding box annotations as a result as shown in Figure 5.

- **Rough size of the vehicles' bounding boxes**
  As we used fixed sized bounding box for each vehicle type, there are some vehicles that does not fit perfectly because they are bigger or smaller than the bounding box. For example, there exist some articulated buses on the road which are longer than the bounding box for the bus type. On the other hand, some compact passenger cars are shorter in length compared to other passenger cars so that their bounding boxes are bigger than themselves. Similar effect happens to heavy and medium trucks since those types of vehicles have different sizes.
- **Incorrect vehicle orientation for temporarily parked vehicles**
  The azimuth of the vehicle is calculated when it starts to move in the video. Therefore, for the vehicles which are temporarily parked on the side lane of the roads, the azimuth has different direction in the video from the direction during the periods it was parked. This happens usually for taxis and buses especially when they are parked in the beginning of the video and start to move in between the recording session. The bounding boxes for those vehicles are in the beginning in a wrong orientation from its actual position. If vehicles are parked throughout the whole recording session, they are not annotated; like in pNEUMA dataset in which they are not tracked at all.
- **Annotations of occluded vehicles**
  In the pNEUMA dataset some of the vehicles are tracked even though they are not

Figure 5: Example of imperfect bounding boxes for both axis-aligned position and rotated position.



(a) Rough bounding boxes

(b) Incorrect vehicle orientations

(c) Annotations of occluded vehicles

visible in the scenes. This may happen if a vehicle appears in the scene for a short period, then gets hidden for various reasons (buildings, trees, etc.) However, it leads to having bounding box annotations over the building at the position where the vehicles are expected. While there are not many occluded spots in the videos, we would like to point this noise for the potential future user of pNEUMA Vision.

4 Experiment

4.1 Method: Density Estimation of the regions

With the advent of image processing techniques with deep learning, many high-performance models have become available to detect or segment objects. Many studies on traffic monitoring using UAS are detecting vehicles in the images first and then compute other traffic information including number of vehicles, speed of the vehicles, etc. Zhu *et al.* (2018) and Khan *et al.* (2018) use deep learning based object detection algorithm using convolutional neural network to detect vehicles to count their number. Based on the vehicle detection result, Zhu *et al.* (2018) computed the density and Khan *et al.* (2018) tracked vehicles and identified the shock wave occurred at one of the lanes in the road. Mou and Zhu (2018) used instance object segmentation methods to count the number of vehicles in the large parking lot. One of the disadvantage of detecting or segmenting the vehicles beforehand is that the noises in the detection or segmentation results such as false positive detection can propagate into the next step of computing traffic parameters such as density. Additionally an extra processing can be required to identify parts of the same vehicle when they show up in the multiple input images. In our method, we trained end-to-end deep neural networks to estimate the density in the region in order to minimize the such error propagation between multiple computing steps.

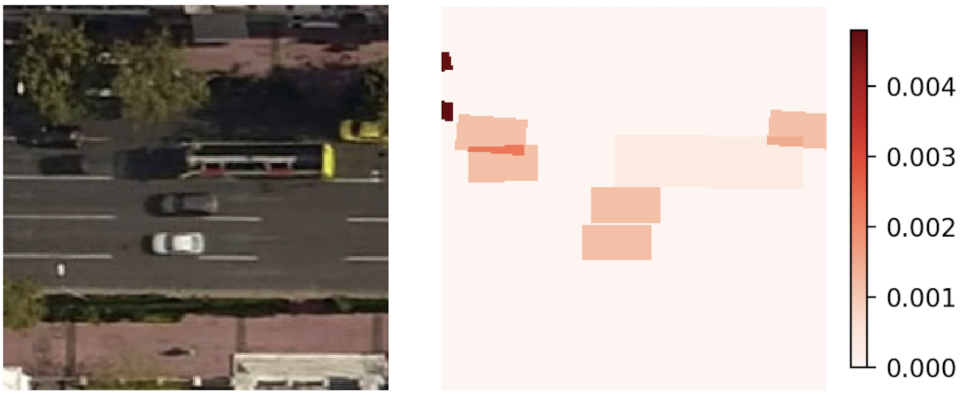
4.1.1 Density map

The density map is often used in crowd counting problem where each position of a person is indicated as a dot with value of 1. In this scenario the density map is generated by applying 2D Gaussian kernel to the ground truth dot mapWan and Chan (2019). Since vehicles are usually much larger in pixel numbers than a person in the crowd and sizes of the vehicles are distinctive from the aerial view, the density map can be defined in another way. We have defined the density map D where each pixel is the sum of proportion of the vehicles sizes and thus sum of the density map is equal to the number of vehicles in the region.

$$D_{i,j} = \sum_k \frac{1}{S_k}, i \in W, andj \in H \qquad and \qquad N_{vehicle} = \sum_i \sum_j D_{i,j} \tag{3}$$

In the equation 3, $D_{i,j}$ is the (i,j) pixel value of the density map D of size W x H. S_k refers to the size of the k-th vehicle whose bounding box is on the (i,j) of the image. The size of the vehicle is set depends on its type. In other words, the larger vehicles such as buses and trucks have smaller pixel values in the density map while smaller vehicles such as cars or motorcycles have larger pixel values. Note that we used the rotated bounding boxes as a reference to create the density map. When two bounding boxes overlap, we sum up all the pixel values so that the number of vehicles in the region does not change. As shown in Figure 6, pixels belongs to a bus have smaller values than that of passenger cars so their pixels are marked as lighter color in the density map. The sum of the pixels in $D_{i,j}$ gives the number of vehicles in the given image $N_{vehicle}$.

Figure 6: Example of density map where the color indicates the value of pixels and smaller vehicles have higher values while larger vehicles have lower values.



4.1.2 Density map based vehicle counting using object segmentation network

In this paper, we propose a framework which predicts the number of vehicle as an output. Zhang *et al.* (2017) modified fully convolutional network for semantic segmentationLong *et al.* (2015) by combining extracted feature maps from shallow layers to improve the segmenting the multi-size vehicles captured from the web camera. Since this web camera has been installed above the road and it creates one-point perspective, the size of the vehicles vary depends on its location due to the point-of-view. For example, the cars that are further away from the camera are under low resolution like ten pixels and suffer from occlusion while the cars that are close to the camera are in high resolution and easy to distinguish. Unlike this study, the vehicles observed by UAS from a top-bottom view like pNEUMA Vision do not suffer from variance in resolution and their size does not depend on their location. Still, however, the size of the vehicles is different depending on their types so we have adopted atrous convolution for semantic image segmentationChen *et al.* (2017) which has significant strengths in extracting features of objects in multiple sizes.

Our proposed network is structured as an encoder-decoder which infers the number of vehicles through predicted density map. Since directly regressing the count of the vehicles can lead to the large errors, we predicted the density map as an intermediary objective to achieve better global optima and thus improving the robustness of the network.

In the encoder part, ResNet101He *et al.* (2015) is used to extract low-level feature maps from the input images and applied atrous convolution for the deeper layers. In the decoder part, multi layers of deconvolution has been used to make the feature maps to be the same size as the ground truth density map. Finally, the decoder generate predicted density map. We applied pixel-wise L2 loss on this predicted density map which is denoted as L_{dm} in Equation 4. Also we also applied L2 loss on the predicted number of vehicles which is denoted as L_{vc} in Equation 4. Then the total loss used to train the network, L, is the weighted sum of L_{dm} and L_{vc}.

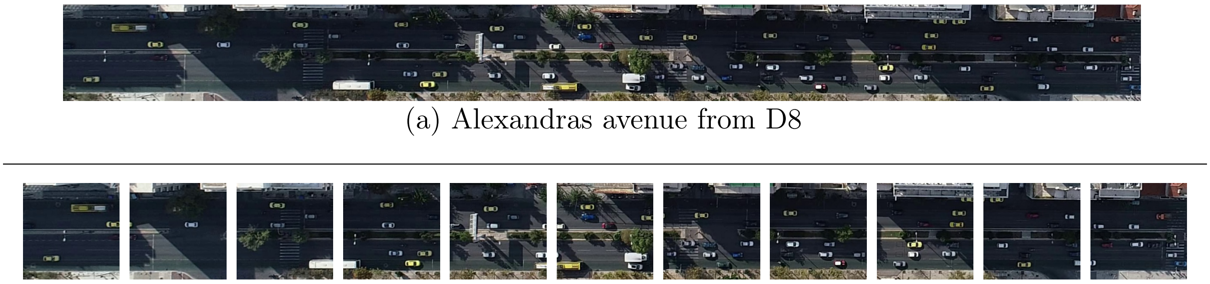
$$L = \lambda_{dm} L_{dm} + \lambda_{vc} L_{vc} where L_{dm} = \sum (D^{gt} - D^{pred})^2 and L_{vc} = \sum (N^{gt}_{vehicle} - N^{pred}_{vehicle})^2 \quad (4)$$

4.2 Setup

We choose two main arterial roads in Athenes center called Panepistimiou street and Alexandras avenue for our experiment. Those streets are captured from D2, D3 and D8. We manually cropped the roads part in a size of 227x227 so that Panepistimiou street is divided into 28 segments of a total length of 840m and Alexandras avenue is divided into

11 segments of a total length of 330m. The part of the Alexandras avenue(D8) chosen for the experiment and its cropped segments are shown in Figure 7. For training the network we used the images of Panepistimiou street(D2 and D3) captured between 8 a.m. to 9:30 a.m. The images of Panepistimiou street(D2 and D3) captured between 9:30 a.m. and 10:00 a.m. are used as a validation set. We composed the two different testing set one from Panepistimiou street(D2 and D3) captured between 10:00 a.m. to 10:30 a.m. and the other from Alexandras avenue captured between 9:30 a.m. and 10:00 a.m.. To summarize, 453k images, 151k images, and 302k images of size 227x227 are used for training, validation, and testing respectively. The network has been trained using GPU model RTX-2080Ti with Ubuntu 18.04 LTS.

Figure 7: The road image and its segments used for the experiment



(a) Alexandras avenue from D8

(b) Examples of the road segments of Alexandras avenue

5 Results and Discussion

5.1 Vehicle counting

The task of our method is counting the number of vehicles in the image and the performance is measured using the mean absolute error(MAE) of the vehicle counting. We analysed the result in two different ways; One is the segment of the road where each segment represents the initial input image to the network of size 227x227 and the other is the integrated road segments which represent the whole arterial roads. In order to compare our result with the most prevalent approach of vehicle number counting which is object detection based counting, we trained YOLOv4Bochkovskiy *et al.* (2020) network which is well known for its object detection performance. We used same training, validation, and testing set for

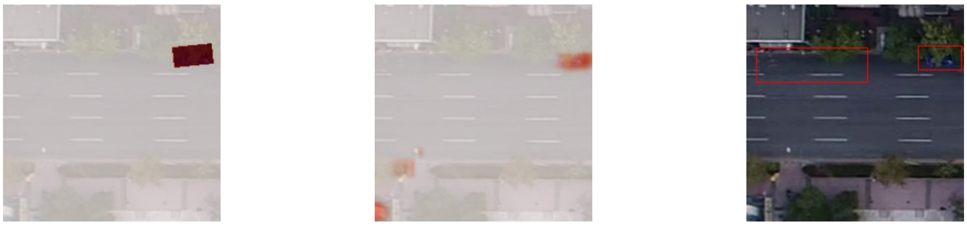
YOLOv4 as we did for our method. The Table 1 shows the comparison of these results.

Table 1: Result of mean absolute error of vehicle counting of each method

		Methods	
Test set	Unit	DMM	VDM
Panepistimiou street (D2, D3)	Road segment	**1.041**	1.3033
	Integrated road segment	**10.203**	17.778
Alexandras avenue (D8)	Road segment	1.533	**1.349**
	Integrated road segment	**5.450**	6.653

From the result of a single road segment, we could observe that our proposed method, density map based method (DMM), could count the number of vehicle slightly better than the vehicle detection based method (VDM) does. Note that in the ground truth, each road segment contained six vehicles on average for Panepistimiou street and nine vehicles on average for Alexandras avenue. Since the training set is composed with roads from D2 and D3, the performance is slightly better when the testing set is also from the same scene(D2 and D3 - Panepistimiou street) for both methods. The false positive detections as shown in Figure 8 were the main reason for having errors in counting vehicles.

Figure 8: Example of false positive case.

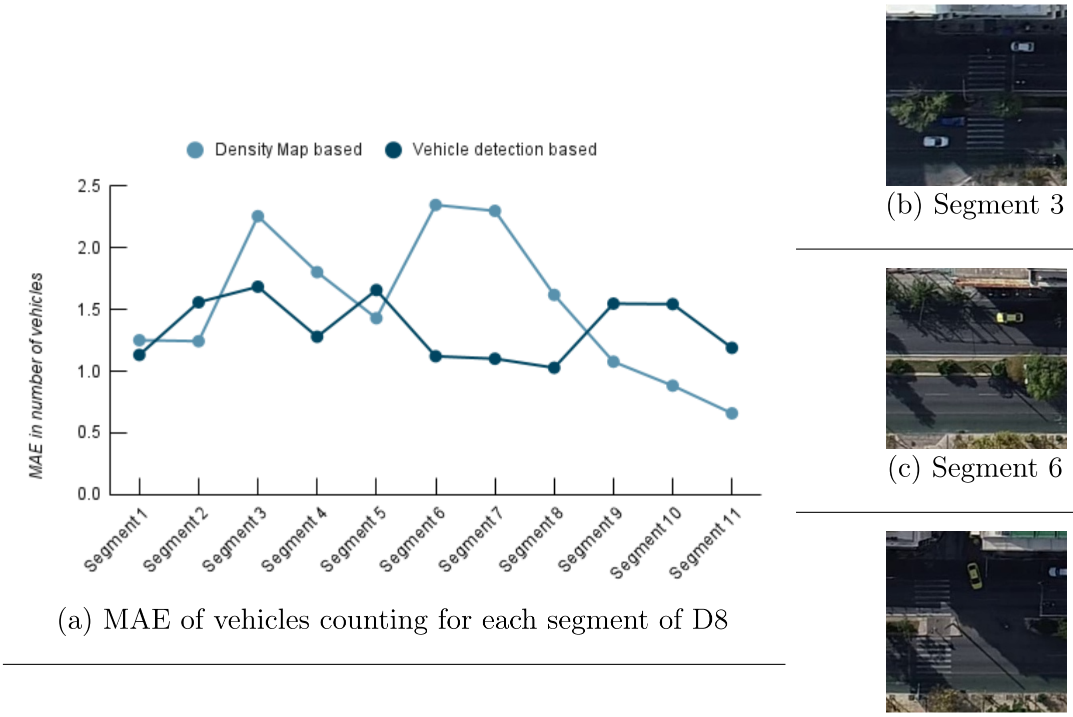


(a) Ground truth (b) Predicted density map (c) Vehicle detection result

In the investigation of the MAE for each segment from the Alexandras avenue(D8), we could observe that in certain segments are suffering from high error in vehicle counting, especially for the density map based approach. As shown in Figure 9, the some road segments which have MAE higher than two, include obstacles in the middle of roads such as trees occluding some part of road. Especially when the vehicles are occluded under the trees, it is difficult to identify them with the specific trained network, whereas the vehicles still exist in the ground-truth annotations.

From the result of an integrated road segment, we could observe that our proposed method could count the number of vehicle better than the VDM does for both streets. Note that

Figure 9: The road segments of D8 that suffer from high errors(higher than 2) with DMM.



(a) MAE of vehicles counting for each segment of D8

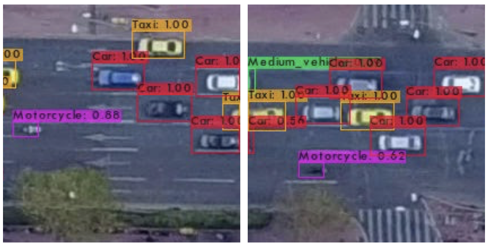
(b) Segment 3

(c) Segment 6

(d) Segment 7

in the ground truth, there were 174 vehicles on average on Panepistimiou street and 99 vehicles on average for Alexandras avenue. When a vehicle is located in two segments, as shown in Figure 10, the VDM detects and counts both part of the vehicles as one and thus double counting for a single vehicle. On the other hand, our proposed method gives proportional value of the car so that it suffers less from double counting when aggregated all the segments of the road. This feature is useful when we have limitation for the input image size for the network as there is no need for the post processing of the outcome.

Figure 10: Example of double counting of vehicles with vehicle detection method.



5.2 Queue identification on the lane

One of the main advantages of the DMM over VDM is that it provides how much the road is occupied by the vehicles. Unlike VDM which counts each detection of vehicle as one vehicle, DMM is trained to learn how much portion of vehicle belongs to each pixels.

In this section, we have extended our results for analysing queues which is considered as one of the important phenomena in urban traffic management. We can infer the status of queues by the number of vehicles in the portion of a lane in a certain period of time. We can assume that the queue is building if there are more vehicles in the region for some period of time. When the number of vehicles are decreasing, we can assume that the queue is dissolving. In this sense, we have applied moving average onto the results to analyze the change in the amount of vehicles on the road throughout the time. The window size for moving average has been set as 100 pixels in road length and the step size as 30 pixels. Considering that normal passenger vehicles and buses are approximately 40 and 110 pixels in length respectively, the expected number of vehicles in this size of window is normally between one to two depending on the size of vehicles.

The Figure 11 shows some lanes from Alexandras avenue(D8) which we used for our analysis. We cropped the region of each lane from the ground truth, the predicted density map, and the vehicle detection result from YOLOv4 and then calculated the average number of vehicles within the pre-defined window for moving average along the whole road. The results of applying moving average to identify the queues in the lane are shown in Figure 12. Note that the travel direction is from 0 to 2500th pixel and the lane 1, 2, and 3 are colored in blue, magenta, and yellow in the Figure 11 respectively.

Figure 11: Lanes in the Alexandras avenue(D8) which used to analyze the queue.

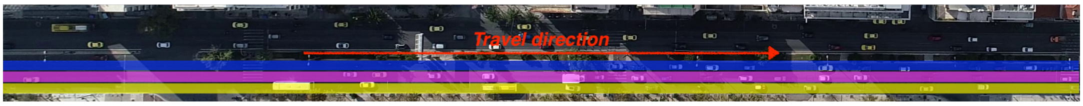
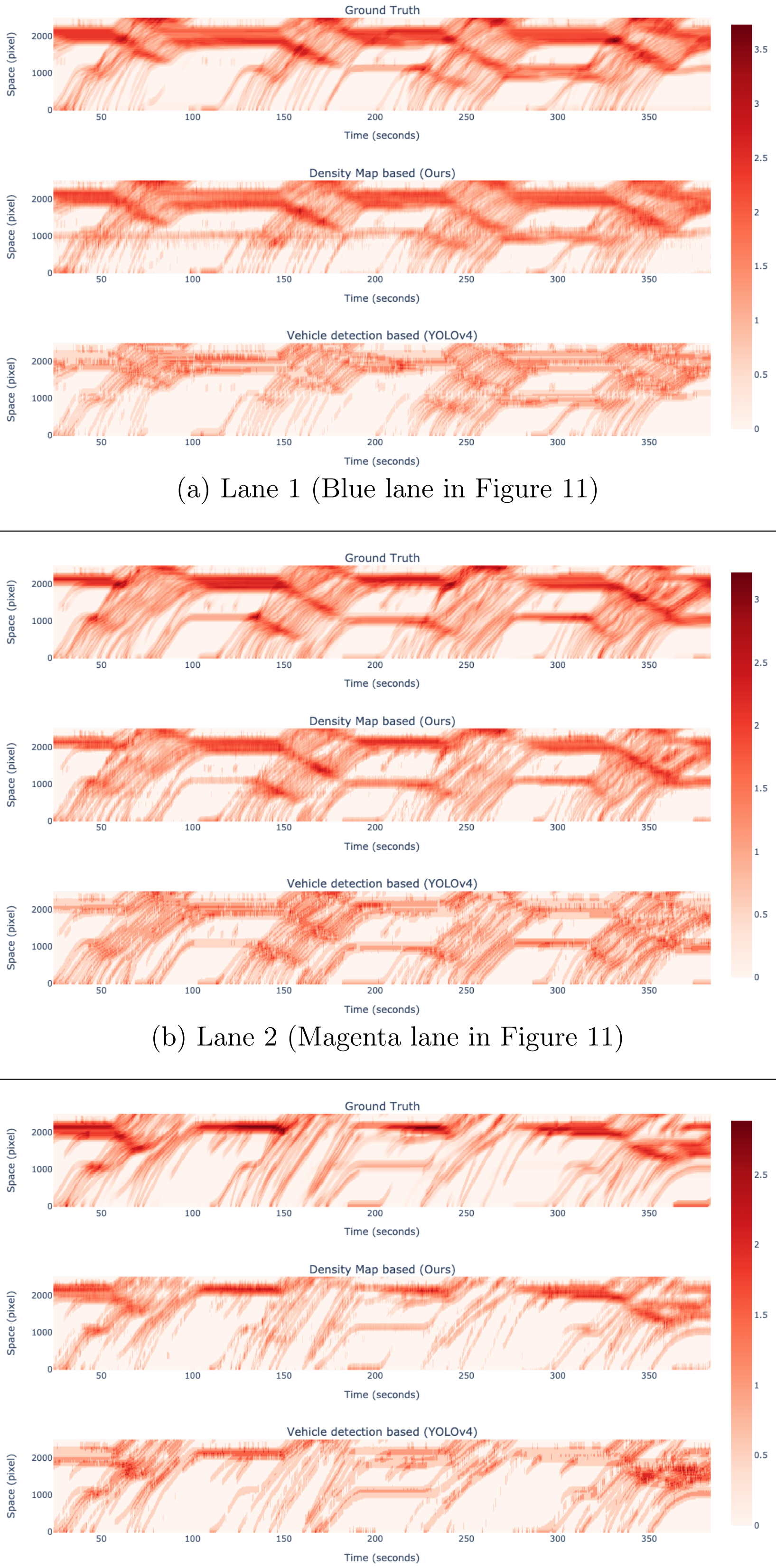


Figure 12: Time-Space-Density diagram for lanes in D8 indicated in Figure 11. The color indicates the number of vehicle in the spatial window.



(a) Lane 1 (Blue lane in Figure 11)

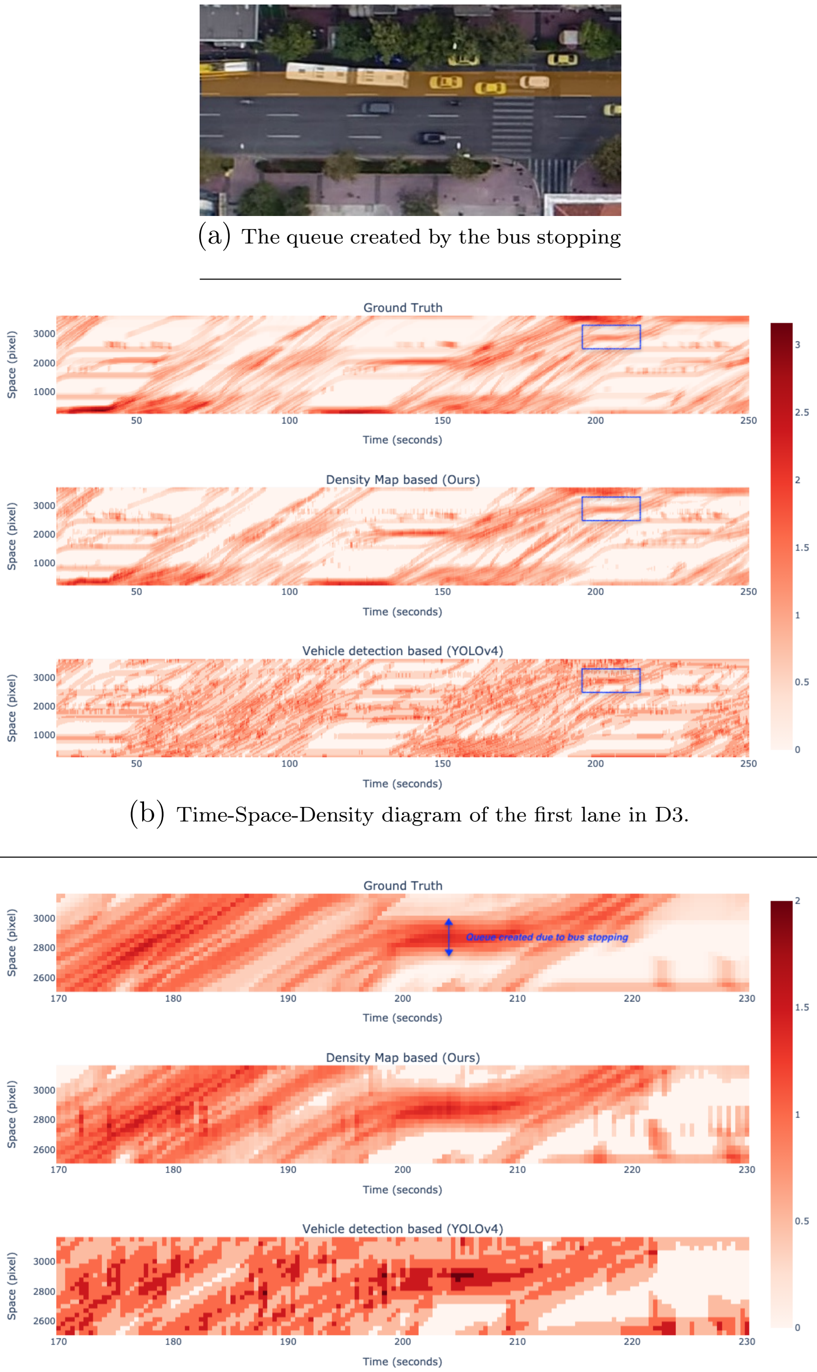
(b) Lane 2 (Magenta lane in Figure 11)

(c) Lane 3 (Yellow lane in Figure 11)

The Time-Space-Density diagrams from Figure 12 clearly show how traffic evolves throughout the time. Also our proposed method better indicates the region where the densities are higher than the VDM. With VDM, it is difficult to know how much each detected bounding box corresponds to how much portion of the vehicle. Especially when the size of the vehicles varies and there is no prior knowledge of whether the detected vehicle is only the part, it is hard to tell with how many number of vehicles are occupying the road. On the other hand, with DMM, the predicted density map provides knowledge of in what portion of the vehicles takes up each pixel. Therefore, we can directly derive the density of the region without any further processing.

One of the interesting phenomena that we could observe in the result is a queue created due to a bus decelerating and stopping to pick up passengers. As shown in Figure 13 (a), the white bus in the first lane has decelerated to stop at the bus stop followed by two taxis. Since there was another bus (the yellow bus in front of it) which was picking up the passengers, the white one should stop for some seconds on the lane and thus create the queues for the vehicles behind it. Even though this is a small example of the queue created by the bus stopping, we could clearly spot out this event from the Time-Space-Density diagram in Figure 13 (b) and (c). The zoomed-in view of the diagram clearly shows how the DMM can illustrate the queue formation better than the VDM, as the actual vehicles' dimensions and spacings are also identified with the DMM.

Figure 13: Queue identified from the Time-Space-Density diagram due to the bus stopping.



(a) The queue created by the bus stopping

(b) Time-Space-Density diagram of the first lane in D3.

(c) Zoomed-in view of the selected in blue square area.

6 Conclusion

In this paper, we introduce an extensive UAS urban traffic imagery dataset named pNEUMA Vision which includes bounding box annotations of multiple types of vehicles. The opportunities that arise from the pNEUMA Vision dataset are huge as it is one of the largest datasets that can be found in an urban setting. We believe that different applications can arise to enhance our understanding of traffic congestion mechanisms but also to contribute to different topics of sustainable mobility management. Similarly, the computer vision community can have access to a very large amount of labelled data and high-resolution images and increase the accessibility of available datasets in this domain of research.

Specifically, the different types of vehicles that were monitored during the pNEUMA experiment, can turn pNEUMA Vision into a benchmark dataset for vehicle detection, especially from a vertical point-of-view which is not usual in the existing datasets that are used for the same purposes. Cars, taxis, buses, powered-two wheelers (PTWs), bikes, medium and heavy vehicles are all present in the city centre of Athens. Apart from the variety of vehicles, more challenging problems can be tackled, for example different lighting conditions or when the same type of vehicle can appear in different sizes. A characteristic example are buses, because, as discussed before, articulated/mini buses have a bigger/smaller length than normal sized ones, or PTWs with scooters being smaller than big capacity motorcycles.

Another opportunity in vehicle detection arises from the different extracted features of the detected object. In many cities worldwide, taxis except for different driving behaviour in the traffic flow, they also have a characteristic colour compared to the rest of the vehicles. In pNEUMA Vision, all cars with the characteristic yellow colour are classified as taxis and thus similar algorithms can be tested towards this direction. The detailed trajectory data of taxis combined with pNEUMA Vision and spatial data has the potential to disclose important results in terms of their operation, by identifying taxi hot spots in both time and space.

While the above applications are focused on computer vision and do not require any significant background in traffic flow theory or transportation engineering in general, the size of the pNEUMA experiment allows the usability of pNEUMA Vision to be enhanced for traffic related purposes. Firstly, a combination of pNEUMA and pNEUMA Vision can also be used for trajectory data generation. Synthetic vehicle trajectories can be produced similar to the real vehicle trajectoriesChoi *et al.* (2021) or for parts of the network where occlusion does not allow the monitoring of the whole network. Additionally, the identification of queues, bottlenecks and spillbacks can be an extremely useful topic for real-time implementation in traffic management schemes. It should be noted that while one can use the pNEUMA dataset to identify the same phenomena using vehicles' trajectories and the fundamentals of traffic theory, pNEUMA Vision can allow the detection of problems using a theory agnostic approach which can still be extremely useful for managing traffic operationsVlahogianni *et al.* (2021).

By combining the two datasets, significant research opportunities in different aspects of traffic operations can also emerge. One of the main advantages of a multi modal environment is the interactions between buses and the rest of the vehicles. By fusing spatial data from external sources for the identification of bus stops, their duration, and the interactions with the rest of the vehicles can now be studied in detail. For the parts of the network that bus lanes are available, one can quantify the effect of violating vehicles in traffic or the effect of the bus lane in the capacity and thus flow of the arterial during different traffic conditions. In parallel, the identification of illegally parked vehicles could provide results in similar directions. Significant findings can also be revealed regarding car following models, not only for operation purposes but for traffic safety and automated vehicles too.

Additionally, although the signal timing in intersections can already be accurately estimated using solely the pNEUMA dataset, an additional possibility can be detected for traffic safety as one could also easily identify the red-light violations and associate driving risk with traffic conditions or individual driving behaviour.

Finally, another potential opportunity can arise for transportation infrastructure monitoring, for example by associating the horizontal road conditions or road markings with traffic safety indicators. Although remote sensing has been quite popular pavement monitoring and loophole detectionSchnebele *et al.* (2015), similar ideas have been proposed with the utilization of dronesSilva *et al.* (2020). This dataset, except for the identification of such cases for a broad urban road network, it allows their association with multi-modal traffic volumes.

As one of the examples for using pNEUMA Vision, we propose a density map based vehicle counting in the arterial roads. One of the advantages of density map based vehicle number estimation is that it can be directly utilized with traffic engineering tools to estimate important congestion phenomena, identification of the queue on the urban arterial roads. The comparison between density map based method to vehicle detection based method which is the most dominant method for vehicle counting, via Time-Space-Density diagram shows that our method gives better performance in terms of identifying queues.

The future work should be made towards developing a method to automatically identify more traffic related phenomena using the density map based method, for example shockwaves, spillbacks, etc.. As shockwave identification is based on sharp changes in density, false positive or false negative from object detection can influence the accuracy. Hence, it is more desirable to do density estimation directly than do individual vehicle detection to identify the shockwaves.

7 References

Ahmed, A., F. Outay, S. O. R. Zaidi, M. Adnan and D. Ngoduy (2020) Examining queue-jumping phenomenon in heterogeneous traffic stream at signalized intersection using uav-based data, *Personal and Ubiquitous Computing*, 1–16.

Apeltauer, J., A. Babinec, D. Herman and T. Apeltauer (2015) Automatic vehicle trajectory extraction for traffic analysis from aerial video data, *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, **40** (3) 9.

Barmpounakis, E. and N. Geroliminis (2020) On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment, *Transportation research part C: emerging technologies*, **111**, 50–71.

Barmpounakis, E. N., E. I. Vlahogianni and J. C. Golias (2016a) Extracting kinematic characteristics from unmanned aerial vehicles, *Technical Report*.

Barmpounakis, E. N., E. I. Vlahogianni and J. C. Golias (2016b) Unmanned aerial aircraft systems for transportation engineering: Current practice and future challenges, *International Journal of Transportation Science and Technology*, **5** (3) 111–122.

Bochkovskiy, A., C.-Y. Wang and H.-Y. M. Liao (2020) Yolov4: Optimal speed and accuracy of object detection.

Chen, L.-C., G. Papandreou, F. Schroff and H. Adam (2017) Rethinking atrous convolution for semantic image segmentation.

Choi, S., J. Kim and H. Yeo (2021) Trajgail: Generating urban vehicle trajectories using generative adversarial imitation learning, *Transportation Research Part C: Emerging Technologies*, **128**, 103091.

Coifman, B., M. McCord, R. G. Mishalani, M. Iswalt and Y. Ji (2006) Roadway traffic monitoring from an unmanned aerial vehicle, paper presented at the *IEE Proceedings-Intelligent Transport Systems*, vol. 153, 11–20.

DataFromSky (2016) Advanced traffic analysis of aerial video data, *Technical Report*. Accessed on March 30, 2021.

Du, D., Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang and Q. Tian (2018) The unmanned aerial vehicle benchmark: Object detection and tracking, paper presented at the *Proceedings of the European Conference on Computer Vision (ECCV)*, 370–386.

He, K., X. Zhang, S. Ren and J. Sun (2015) Deep residual learning for image recognition.

Hsieh, M.-R., Y.-L. Lin and W. H. Hsu (2017) Drone-based object counting by spatially regularized regional proposal network, paper presented at the *Proceedings of the IEEE International Conference on Computer Vision*, 4145–4153.

Khan, M. A., W. Ectors, T. Bellemans, D. Janssens and G. Wets (2017) Unmanned aerial vehicle–based traffic analysis: Methodological framework for automated multivehicle trajectory extraction, *Transportation research record*, **2626** (1) 25–33.

Khan, M. A., W. Ectors, T. Bellemans, D. Janssens and G. Wets (2018) Unmanned aerial vehicle-based traffic analysis: A case study for shockwave identification and flow parameters estimation at signalized intersections, *Remote Sensing*, **10** (3) 458.

Krajewski, R., J. Bock, L. Kloeker and L. Eckstein (2018) The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems.

Long, J., E. Shelhamer and T. Darrell (2015) Fully convolutional networks for semantic segmentation.

Mou, L. and X. X. Zhu (2018) Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network, *IEEE Transactions on Geoscience and Remote Sensing*, **56** (11) 6699–6711.

Robicquet, A., A. Sadeghian, A. Alahi and S. Savarese (2016) Learning social etiquette: Human trajectory understanding in crowded scenes, paper presented at the *European conference on computer vision*, 549–565.

Salvo, G., L. Caruso and A. Scordo (2014a) Gap acceptance analysis in an urban intersection through a video acquired by an uav, *Recent Advances in Civil Engineering and Mechanics*, 199–205.

Salvo, G., L. Caruso and A. Scordo (2014b) Urban traffic analysis through an uav, *Procedia-Social and Behavioral Sciences*, **111**, 1083–1091.

Schnebele, E., B. Tanyu, G. Cervone and N. Waters (2015) Review of remote sensing methodologies for pavement management and assessment, *European Transport Research Review*, **7** (2) 1–19.

Silva, L. A., H. Sanchez San Blas, D. Peral García, A. Sales Mendes and G. Villarubia González (2020) An architectural multi-agent system for a pavement monitoring system with pothole recognition in uav images, *Sensors*, **20** (21) 6205.

Vlahogianni, E. I. (2015) Computational intelligence and optimization for transportation big data: challenges and opportunities, in *Engineering and Applied Sciences Optimization*, 107–128, Springer.

Vlahogianni, E. I., J. Del Ser, K. Kepaptsoglou and I. Laña (2021) Model free identification of traffic conditions using unmanned aerial vehicles and deep learning, *Journal of Big Data Analytics in Transportation*, **3** (1) 1–13.

Wan, J. and A. B. Chan (2019) Adaptive density map generation for crowd counting, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1130–1139.

Zhang, H., M. Liptrott, N. Bessis and J. Cheng (2019) Real-time traffic analysis using deep learning techniques and uav based video, paper presented at the *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–5.

Zhang, S., G. Wu, J. P. Costeira and J. M. Moura (2017) Understanding traffic density from large-scale web camera data, paper presented at the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5898–5907.

Zhu, J., K. Sun, S. Jia, Q. Li, X. Hou, W. Lin, B. Liu and G. Qiu (2018) Urban traffic density estimation based on ultrahigh-resolution uav video and deep neural network, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **11** (12) 4968–4981.

Zhu, P., L. Wen, D. Du, X. Bian, Q. Hu and H. Ling (2020) Vision meets drones: Past, present and future, *arXiv preprint arXiv:2001.06303*.