

Truck traffic monitoring with satellite images

Author:

Lynn H. Kaack
Department of Engineering and Public Policy
kaack@cmu.edu

DAP Advisors:

George H. Chen
Heinz College of Information Systems and Public Policy
Machine Learning Department
georgechen@cmu.edu

M. Granger Morgan
Department of Engineering and Public Policy
granger.morgan@andrew.cmu.edu

November 28, 2018

Abstract

The road freight sector is responsible for a large and growing share of greenhouse gas emissions, but reliable data on the amount of freight that is moved on roads are scarce. Many low- and middle-income countries have limited ground-based traffic monitoring and freight surveying activities. We show that we can use an object detection network to count trucks in satellite images and predict average daily truck traffic from those counts. In this proof of concept, we describe a complete model, test the uncertainty of the estimation, and discuss the transfer to developing countries.

1 Introduction

As noted by the United Nations, despite an exponential growth in the availability of data in recent decades, many people and critical aspects of their lives and environment remain unmeasured [13]. Especially across the developing world, a key barrier to identifying opportunities for mitigating climate change is the lack of sufficiently granular, high-quality data. Heavy- and medium-duty trucking accounts for 7% of total world energy-related CO₂ emissions [30], with much of the growth occurring in developing countries [15]. In order to successfully implement policies and make targeted investments, reliable data about the volume of freight that is moved on roads is crucial. More than half of all countries do not collect national road freight activity data and where estimates exist, they are typically survey-based and often inadequate [15]. Knowing truck movements is also important for a variety of economic analyses and for road maintenance planning, even if only based on short-duration counts [33], but such ground-based traffic monitoring is costly and not performed in many countries.

In this paper, we propose a remote sensing approach to obtain vehicle counts from high-resolution satellite images. As satellite images become both cheaper and are taken at a higher resolution over time, we suspect our proposed approach to be scalable at an affordable cost within the next few years to much larger geographic regions. We take advantage of recent advances in deep convolutional neural networks for object detection. These methods have already been successfully applied to detecting vehicles in satellite images [29, 14, 3, 4, 23]. Most work has focused on cars, and to a lesser extent on multiple vehicle classes including trucks [20, 29]. Note that a satellite image is only for a single snapshot in time, whereas conventional traffic estimates are taken over a much longer period of time. Thus, our approach separately models how traffic changes with time.

We begin by providing a brief overview of traditional ground-based traffic monitoring and remote sensing alternatives (Section 2). We then introduce our framework, which consists of a truck detection model (Section 3.1) and a temporal traffic monitoring model (Section 3.2). We validate and test our approach using data on the New York Thruway (Section 4) and assess how the model transfers to data from Brazil (Section 5). We conclude with a qualitative discussion of how well the model translates to developing countries and outline future work (Section 6).

2 Traffic monitoring and freight surveying

The US Federal Highway Administration (FHWA) highlights the importance of vehicle counting for traffic monitoring, as it provides statistics such as the Annual Average Daily Truck Traffic (AADTT) [33]. Ground-based automatic vehicle counting devices include pneumatic tubes, inductive loop detectors, magnetic sensors, video detection systems, and several others. Installation and maintenance for some of these systems requires pavement cuts and lane closures. Traffic monitoring is usually based on continuous counts, which also provide the basis for periodic (e.g., hour of the day) factors applied to short duration counts. Typical short duration detection periods are between 24 hours and a week long [33].

Traffic monitoring with remote sensing As ground-based detection devices can be prone to failure and are too costly to install and maintain in some countries, there is a need for alternative monitoring technologies, such as through GPS data from cell phones [11] or with aerial or high-resolution satellite images. Even lower-resolution satellite images can provide sufficient resolution [19, 7] and there is potential for using drones [16]. With remote sensing, a large number of roads can be covered at the same instance, many of which are not equipped with costly sensors [19, 7] (e.g., rural or remote roads). Also, areas that are difficult to access, for example due to a disaster or conflict, can be monitored [8]. A weakness of the method is that traffic fluctuations on small time scales as well as time-of-day, day-of-week, and seasonal traffic patterns can distort the accuracy of the estimate of the AADTT [33]. In addition, this method requires advanced analytical and computational resources. The uptake of remote sensing methods for transportation applications has been slow but it promises to offer cost-effective and scalable options for a multitude of applications [2, 1].

Freight surveying Data on road freight activity, measured in tonne-km, are typically obtained through national surveys of shipping companies, which need to provide information on origin, destination, weight, and other indicators of all shipments [15]. As the road freight sector is fragmented with most companies operating very few trucks, this approach can be costly and relies on high compliance rates. Less than half of the countries in the world collect this type of information [15].

3 Problem setup

Our framework consists of a truck detection model and freight monitoring model. The detection model counts the number of freight vehicles on roadways in a satellite

image, and the monitoring model translates these counts into the AADTT and other variables of interest.

3.1 Detection model

Object detector The object detection model provides the vehicle count from an image. Huang et al. identified three object detection meta-architectures, which are Faster Region-based Convolutional Neural Networks (Faster R-CNN), Single Shot Detectors (SSD) and Region-based Fully Convolutional Networks (R-FCN) [12]. They have tested models based on these meta-architectures for speed and accuracy and have found that Faster R-CNN often achieves the highest accuracy, while SSD excels in speed. We use Faster R-CNN [28] with 50- and 101-layer Resnets [10] for our application. Faster R-CNN first proposes regions with the Region Proposal Network (RPN) and then uses the Fast R-CNN detector [9] for object detection, sharing convolutional layers. We also use an SSD Inception V2 [21] for comparison. We use the default implementations for the COCO image dataset from the Tensorflow Object Detection API [12] and pre-trained convolutional layers.

Road filter We only want to count trucks that are driving on the road of interest, and exclude those sitting in parking lots or traveling on smaller roads. To filter out irrelevant predictions from the detection model, we use geospatial data. Those data are ubiquitous, and also available for main transit highways in developing countries. We count a truck if at least one corner of its bounding box is within 8 meters of the center of the road, which approximately accommodates a four-lane highway. This filter is applied to both the annotated validation and test datasets and the predictions.

3.2 Freight monitoring model

To use a snapshot image to approximate ground-based vehicle counts, we assume that all n_I vehicles travel at the same constant speed within the interval captured by the image. From that we infer the time t_I that it takes for a vehicle to travel from the start to the end point in the interval. A detector installed in the end point should count n_I vehicles in t_I . The FHWA recommends that traffic density variation factors $f_{h,d,m}$ be applied when using less-than-a-day counts to compute the AADTT [33], so as to account for time-of-day, day-of-the-week and monthly variations. We can approximate the average daily (bidirectional) counts as $AADTT \approx n_I \cdot \frac{24h}{t_I} \cdot f_{h,d,m}$. Detailed information about traffic patterns, and access to satellite images taken at

different times for the same location can reduce the error of the estimate. By making assumptions about the distribution of payloads of the freight vehicles [33], one could also estimate the freight activity through truck counts.

4 Experiments

We have run our experiments on images and toll data of the NY Thruway. We have two kinds of ground-truth data: the true labels in the satellite images and the ground-based counts from toll data. Both of the submodels are independent and we validate them separately to choose the best model specifications and parameters. We then test how well the whole model can estimate the AADTT on a held-out section of the Thruway.

Data We curated our own collection of 31cm-resolution, RGB-color satellite images (Appendix A.2), since a large satellite image database ("xView" [18]) with several thousand labeled truck instances proved too inaccurate and other satellite image datasets contained only small numbers of trucks [23, 27]. For training, we used images of several regions in the Northeastern US, primarily the NY Thruway, with a total of 2050 truck examples. For validation, we worked with images from 3 sections of the Thruway, some partially covered by fog, that contain 216 truck examples (81 on road). For the road filter, we used a shapefile of the Thruway provided by the State of New York [26].

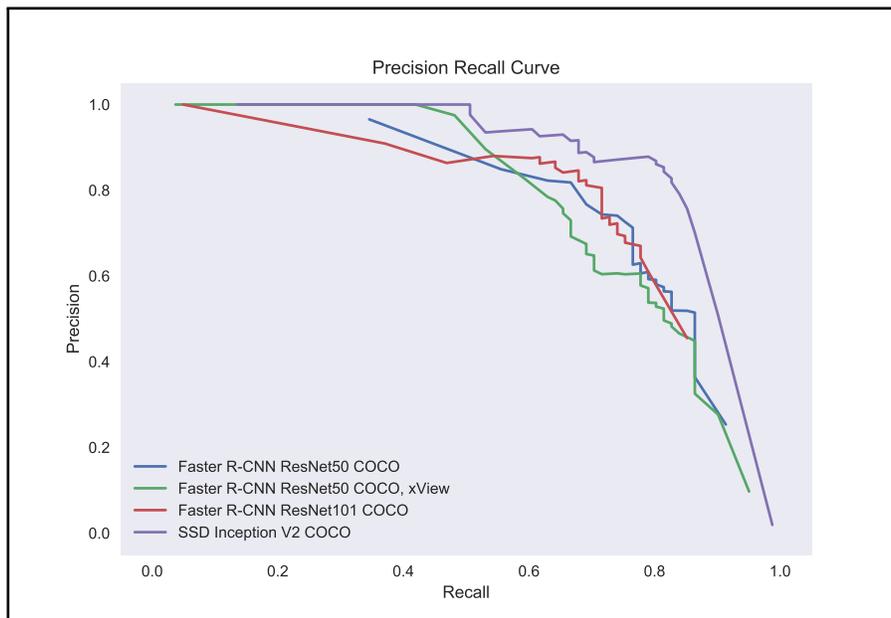
The toll data for the Thruway contain the entrance and exit toll booth for the vehicle, hour and date of entrance, the vehicle class and the number of vehicles with these specifics. We use data with a time resolution of 1 hr for the year 2016 [24] for training and 2017 for testing [25] (Appendix A.3).

Detection model We ran four experiments with Faster R-CNN Resnet 50 and 101 and SSD Inception V2, where we (1) pre-trained with COCO image data and (2) pre-trained with COCO and xView (Appendix B.1). From Table 1 and Fig. 1, we can see that the models perform better if the experiment is constrained to the road. The full image can contain more difficult examples, for example clustered trucks on parking lots or less typical trucks in junk yards or construction sites. A model that was not pre-trained proved difficult to train, but when pre-trained on COCO data it showed good results. Additional pre-training on xView marginally improved the precision and recall of the smaller net for predictions on the full image, as it has learned more context with images that do not contain trucks. The additional training data and

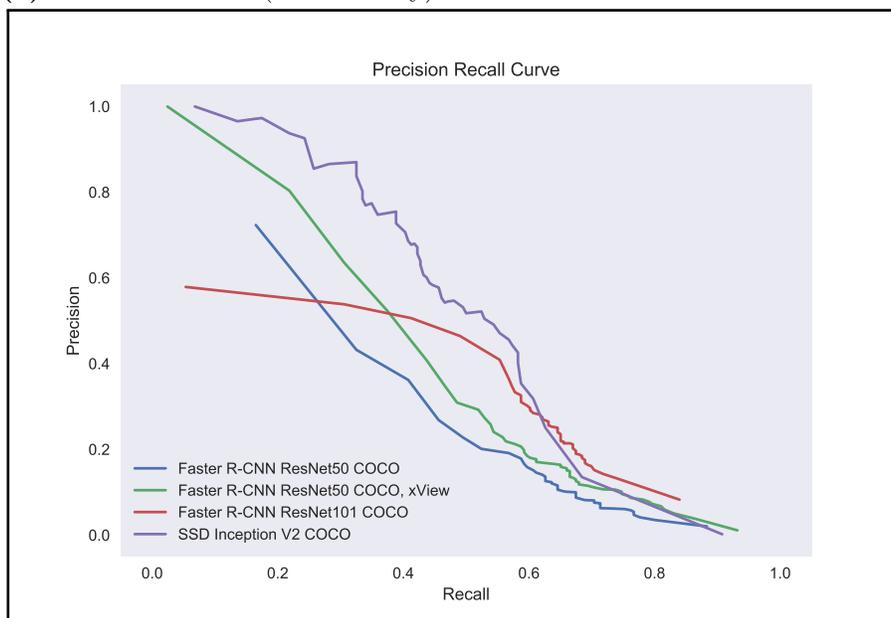
Table 1: Performance for optimal count prediction probability.

Faster R-CNN	Pre-train	Fine-tune	Av. Count Err.	p_{pred}	Av. Prec.	Av. Rec.
With road filter						
ResNet50	COCO	~ 1000 trucks	0.111	0.925	0.741	0.750
ResNet50	COCO & xView	~ 1000 trucks	0.247	0.880	0.654	0.746
ResNet101	COCO	~ 2000 trucks	0.160	0.600	0.716	0.806
SSD Inception V2	COCO	~ 2000 trucks	0.123	0.175	0.802	0.861
On entire image						
ResNet50	COCO	~ 1000 trucks	0.510	0.995	0.325	0.432
ResNet50	COCO & xView	~ 1000 trucks	0.403	0.980	0.374	0.524
ResNet101	COCO	~ 2000 trucks	0.466	0.990	0.413	0.506
SSD Inception V2	COCO	~ 2000 trucks	0.262	0.200	0.544	0.491

layers of the third model appeared not to have improved the performance on road, but on the entire image. SSD Inception V2 shows the best results. We optimized the threshold for the prediction probability to achieve minimal count error on the validation data constrained to the road (Table 1 and Appendix B.1). SSD Inception V2 achieved a marginally higher minimal count error than Faster R-CNN Resnet 50, but based on the clearly higher precision, recall and speed, we chose to use SSD Inception V2 with prediction probability $p_{pred} = 0.175$ to test the model.



(a) With road filter (on Thruway)



(b) On entire image

Figure 1: Precision-recall curves for validation images. All of the models perform better when used for on-road predictions, as those often contain less difficult examples.

Table 2: Test results using SSD Inception V2.

Test	Exits Section	ann. count	pred. count	$AADTT_{true}$	$AADTT_{pred}$	ϵ_{test}
1	Exit 25A to 26	10	8	2290	2245	-2%
2	Exit 44 to 45	36	12	1710	2788	63%
3	Exit 35 to 36	23	15	2149	4111	91%

Freight monitoring model The data are seasonal with no trend, which is why we use linear regression models with time fixed effects to estimate the factors of time-of-day, day-of-week, and monthly variation, and the variance of the random component of the traffic counts. Those models are informed by the recommended practices of the FHWA [33, 17]. We found that the linear model, which yields the lowest cross-validated error, includes one interaction term:

$$\text{hourly count} \sim \text{month} + \text{day of week} + \text{hour} + \text{day of week} \times \text{hour}.$$

We train on 2016 data from three sections of the Thruway to estimate the time-varying factors for the test cases in 2017. We normalize to account for an equal number of weekdays between years. We find that individual hours in validation data can vary substantially from the predicted value, depending on which exits are used for training (Appendix B.2). We expect the error to be even larger if factors are applied to other roads, and countries.

Test results We estimate the AADTT for 2017 for three test sections using the number of trucks counted with the detection model and the predicted seasonal variation factor for the time stamps of the images. The model achieves a quite small error for the first test case, but is overpredicting in the two other cases (Table 2). This discrepancy of model performance is expected, given that a snapshot image corresponds to a single, very short counting time, and is sensitive to traffic fluctuations. For example, the very high number of annotated trucks in Test Case 2 is partially due to trucks parked on the shoulder of the highway that belonged to a building ground. We see that the detection model, however, reports counts that are systematically lower than the true count.

5 Generalizing the model to another country

We work with a highway in Brazil (BR-116) to test how well the model transfers to another country. This test case is suitable, as it is located in a developing country

Table 3: Test results for Brazil count station km 109 with $AADTT_{true} = 15798$.

Test	Image date	Day of week	Time	Count	$AADTT_{pred}$	ϵ_{test}
With Thruway factors and detection model counts						
1	03/12/2018	Monday	10:26	10	3042	-81%
2	08/11/2016	Thursday	10:28	12	3651	-77%
3	06/16/2016	Thursday	10:21	14	4259	-73%
With Thruway factors and true image counts						
1	03/12/2018	Monday	10:26	32	9735	-38%
2	08/11/2016	Thursday	10:28	24	7301	-54%
3	06/16/2016	Thursday	10:21	31	9431	-40%

but there are sufficient data available to analyze how well the model and each of its components generalize.

Data For the detection model, we use image data from DigitalGlobe, Inc., for three different time stamps at the same location. For the monitoring model, we work with traffic data from continuous and short term counters available through the Brazilian agency Departamento Nacional de Infraestrutura de Transportes (DNIT) [5]. Our first test case is a section of the BR-116 between two exits, where a counter was located at km 109. We also retrieve geospatial data of roads in Brazil from DNIT [6]. These data are centered in one of the lanes, which is why we needed to expand the range used in the road filter to 30m to ensure that both lanes pass the filter. This could result in some errors if trucks are parked close to the road.

Experiment We begin by analyzing the performance of the whole model on the new image data, using the trained detection model and parameter settings as well as the traffic variation factors from the NY Thruway. We estimate the AADTT both on the count obtained through the detection model and, in order to assess the monitoring model separately, we also estimate it based on the true number of trucks seen in the images. For the ground truth we use an AADTT based on a count of 21 days in 2017 as reported by DNIT [5].

Test results From the results in Table 3, we see that the detection model under-predicts the number of trucks in the new images. The time variation factors, however, seem to transfer better. If the true number of trucks visible in the images is used, the estimated AADTT is closer to the true value.

Discussion These results clearly indicate that additional fine-tuning of the detection model on images of the new location is necessary. The poor performance of the detection model on data in Brazil is most likely due to the occurrence of new truck types that are specific to Brazil and were not contained in the training dataset (Fig. 2). To understand how well the monitoring model generalizes, we will compare the factors obtained through ground truth traffic counts on the BR-116 with those from the NY Thruway. We also plan to produce AADTT estimates with factors from local data. Besides testing each of the three images separately, we will also report the results averaged over all three.



Figure 2: While there is considerable variability in the training data from North-eastern US, which also include winter scenes, the detection model does not generalize well. These images show that trucks seem to look different in Brazil compared to the US. Green boxes indicate annotated examples.

6 Discussion and conclusion

We find that we can use machine learning to count trucks with reasonable accuracy. If provided with typical traffic patterns, a snapshot image can yield predictions of average daily traffic volumes that are acceptable, given the data limitations. Those results could be improved in particular by using multiple satellite images taken of the same section at different times, where available. The method currently still requires

access to images, knowledge, and computing resources that might be difficult for some countries, but this could change in the near future.

While these initial results are promising, both the detection model and the monitoring model are likely to entail much higher uncertainty when transferred to developing countries. We will explore in detail, how the model transfers between the two countries, in order to understand how it would perform in developing countries with poor data. From an initial test on a highway in Brazil, we found that in particular the detection model does not generalize well. Distinct truck types and road surfaces likely impact the prediction accuracy of the detection model, and additional training seems necessary. Also, seasonality patterns of traffic can be expected to differ significantly from Northeastern US. We will add a more detailed analysis of the monitoring model, where we will compare the traffic factors from the Thruway with ground truth data from Brazil. We will also perform a sensitivity analysis of the assumptions.

Acknowledgments This work was supported by the Center for Climate and Energy Decision Making through a cooperative agreement between the National Science Foundation and Carnegie Mellon University (SES-0949710). We thank DigitalGlobe, Inc., for providing satellite image data. We are also grateful for computing resources that were made available through Google Cloud Platform and Microsoft AI for Earth. We are also grateful for assistance from Francisco Ralston Fonseca with Brazilian data.

References

- [1] Brent Bowen, Karisa Vlasek, Cindy Webb, et al. An assessment of remote sensing applications in transportation. In *45th Annual Transportation Research Forum, Evanston, Illinois, March 21-23, 2004*, number 208247. Transportation Research Forum, 2004.
- [2] Raj Bridgelall, James B Rafert PI, and Denver D Tolliver. Remote sensing of multi-modal transportation systems. Technical report, Mountain Plains Consortium, 2016.
- [3] Xueyun Chen, Shiming Xiang, Cheng-Lin Liu, and Chun-Hong Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11(10):1797–1801, 2014.
- [4] Zhipeng Deng, Hao Sun, Shilin Zhou, Juanping Zhao, and Huanxin Zou. Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2017.

-
- [5] Departamento Nacional de Infraestrutura de Transportes . Plano Nacional de Contagem de Tráfego - Dados de Tráfego. Available at <http://servicos.dnit.gov.br/dadospnct/DadosTrafego>.
- [6] Departamento Nacional de Infraestrutura de Transportes. Shapefile of roads in Brazil. Available at <http://www.dnit.gov.br/mapas-multimodais/shapefiles>.
- [7] Line Eikvil, Lars Aurdal, and Hans Koren. Classification-based vehicle detection in high-resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(1):65–72, 2009.
- [8] A Gerhardinger, D Ehrlich, and M Pesaresi. Vehicles detection from very high resolution satellite imagery. *International Archives of Photogrammetry and Remote Sensing*, 36(Part 3):W24, 2005.
- [9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] Juan C Herrera, Daniel B Work, Ryan Herring, Xuegang Jeff Ban, Quinn Jacobson, and Alexandre M Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568–583, 2010.
- [12] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, 2017.
- [13] Independent Expert Advisory Group on a Data Revolution for Sustainable Development. A world that counts: mobilizing the data revolution for sustainable development. *United Nations Publication*, 2014.
- [14] Qiling Jiang, Liujuan Cao, Ming Cheng, Cheng Wang, and Jonathan Li. Deep neural networks-based vehicle detection in satellite images. In *Bioelectronics and Bioinformatics (ISBB), 2015 International Symposium on*, pages 184–187. IEEE, 2015.
- [15] Lynn H Kaack, Parth Vaishnav, M Granger Morgan, Inês L Azevedo, and Srijana Rai. Decarbonizing intraregional freight systems with a focus on modal shift. *Environmental Research Letters*, 13(8):083001, 2018.

- [16] Konstantinos Kanistras, Goncalo Martins, Matthew J Rutherford, and Kimon P Valavanis. Survey of unmanned aerial vehicles (uavs) for traffic monitoring. In *Handbook of unmanned aerial vehicles*, pages 2643–2666. Springer, 2015.
- [17] Robert Krile, Fred Todt, and Jeremy Schroeder. Assessing Roadway Traffic Count Duration and Frequency Impacts on Annual Average Daily Traffic Estimation. Technical Report FHWA-PL-16-012, Federal Highway Administration, Washington, D.C., United States, 2016.
- [18] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018.
- [19] Siri Øyen Larsen, Hans Koren, and Rune Solberg. Traffic monitoring using very high resolution satellite imagery. *Photogrammetric Engineering & Remote Sensing*, 75(7):859–869, 2009.
- [20] Kang Liu and Gellert Mattyus. Fast multiclass vehicle detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, 12(9):1938–1942, 2015.
- [21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [22] Mark Marsden, Kevin McGuinness, Suzanne Little, Ciara E. Keogh, and Noel E. O’Connor. People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. 11 2017.
- [23] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision*, pages 785–800. Springer, 2016.
- [24] New York State Thruway Authority. NYS Thruway Origin and Destination Points for All Vehicles - 1 Hour Intervals: 2016, February 2018. Available at <https://catalog.data.gov/dataset/nys-thruway-origin-and-destination-points-for-all-vehicles-1-hour-intervals-2016>.
- [25] New York State Thruway Authority. NYS Thruway Origin and Destination Points for All Vehicles - 1 Hour Intervals: 2017, February 2018. Available at <https://catalog.data.gov/dataset/nys-thruway-origin-and-destination-points-for-all-vehicles-1-hour-intervals-2017>.
- [26] New York State Thruway Authority. *NYSTA Route System*. New York State Thruway Authority, GIS, Albany, New York, 2018. Available at <https://gis.ny.gov/gisdata/inventories/details.cfm?DSID=440>.

-
- [27] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016.
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [29] Lars Wilko Sommer, Tobias Schuchert, and Jürgen Beyerer. Fast deep vehicle detection in aerial images. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 311–319. IEEE, 2017.
- [30] Jacob Teter, Pierpaolo Cazzola, and Timur Gül. *The Future of Trucks*. International Energy Agency, 2017.
- [31] Thruway Authority. Interchange/Exit Listings. Available at <https://www.thruway.ny.gov/travelers/interchanges/index.html>.
- [32] Tzutalin. Labeling, 2015. Available at <https://github.com/tzutalin/labelImg>.
- [33] US Federal Highway Administration. Traffic Monitoring Guide, 2016.

A Data preparation

A.1 Satellite images

We use satellite images from Digital Globe, which are taken frequently for many locations and by a number of different satellites. We only work with 3-channel RGB images of the satellite "World View 3" (VW03), as it has the highest resolution. We found that there are difficulties identifying for example black cars in images of the other satellites. This also makes our dataset comparable to xView, which also uses VW03 imagery. The images are not cloud-free but we attempt to select images with nearly no cloud-cover in relevant areas.

For training, we use images of several regions in the Northeastern United States. These include images of the Thruway but also other highways. We also annotate the regions around the highways such as parking lots and logistics centers.

A.2 Annotations

We use the python-based annotation software "LabelImg" [32] to label the more than 2000 truck examples. We mark each truck with a bounding box and a class label "Truck." Below we describe in detail, which types of vehicles we include as trucks. For cloudy images, we tried to label also those trucks that are hardly visible through the cloud or in the shade of the cloud.

For the training data, we labeled large 3000×3000 pixel images, from which we created 300×300 pixel chips, where we only retained chips with truck examples. Note that this procedure reduces the number of truck examples somewhat with respect to the large images, as bounding boxes are cut and those examples are lost. We chose truck examples conservatively and prioritized accuracy of labeled examples over labeling as many as possible. This means, when in doubt, we chose not to label truck, unless there are very obvious or interesting truck examples in the immediate vicinity. We have sometimes omitted parking areas and junk yards, where vehicles are very close together such that there is no pavement visible. These examples might be less useful for learning on highways but might aid for situations with dense traffic.

In contrast, for test images it was important to label all likely trucks as such, as the whole image is evaluated with all examples it contains. This includes trucks that are partially obstructed through trees, bridges etc.

We only labeled semi trucks with a trailer, and medium-duty trucks. This also includes car carrier trailers, flatbed trucks or oversized transports such as windturbine blades. We do not annotate pickup trucks, even if they pull a trailer, and omit vans, buses, caravans, and RVs. We also do not label tractors or trailers separately, only the combination. Fig. 3a shows an example of easily identifiable trucks of different sizes. The examples in Fig. 3b and 3c are partially obstructed by clouds or trees but they can still be identified as trucks. Buses and RVs can easily be confused with a truck. The example in Fig. 3d shows a bus

(or a long RV) and something that is likely an RV that is pulling a car, both of which could be taken for a truck. Fig. 3e and 3f have more of those examples, including a number of yellow school buses. We also include trucks with special trailers such as flatbeds (Fig. 3g), or oversized load (Fig. 3h). We also label trucks with empty trailers, as sometimes load cannot be distinguished from the empty trailer. Smaller trucks and their similarity with vans are particularly difficult (Fig. 3i). We consider everything that has a box that is elevated from the driver’s cabin as a truck but errors cannot be excluded. Also, there are many examples of small parked trailers that have a white attachment, which could also be a small drivers cabin (Fig. 3k). For trucks that are docked to a building, we excluded those where only the trailer is visible but included those, where the tractor is still attached (Fig. 3b).

In the images from Brazil, we found what seemed to be yellow and white busses, which appear in multiple locations Fig. 3l). After we have confirmed with Google Street View that such busses frequently travel the highway, we have not labeled these as trucks.

A.3 Thruway vehicle counts

The toll data for the New York Thruway [25, 24] contain the entrance and exit location for every vehicle and the time it has entered the Thruway as recorded in the toll collect system. We count all vehicles with 3 or more axles that are indicated as high vehicles as trucks. To determine when a vehicle has passed a location between entrance and exit, we need to make assumptions about the speed it has traveled and the distance between highway exits. We assume that every vehicle travels 65 mi/hr. The information about Thruway mileposts are available through the State of New York Thruway Authority [31]. We determine the counts by summing up all the vehicles that have entered a section between two highway exits. For example, if we are interested in analyzing the stretch of road between Exit 30 and 31, we determine the hourly counts by summing up the number of vehicles that pass Exit 30 in one direction and pass Exit 31 in the other direction within that hour. We are not only counting the vehicles that pass the particular exit but also the vehicles that enter the Thruway on that exit. The ones that leave at the exit before they enter the section are not counted. See Fig. 4 for an illustration.

B Model details

B.1 Detection model

Model specifications

1. Faster R-CNN Resnet 50, convolutional layers pre-trained on COCO image data

2. Faster R-CNN Resnet 50, convolutional layers pre-trained on COCO image data and subsequently on xView
3. Faster R-CNN Resnet 101, convolutional layers pre-trained on COCO image data.
4. Single Shot Detector with Inception V2, pre-trained on COCO image data.

Validation We count a truck as detected, if its bounding box has an intersection over union (IoU) with the ground truth of at least 0.3. We compute the average precision and recall over all validation images together, and do not average the performance over each image separately.

We determine the optimal threshold for the prediction probability to minimize the error of total counts in an image (true positives and false positives). We compute the mean absolute count error over all validation images as the weighted sum of the relative absolute count error of each of N images

$$\begin{aligned}
 \epsilon_{Count} &= \sum_{i=1}^N w^{(i)} \epsilon_{Count}^{(i)} \\
 &= \sum_{i=1}^N \frac{f_{true}^{(i)}}{\sum_{i=1}^N f_{true}^{(i)}} \cdot \frac{|f_{pred}^{(i)} - f_{true}^{(i)}|}{f_{true}^{(i)}} \\
 &= \frac{\sum_{i=1}^N |f_{pred}^{(i)} - f_{true}^{(i)}|}{\sum_{i=1}^N f_{true}^{(i)}}, \tag{1}
 \end{aligned}$$

where f is the number of trucks counted in image i . While the mean absolute count error is often used [22], we chose to use a weighted sum to account for the relative importance of images with more validation examples, as we have large images with a diverse number of truck examples. The measure does not have guarantees to be convex and might not have a unique minimum due to the discrete nature of counting. We compare this for various probability thresholds (Fig. 5). We see that Faster R-CNN Resnet 50, that is only pre-trained on COCO, achieves lowest average absolute count error for a prediction probability of ~ 0.925 . It can achieve simultaneously the largest precision and recall (Table 1 and Fig. 1).

B.2 Monitoring model

The uncertainty of estimating the AADTT from a 1-hr count is composite of the inter-year variation and the prediction error from the fitted model. Typically, there will be few ground-based counting stations on a highway, but the toll data provides counts for every section. If we use all sections on the Thruway for training, or a section that is very close to the test section, we will overstate our ability to estimate the time factors. We therefore train the

model on three sections (Exits 19-20, 30-31, and 41-42). Those exits are located relatively far away from each other and from the test sections Exit 25A-26 and Exit 44-45. We train on data from 2016, and test on 2017. Data processing is costly for the large toll datasets. For validation, we have trained on the first two sections together, and each separately, and tested how well they predict ground truth data for Exit 41-42 in 2017. We have also trained only on Exit 41-42 in 2016. As expected, this latter trivial case yields the most accurate time-varying factors, where we find a median out-of-sample relative prediction error of -1% and 5th and 95th percentile of -19% and 36% over all hours of the year. However, we are interested in more realistic cases, where we do not have perfect data for the validation region. Training on Exits 19-20 (2016) to predict factors for Exits 41-42 in 2017 results in a mean error of 0.09 [-0.18, 0.83]. Training on Exits 30-31 (2016) yields -0.32 [-0.47, 0.00]. Training on both sections gives the best result of -0.10 [-0.30, 0.35]. We conclude that averaging over more sections makes predictions more robust and therefore use the counts for all three exits in 2016 to train.

B.3 Testing

Test sections:

1. Exit 25A to 26 on the NY Thruway in an image from 2017-08-04/16:02
2. Exit 44 to 45 on the NY Thruway in an image from 2017-09-12/16:27.

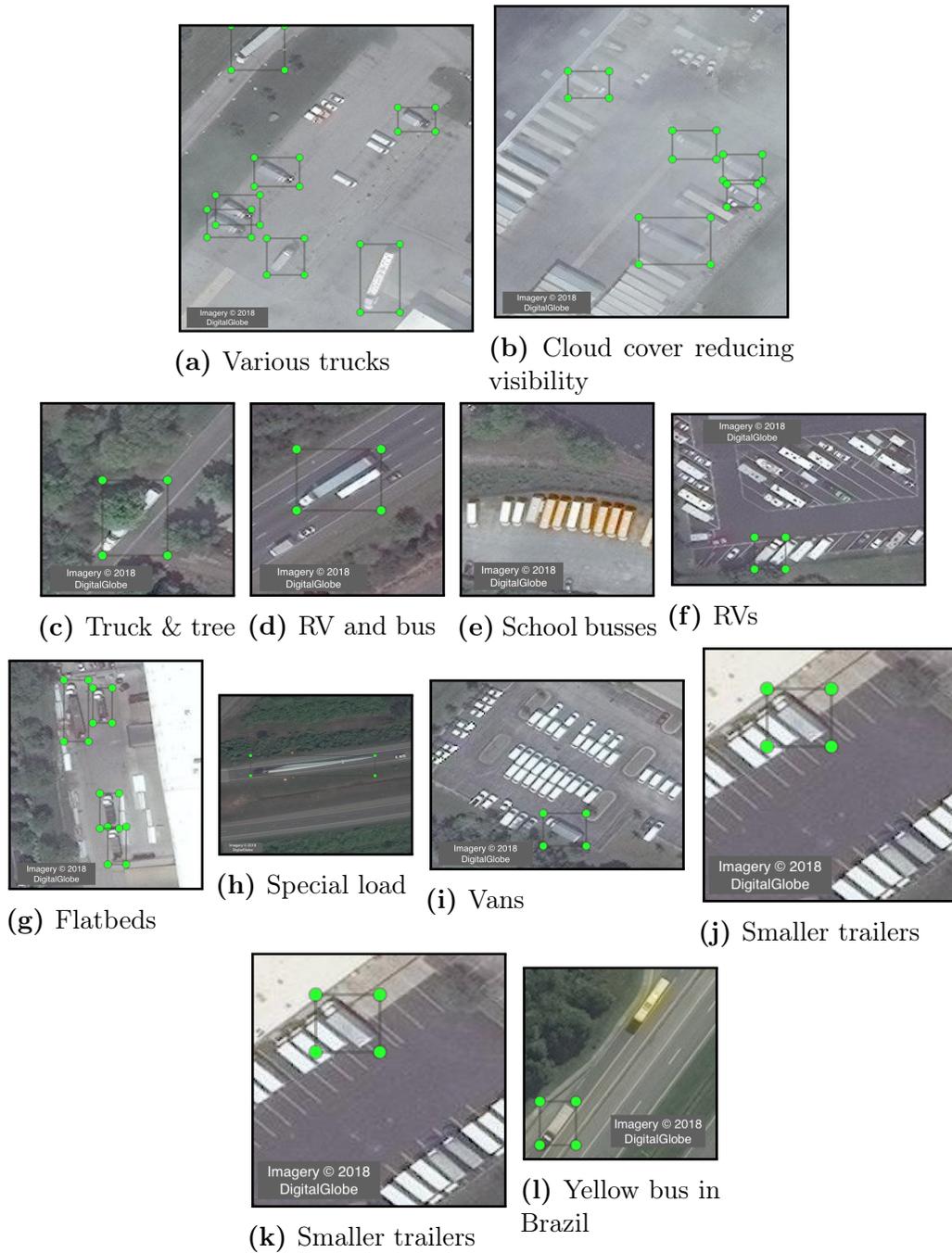


Figure 3: Image chips that illustrate what is labeled as a "Truck" with a bounding box. Imagery © 2018 DigitalGlobe, Inc.

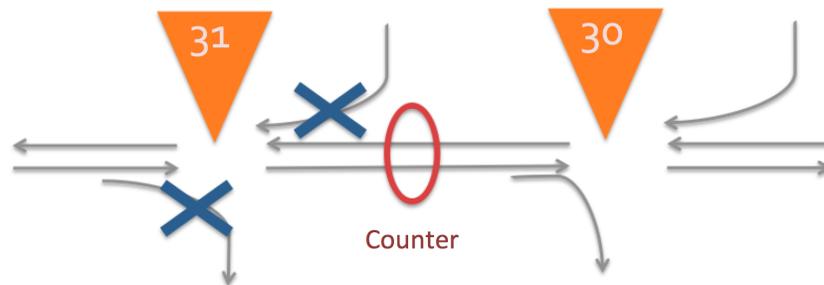
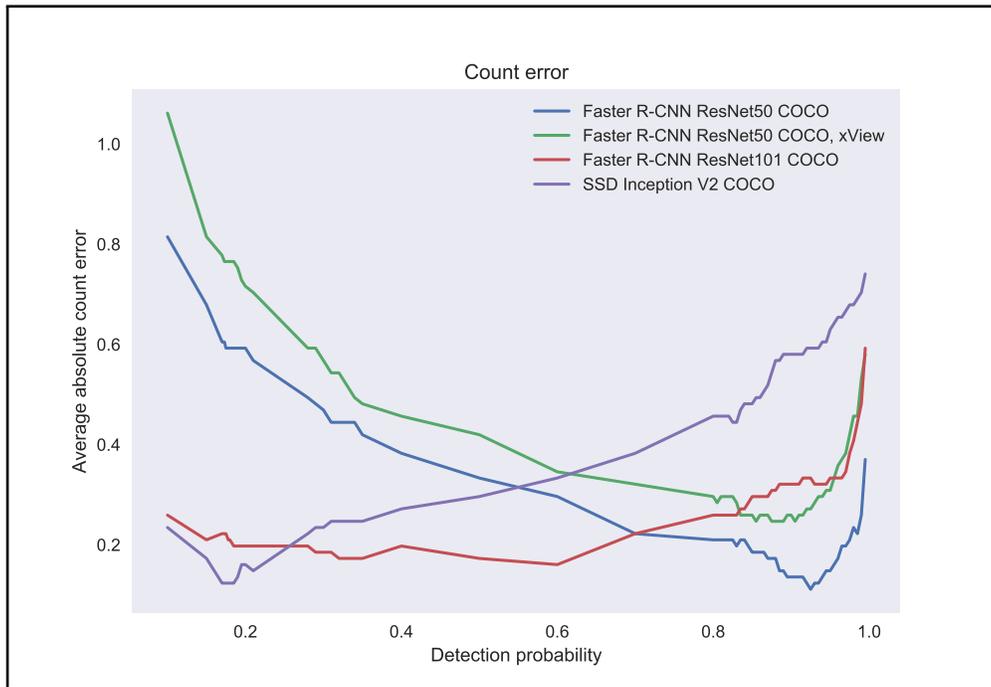
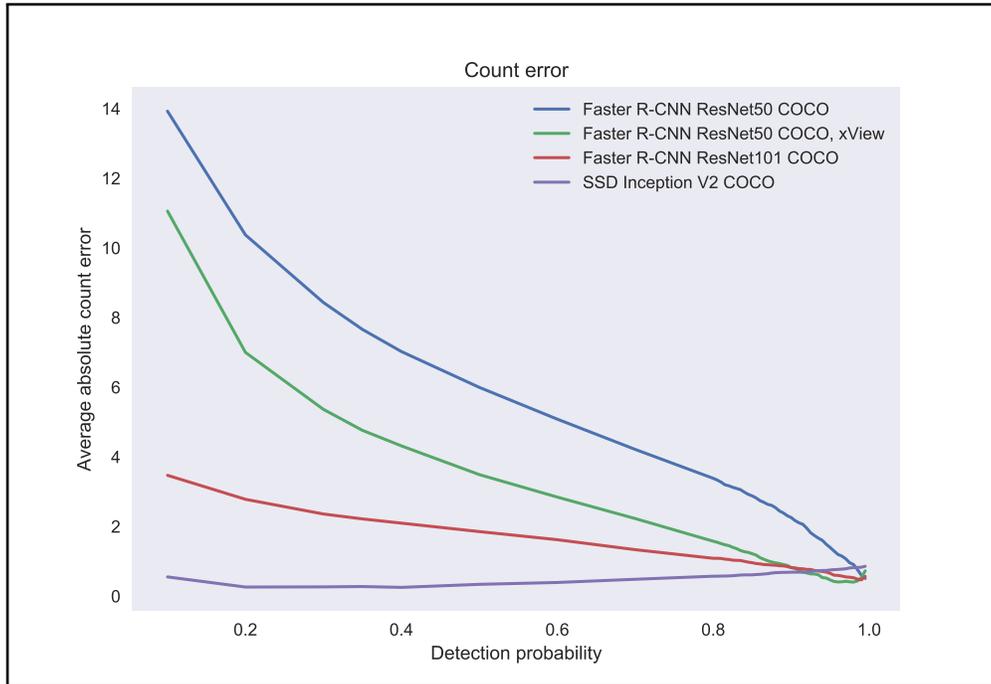


Figure 4: Schematic illustration of how we compute the traffic flow for the monitoring model from toll data. The orange cones indicate highway exits, and the grey lines are road sections. We do not consider those that are crossed out.



(a) With road filter



(b) On entire image

Figure 5: Absolute error of total truck counts, which also includes false positives, over detection probability. We see that the Resnet 50 model that is only pre-trained on COCO achieves lower count error.