2020 IEEE Intelligent Transportation Systems Conference (ITSC)
Rhodes, Greece. September 20-23, 2020 (Virtual)

Presenting the Multi-view Traffic Intersection Dataset (MTID): A Detailed Traffic-Surveillance Dataset

Morten B. Jensen[1], Andreas Møgelmose[2] and Thomas B. Moeslund[2]

***Abstract*— This work presents a novel and unique traffic surveillance dataset, the MTID. When capturing data for traffic surveillance, the positioning of the camera is crucial, and depending on the task different approaches provide different advantages. Multiple viewpoints, however, are rarely compared. Our dataset gives the ability to analyze the difference between two viewpoints in great detail.**

A complex traffic scene has been captured simultaneously from two different viewpoints: that of a camera mounted on existing infrastructure, and an that of a drone. The frames from each video capture have been synchronized in time and all road users have been carefully annotated down to pixel-level accuracy. The dataset consists of 3100 frames from each viewpoint, containing 18883 individual annotations on the pole viewpoint, and 50274 individual annotations on the drone viewpoint. The dataset is freely available online*.

Apart from the dataset, which is our main contribution, we also provide benchmark detection results for four different groups of road users for other researchers to compare their results with. We show that the detection problem is challenging, as we achieve mAPs of only 22.62% and 27.75% using a pre-trained state-of-the-art detector on the two viewpoints.

I. Introduction

Traffic surveillance is a subject of increasing importance, as our cities grow larger and more people become able to afford cars. Traffic surveillance can help not only enforce traffic laws (with e. g. red light cameras), but also in counting traffic and evaluating traffic flows in cities. Information gained from traffic surveillance can be used to improve both throughput and safety on our roads.

From a computer vision point of view, traffic surveillance is very challenging due to the large amount of diverse scenes, objects and in particular objects overlapping each other, resulting in occlusion. The easiest way to prevent occlusion is to carefully consider the placement of sensors. Intuitively, a proper birds-eye view of the infrastructure seems the best for counting and tracking road users. In an attempt to achieve this, the sensors are placed as high as possible, usually on either existing infrastructure, e.g. traffic light poles, or on some portable or temporary pole [9]. In some scenarios, the portable pole can provide a better view-angle in case the existing infrastructure options are limited. These options provide a side-top view-angle, which, while not perfect, gives something akin to a birds-eye viewpoint. For the purposes of this work, there is essentially no difference between the pole and infrastructure mounting options. However, this perspective still proves problematic when large objects, e.g. trucks and buses, occlude large areas in the scene.

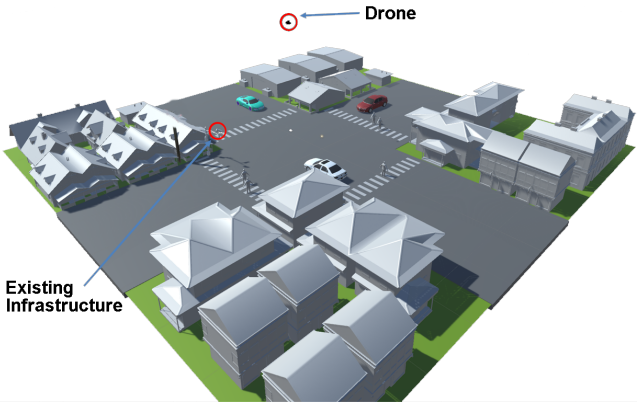


Fig. 1: A traffic intersection equipped with two different capturing viewpoints by mounting a camera in existing infrastructure and using a drone equipped with a camera.

An obvious attempt at a solution to this is to instead put the camera on a drone, which would be able to fly directly above the scene and provide a perfect birds-eye view and thus solve all above issues. This approach, however, is not without issues either. Large objects such as motor vehicles, trucks, and buses are still easily detectable in such footage, but in a direct top-down view, the footprint of vulnerable road users, in particular pedestrians and cyclists, is drastically smaller. An illustration of the infrastructure and drone approaches for capturing data at traffic intersections is shown in figure 1.

This paper attempts to provide insights into the differences between these viewpoints, and its main contribution is a carefully annotated dataset of the same traffic scene captures from both a infrastructure-mounted viewpoint and a drone viewpoint. The dataset has been annotated not only with bounding boxes, but with pixel-level precision.

This dataset is interesting for two main reasons:

1) It provides a unique opportunity to compare two different viewpoints of the same scene in a rigorous manner.
2) It puts forth a very challenging detection and segmentation problem, especially with regards to pedestrians and cyclists in the birds-eye viewpoint.

Apart from providing the dataset and the detailed annotation,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 635895. This publication reflects only the authors' view. The European Commission is not responsible for any use that may be made of the information it contains.

[1]Morten B. Jensen is with BEUMER Group, Aarhus, Denmark mbornoe@gmail.com

[2]Andreas Møgelmose (anmo@create.aau.dk) and Thomas B. Moeslund (tbm@create.aau.dk) are with the Visual Analysis of People Lab, Aalborg University, Denmark

*The dataset can be downloaded from http://vap.aau.dk/mitd

we also show benchmark detection results on four different types of road users for other members of the scientific community to compare against.

II. Related work

Making the computer intelligent and able to reduce the dataset to only interesting time sequences usually refer to the computer vision stages: object detection, classification and tracking. In this paper, we will only look into object detection. Object detection has been a large computer vision area for ages and has to some extend been solved with simple algorithms such as the Viola-Jones framework [17].

The Haar-like features were used for quite some time until Histogram of Oriented Gradients (HOG) was introduced, which when combined with Support Vector Machine (SVM) outperforms the Viola-Jones framework [4]. Additionally, HOG+SVM has in newer updated versions been able to perform very well with a rather small selection of training images due to inclusion of newer optimization methods.

Most recently, of course, deep learning object detectors have emerged and outperformed the traditional detection schemes by far. Examples include Mask R-CNN [7], Single Shot MultiBox Detector (SSD) [12], and You-Only-Look-Once (YOLO) [15].

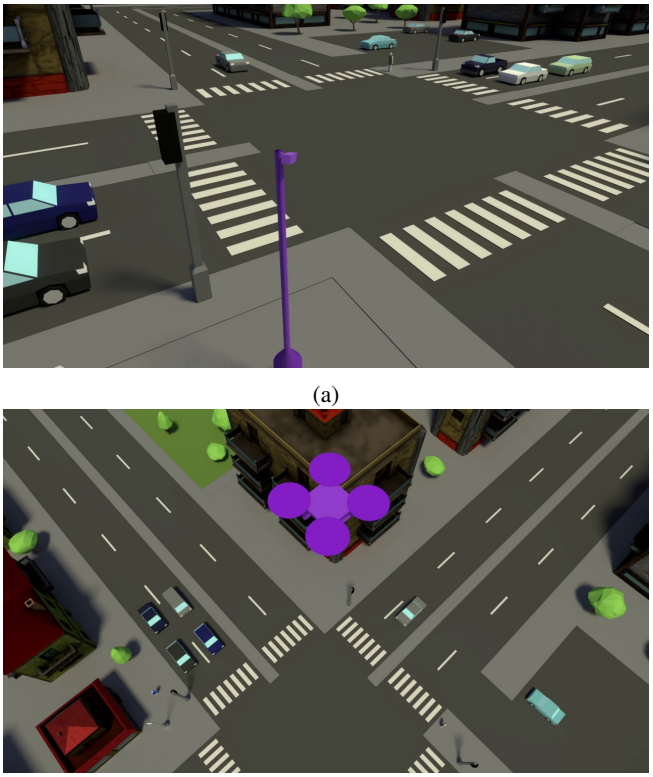
Machine learning is key for these legacy methods, as well as for almost every recent publication in major journals and conferences. Because of this, all of the aforementioned methods rely heavily on the underlying training dataset. In the past decade, multiple datasets and challenges have been offered to the public. This includes general sets, such as Pascal VOC [5], COCO [11], VIRAT [14], and ImageNet [16], but also more application-specific datasets have appeared within the traffic domain, including for example the LISA Traffic Sign [13] and Traffic Lights [8] datasets, KITTI [6], and AAU RainSnow [1]. Drone datasets do exist for the traffic domain. Relevant examples are [10], [2], [3]. None of these provide any comparison to other viewpoints.

In this paper we present a new dataset, which is unique in presenting two different viewpoints on the same scene. It contains a view from existing infrastructure and a view from a drone, as illustrated in figure 1. To compare the view-angles, we annotate the dataset with both bounding box and instance segmentation annotations. We furthermore apply the state-of-the-art object detector Mask R-CNN on the dataset with the purpose of comparing the performance on different objects in different view-angles.

III. Challenges in capturing traffic surveillance data

The current widespread solution for recording traffic video data at traffic intersections is by the use of cameras mounted in existing infrastructure, e.g. poles, as seen in figure 2a.

As mentioned in the introduction, this view-angle is not always ideal as larger objects like buses and trucks will occlude smaller objects. This is illustrated in figure 3, where the red car is completely hidden from view in one viewpoint, while being visible in the other. This occlusion is a quite



(a)

(b)

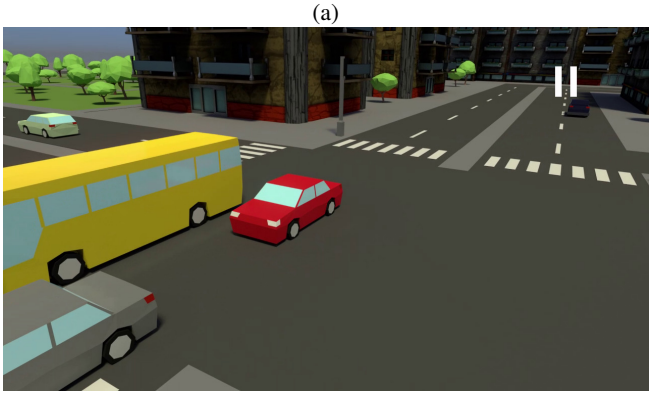
Fig. 2: Two different mounting solutions for traffic surveillance cameras, leading to two vastly different viewpoints on the scene. (a) shows a camera mounted in existing infrastructure. (b) shows a drone hovering above a traffic intersection.

well known problem, which can sometimes be partly solved by deploying a portable pole providing a higher capturing height and a better position [9]. But in order to really get the full overview of a traffic intersection, a drone is an obvious choice given its birds-eye view as illustrated in figure 2b.

Though the drone perspective clearly provides a better overview resulting in fewer occluded objects in the traffic scene, it is not without issues. It is quite difficult to see smaller objects like cyclists and pedestrians. This is sketched in figure 4, and also clear from the actual footage from the dataset shown later in fig. 5 and 6. The drone altitude can somewhat mitigate this, but many countries have strict regulations of drones which defines some restrictions on these possibilities. The same regulations will also often prohibit flying directly over the intersection itself, requiring the drone to stay off to the side of the intersection. This reintroduces some of the occlusion problems from the infrastructure-based camera mounting, as the drone essentially now simply act as a very tall, invisible pole.

The correct choice among the two different viewpoints depends on the requirements of the footage. The drone can provide data very rapidly without any major setup, but will in some jurisdictions require a permission from the local police as well as a licensed pilot to operation the drone. Drones will

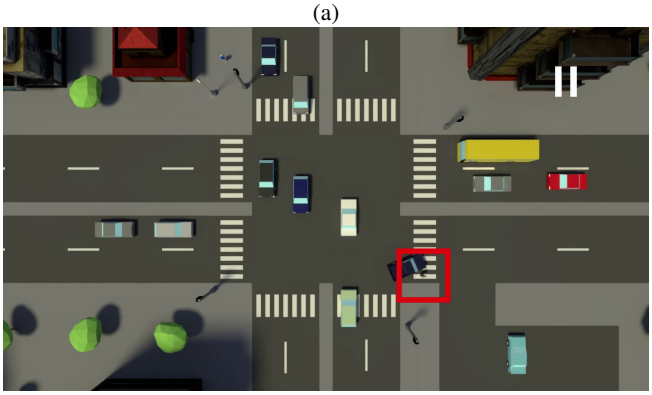
(a)



(b)

Fig. 3: The challenge of capturing reliable data from existing infrastructure. Depending on the location of the camera, entire cars may disappear in occlusions. In this case, the red car is invisible from viewpoint (a), but perfectly visible in viewpoint (b). That does not make viewpoint (b) superior, though, since the same problem would be present there, had the positions of the bus and the red car been swapped.

(a)



(b)

Fig. 4: An illustration of the difficulty with drone-footage. (a) A pedestrian ready to cross the intersection. (b) It is very hard to see and thus detect the pedestrian in the drone view-angle.

also provide a very nice overview of most traffic scenes, but generally do not allow for capturing continuously for a very long period of time. Most batteries in high-end consumer products allow for up to 24-30 minutes of flight time. For some applications and pilot tests this might be sufficient, but for more comprehensive studies, longer recordings are needed. Changing batteries will scale the recording time linearly, but require the drone to land and take-off again. Drones intended for industrial use may allow a power cable attached to it, which allows for longer flight times. Though some expensive industrial drones can remain operational during rainy and windy conditions, most consumer-friendly drones also need to be taken down if the wind speed exceeds 10 m/s or if it starts raining.

So in summary, while the drone seems to provide the best viewpoint (with some caveats, as discussed in section V), it has major drawbacks which will make it useless in certain situations. Sometimes using a drone will be the correct choice, sometimes not.

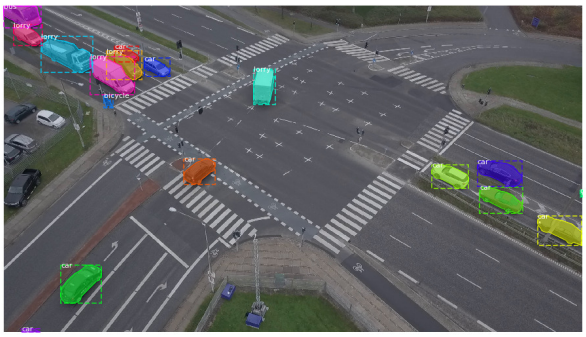
IV. Dataset

The datasets consists of two different view-angles: Existing infrastructure and a drone. The dataset contains 3100 synchronized and annotated frames from each of the two view-angles. Each frame is annotated with both an axis-aligned bounding box as well as a pixel level mask, allowing instance segmentation. The enormous amount of work that is pixel-wise annotation is worthwhile, since the precise positions of the road users are necessary for proper analysis. A diagonally oriented truck will have a very large bounding box, but less than one third of the bounding box is taken up by the actual truck. For collision analysis, simply looking at bounding box overlap is hence not sufficient.

The annotation scheme follows the COCO format and classes, which in this case provides 4 different annotations in the dataset, namely: bicycle, car, bus and lorry. An overview of the specifications of the dataset is shown in table I. The infrastructure viewpoint was captured using an AXIS M1124-E camera with a VGA resolution at 30 FPS. The drone viewpoint was captured using a DJI Mavic Pro Drone using its standard on-board camera. The footage is captured in full HD resolution at 30 FPS. The dataset was captured at an intersection in downtown Aalborg, Denmark.

Sample pictures from each viewpoint are shown in figures 5 and 6. It is worth noting that while permission could not

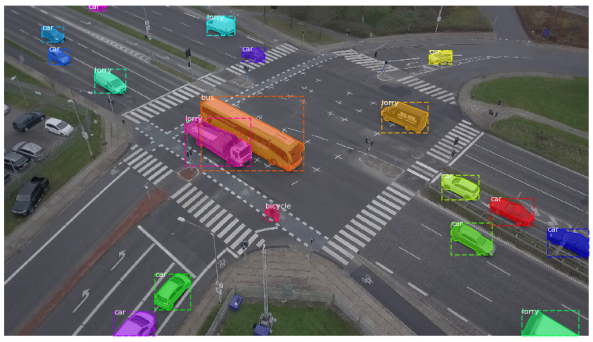
(a) Infrastructure



(b) Drone

Fig. 5: Annotated frame 26 of each of view-angle.

(a) Infrastructure



(b) Drone

Fig. 6: Annotated frame 759 of each of view-angle.

TABLE I: Overview of the dataset.

	Infrastructure viewpoint	**Drone viewpoint**
Resolution	640x480	1920x1080
Number of frames	3100	3100
Bicycles annotated	2003	2909
Cars annotated	9989	32550
Buses annotated	2113	3607
Lorries annotated	4778	11208
Total annotations	**18883**	**50274**

be obtained to fly directly over the intersection, the drone footages does mitigate occlusions to a large extent.

The annotation distribution in the dataset for the infrastructure and drone view-angle are seen in figure 7. We note that the overall shape of the two plots are similar, indicating that they do indeed capture the same scene. There are, however, significantly more annotations in the drone viewpoint. This has two reasons. 1) the drone captures from a much higher vantage point, removing many occlusions, but also 2) the drone simply has a larger field of view.

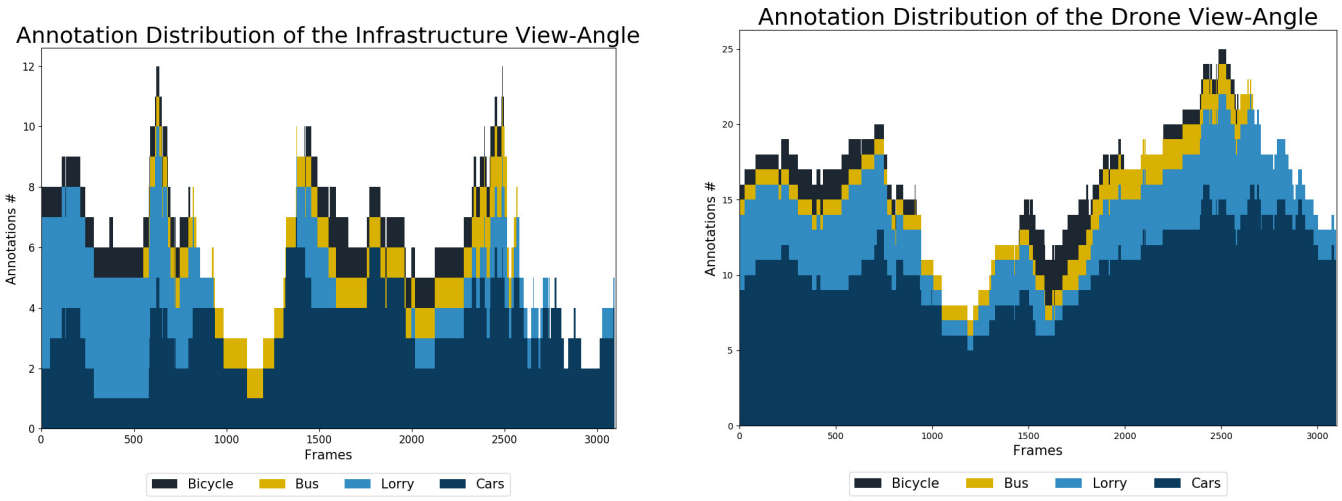
The MTID dataset is freely available at `http://vap.aau.dk/mitd`.

V. Benchmark detection results

For evaluating whether one viewpoint is better for automated detection than the other, we used the Mask R-CNN [7] for object detection. We used a model pre-trained on the Microsoft COCO dataset applied directly to MTID. Given an overlap criterion of 50%, the Mask R-CNN achieved a mean average precision (mAp) of 22.62% for infrastructure and 27.75% for drone, which is shown in the precision-recall curves in figures 8 and 9. Regardless of viewpoint, it is clear from these curves that MTID poses a very difficult detection problem. Mask R-CNN, an otherwise impressive detector, performs poorly on both. This also indicates that while the choice of viewpoint is highly relevant, in particular for human evaluation of footage, both viewpoints are so difficult to handle that automatic detection is hardly feasible - there is definitely an open problem here.

Still, let us have a look at the performance for each viewpoint. Started with the infrastructure-based viewpoint, it is clear from figure 8 that neither of the 4 classes are particularly well-performing. The best performing class is ”car” with an average precision (AP) of 47.49%, followed by bus and lorry on 26.94% and 10.21%, respectively. The Mask R-CNN object detector on the infrastructure view-angle is able to detect a few bicycles, but it turns out to not even be 10% of the total amount of bicycles present.

The same pattern emerges when analyzing the detection performance on the drone viewpoint. Car, bus, and lorry performs the best, and even slightly better than in the infrastructure viewpoint. Still, detection performance is not great in either case. The biggest surprise is that no bicycles at all are successfully detected in the drone viewpoint. This



(a) (b)

Fig. 7: Distribution of annotations for each viewpoint in the dataset. (a) shows the distribution of annotations for the infrastructure-mounted viewpoint, while (b) shows it for the drone viewpoint.

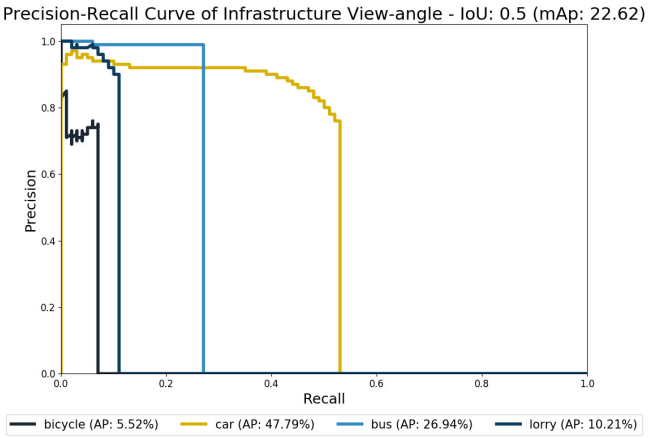


Fig. 8: Precision-recall curve with IoU of 50 % of the infrastructure view-angle.

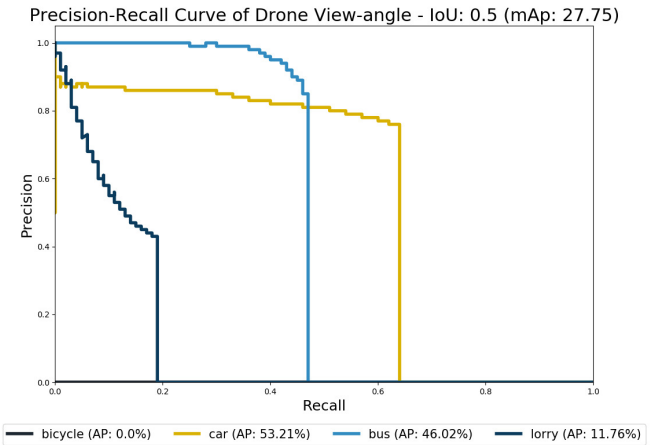


Fig. 9: Precision-recall curve with IoU of 50 % of the drone view-angle.

is rather important, as it negates much of the advantage the drone has. Remember, the main reason for using the drone viewpoint is to minimize lost detections due to occlusions. Bicycles are the smallest of the classes we consider, and hence the class with the biggest risk of occlusions. But if the drone viewpoint means the bicycles are never detected, regardless of occlusions, that point is moot. Obviously, we do not suggest that bicycle detection in the drone viewpoint is impossible, but it is a very challenging task.

Because the biggest issue is with bicycle detection, we will analyze this problem further. Is the bicycle class ever detected in the dataset? In figures 11 and 10, we show detection performance when varying the intersection over union (IoU) or overlap criterion between 0.01 and 1.00. In other words, at 0.01, we allow a very small overlap between detection and ground truth to be considered at true positive, while a value of 1.00 requires perfect overlap between the ground truth and the detection, meaning a perfect detection. A true 1.00 overlap is basically impossible. As we vary the IoU-criterion, we measure the AP and mAP for all the classes.

When examining and comparing the results presented in figure 11 and figure 10, it is clear that even if we allow for detections that barely overlap the ground truths annotations, we still do not detect any bicycles in the drone view-angle. Nor does it significantly improve the AP of the bicycles in the infrastructure view-angle. The take-away from this, is that bicycle detection in the context will at the very least require a detector which is trained (or fine-tuned) for the particular task.

VI. Conclusion

In this paper, we present the MTID: The Multi-view Traffic Intersection Dataset. It is a publicly available dataset providing two different viewpoints of the same traffic scene. Four classes of objects have been annotated throughout the

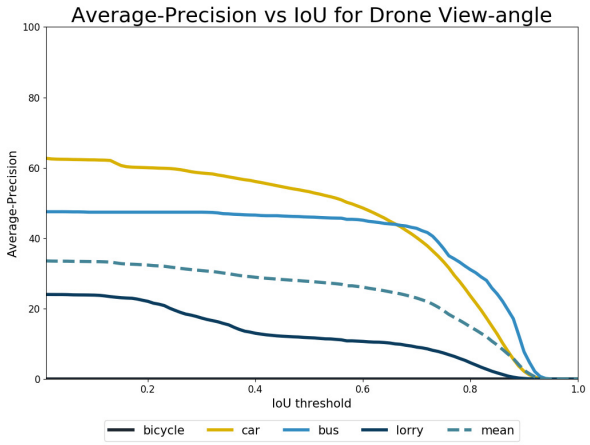


Fig. 10: The AP as a result of varying the IoU criterion in the drone viewpoint.

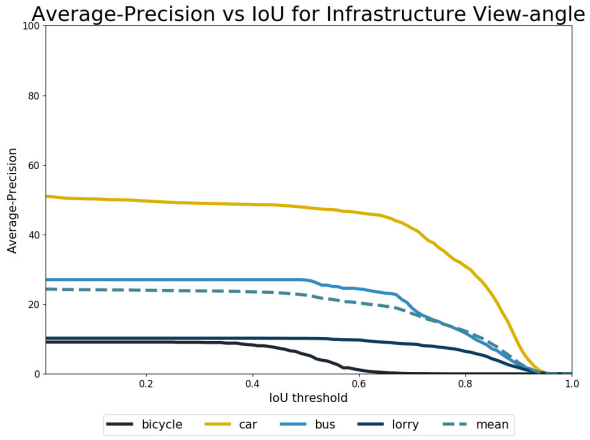


Fig. 11: The AP as a result of varying the IoU criterion in the infrastructure viewpoint.

dataset in both viewpoints: Cars, buses, lorries, and bicycles. 3100 synchronized frames are provided for each viewpoint and in each, all instances of the aforementioned classes have been carefully annotated with both bounding box and at pixel-level. In total 69157 annotations are provided.

We evaluate detection of the four classes in both viewpoints by using a Mask R-CNN pre-trained on the COCO dataset, and find that detection performance for all classes is very poor, regardless of perspective. The infrastructure viewpoint gives an mAP of 22.62%, while the drone viewpoint results in 27.75%. What is also really important to note is that the pre-trained detector failed completely at detecting bicycles.

There are a number of interesting avenues for continuing the work here. First and foremost improving the detection performance, as any other analysis will hinge on better detection. When better detection performance has been obtained, an analysis of whether the drone viewpoint actually helps solve the occlusion problem would be very interesting. This would require linking detections (or annotations) across the two viewpoints and a method of detecting when occlusion occurs. Finally, the dataset can be used to detect various traffic patterns and conflicts.

References

[1] Chris H Bahnsen and Thomas B Moeslund. Rain removal in traffic surveillance: Does it matter? *IEEE Transactions on Intelligent Transportation Systems*, 20(8):2802–2819, 2018.

[2] Margherita Bonetto, Pavel Korshunov, Giovanni Ramponi, and Touradj Ebrahimi. Privacy in mini-drone based video surveillance. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 4, pages 1–6. IEEE, 2015.

[3] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *IEEE International Conference on Robotics and Automation*, February 2020.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan 2015.

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[8] M. B. Jensen, M. P. Philipsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi. Vision for looking at traffic lights: Issues, survey, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):1800–1815, July 2016.

[9] Morten Bornø Jensen, Chris Holmberg Bahnsen, Harry Spaabæk Lahrmann, Tanja Kidholm Osmann Madsen, and Thomas B. Moeslund. Collecting traffic video data using portable poles: Survey, proposal, and analysis. *Journal of Transportation Technologies*, 8(4), 2018.

[10] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018.

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. *SSD: Single Shot MultiBox Detector*, pages 21–37. Springer International Publishing, Cham, 2016.

[13] Andreas Møgelmose, Mohan Manubhai Trivedi, and Thomas B Moeslund. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1484–1497, 2012.

[14] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, June 2011.

[15] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. 1:I–I–518, 2001.