

Detection of Vehicles from satellite images Using Deep Neural Networks

Golej Peter^{1,#}, Horák Jiří¹, Kukuliach Pavel¹, Orliková Lucie¹

¹ Department of Geoinformatics, VSB - Technical University of Ostrava, Ostrava, 708 00, Czech Republic, peter.golej@vsb.cz, jiri.horak@vsb.cz, pavel.kukuliach@vsb.cz, lucie.orlikova@vsb.cz

Abstract

Vehicle detection in urban environments is an important task in remote sensing, intelligent transportation systems, and spatial management. This paper investigates the use of deep convolutional neural networks for the automatic detection of vehicles from high-resolution panchromatic satellite imagery within the built-up area of the city of Ostrava. The proposed approach was implemented using ArcGIS Online and ArcGIS Pro tools, with model training performed in a GPU-enabled Jupyter Notebook environment. Vehicle detection was carried out using the Faster R-CNN model, with a focus on optimizing the backbone architecture, batch size, number of training epochs, learning rate, and data augmentation strategy. In addition to detecting all vehicles as a single class, the model was also evaluated for vehicle classification based on lightness and vehicle type. The results demonstrate that an appropriate combination of training parameters, the use of the ResNet152 backbone, and the application of spatial masks substantially reduce false positive detections and improve overall detection performance, achieving classification accuracy of approximately 70%.

Keywords: Object detection, satellite, vehicles, Faster R-CNN, deep learning

Introduction

Automatic vehicle detection from aerial and satellite imagery is an important research topic with applications in traffic planning, parking monitoring, mobility analysis, and security-related tasks. With the rapid development of deep learning methods—particularly convolutional neural networks (CNNs)—the accuracy and robustness of object detection algorithms have increased significantly [1], [2].

However, urban environments place high demands on detection algorithms. Vehicles appear at varying scales, orientations, and illumination conditions, are often partially occluded, and may be located in shadows cast by buildings or vegetation. Moreover, urban imagery contains objects with similar shapes and spectral characteristics, such as rooftops or technical infrastructure elements, which can lead to false positive detections.

While most existing vehicle detection approaches rely on standard RGB or multispectral satellite imagery, the potential of high-resolution panchromatic data remains underexplored. Compared to pan-sharpened RGB imagery, panchromatic data preserve the original spatial detail and object boundaries without introducing pan-sharpening artifacts, which may be advantageous for detecting small objects such as vehicles. The absence of color information, however, presents unique challenges for deep learning models typically designed for RGB inputs, particularly when distinguishing vehicles from spectrally similar urban features.

The aim of this paper is to analyze the applicability of the Faster R-CNN model within the ArcGIS Online environment for vehicle detection from panchromatic satellite imagery of the city of Ostrava. Specifically, we investigate whether deep learning models such as Faster R-CNN can effectively detect and classify vehicles using single-band panchromatic data converted into a three-band format, and we evaluate the impact of different training settings on detection performance. Special attention is given to data augmentation techniques and the use of building masks as a means of reducing false positive detections.

Data and Method

Satellite Image Data

The input data for vehicle detection consisted of very high-resolution imagery acquired by the WorldView-4 satellite on 13 August 2018. The area of interest was the city of Ostrava, a city of approximately 300,000 inhabitants located in the NE part of the Czech Republic. Panchromatic images with a spatial resolution of 30 cm covering the urban area of Ostrava were used.

The dataset captured a wide range of urban structures, including residential and industrial buildings, parking areas, road networks, and vegetation. As is typical for dense urban environments, the imagery exhibited several challenging characteristics, such as strong shadows cast by buildings, varying object orientations, partial occlusions, and high surface variability.

Prior to model training, the input images were processed into a three-band format analogous to RGB imagery to ensure compatibility with convolutional neural networks implemented in the ArcGIS Online environment [3]. Subsequently, the dataset was divided into training, validation, and test subsets.

Reference Data

Reference data were created through manual annotation of vehicles visible in the imagery. Vehicles were classified according to both radiometric appearance and vehicle type. Based on lightness, three categories were defined: light vehicles (307 samples), dark vehicles (308 samples), and vehicles located in shadow (38 samples). Based on vehicle type, the annotated objects were categorized as cars (630 samples), vans (20 samples), and buses (3 samples). In this study, the term *car* refers to passenger vehicles, while *vans* primarily represent small lorries. Owing to the low frequency of certain categories, particularly buses and vans, the class distribution was highly imbalanced, which influenced the detection performance of the model. Data augmentation was therefore used to partially mitigate the effects of class imbalance.

Model Architecture and Training Setup

Vehicle detection was performed using the Faster R-CNN model, which belongs to the class of two-stage object detection architectures. In the first stage, candidate object regions are generated by a Region Proposal Network (RPN), while in the second stage these regions are classified and refined through bounding box regression [4].

A total of 653 annotated vehicles were used for model development and split into 522 samples for training (80%) and 131 samples for validation (20%) [5]. An additional set of 528 independently annotated vehicles was reserved for final model testing.

Batch size was treated as a critical parameter influencing both computational efficiency and training performance. Batch sizes ranging from 4 to 64 were evaluated. The experiments indicated that a smaller batch size (batch size = 4) provided the most favorable trade-off between computational efficiency and detection accuracy.

Several backbone architectures were tested, including ResNet50, ResNet101, and ResNet152. The backbone network is responsible for feature extraction and has a substantial impact on overall detection performance. Based on the experimental results (Table 1), ResNet152 was identified as the most suitable backbone architecture for the given task.

Preliminary tests using the YOLOv3 architecture supported by ArcGIS Online achieved lower detection performance than Faster R-CNN on the evaluated panchromatic imagery; consequently, subsequent experiments were conducted using Faster R-CNN only.

The optimal learning rate was determined using a learning rate finder, which analyzes the behavior of the loss function across a range of learning rate values [6].

The number of training epochs was set to 150, 300, and 450 in different experiments (Table 1). Although the highest Average Precision was achieved after 450 epochs, the model trained for 300 epochs was selected as the optimal compromise, as longer training times resulted in significantly increased computational costs with only marginal performance gains.

Table 1. Settings for training the Faster R-CNN model.

Input parameters				Optimalization	Results
ID	Backbone	Batch size	Epochs	Learning rate	AP Score (%)
1.	Resnet152	16	150	0,000021	65.6
2.	Resnet152	32	150	0,000012	55.7
3.	Resnet152	64	150	0,000014	54.5
4.	Resnet152	8	150	0,000017	65.6
5.	Resnet152	4	150	0,000021	68.5
6.	Resnet152	4	300	0,000017	71.5
7.	Resnet152	4	450	0,000003	73.4
8.	Resnet101	16	150	0,000025	66.1
9.	Resnet101	4	150	0,000052	70.9
10.	Resnet50	4	150	0,001	58.2

To increase the diversity of training samples, data augmentation techniques [7], [8], [9] were applied using the `get_transforms` function. The applied transformations included horizontal and vertical flipping, image rotation within $\pm 20^\circ$, random zooming of up to 20%, and warp transformations with a magnitude of 0.2. These transformations were applied randomly during each training iteration, meaning that the same image could be processed differently in each batch. This approach effectively increased the size of the training dataset and improved the model's generalization capability.

The behavior of the training and validation loss functions provides important insight into the learning process. When data augmentation was applied, the training and validation loss curves remained closer to each other and exhibited less divergence, indicating improved generalization and a reduced risk of overfitting (Fig. 1b).

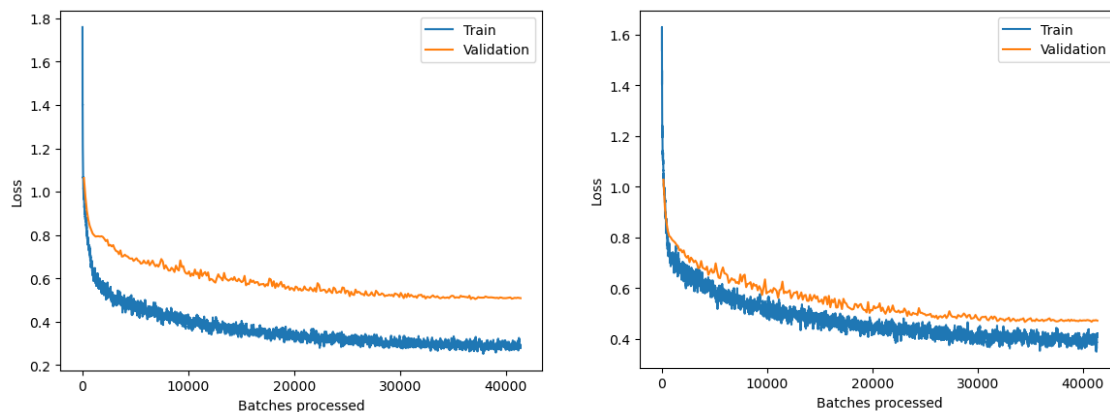


Fig. 1. Training and validation loss functions for the sixth experimental setting listed in Table 1: (a) without data augmentation; (b) with data augmentation applied.

Model Testing and Accuracy Assessment

The trained Faster R-CNN model was tested in the ArcGIS Pro environment using the *Detect Objects Using Deep Learning* tool. Testing was performed on data that were not used during either training or validation. Only detections with a confidence score greater than 50% were considered valid detections.

Three model variants were evaluated: (i) detection of all vehicles as a single class, (ii) detection based on vehicle lightness, and (iii) detection based on vehicle type. Each variant was tested both with and without the application of a building mask. The use of the building mask during evaluation helped to eliminate false positive detections in areas where vehicles are not expected to occur, such as building rooftops.

Accuracy assessment was carried out using the *Compute Accuracy for Object Detection* tool, which provides Precision, Recall, and F1-score metrics [10] for different Intersection over Union (IoU) thresholds. The IoU metric represents the degree of overlap between the bounding box of a predicted object and the corresponding bounding box of the ground reference data.

Results

Detection of All Vehicles as a Single Class

The results of vehicle detection without distinguishing vehicle type or lightness indicate that the application of spatial masks significantly reduces the number of false positive detections. In this context, the *evaluation mask* refers to a mask applied only during post-training accuracy assessment, whereas the *training mask* denotes a mask used during model training to exclude areas such as building rooftops or water bodies.

At lower IoU thresholds, the model achieved relatively high Recall values (71.6%). Precision and F1-score improved notably when evaluation masks were applied. A substantial decrease in accuracy was observed at the stricter IoU threshold of 0.5. Since vehicles occupy only approximately 10–20 pixels in the imagery, even small localization errors can significantly reduce the overlap between predicted and reference bounding boxes.

Table 2. Evaluation of the Faster R-CNN model for vehicle detection.

Model	IoU	Evaluation mask	TP	FP	FN	Precision (%)	Recall (%)	F1 score (%)
without primary mask	0.1	No	378	300	150	55.8	71.6	62.7
	0.2	No	370	308	158	54.6	70.1	61.4
	0.5	No	79	599	449	11.7	15.0	13.1
without primary mask	0.1	Yes	378	213	150	64.0	71.6	67.6
	0.2	Yes	370	221	158	62.6	70.1	66.1
	0.5	Yes	79	512	449	13.4	15.0	14.1
with primary mask	0.1	No	380	196	148	66.0	72.0	68.8
	0.2	No	373	203	155	64.8	70.6	67.6
	0.5	No	78	498	450	13.5	14.8	14.1

Note: The evaluation mask indicates whether a mask was applied during the accuracy assessment of the model.

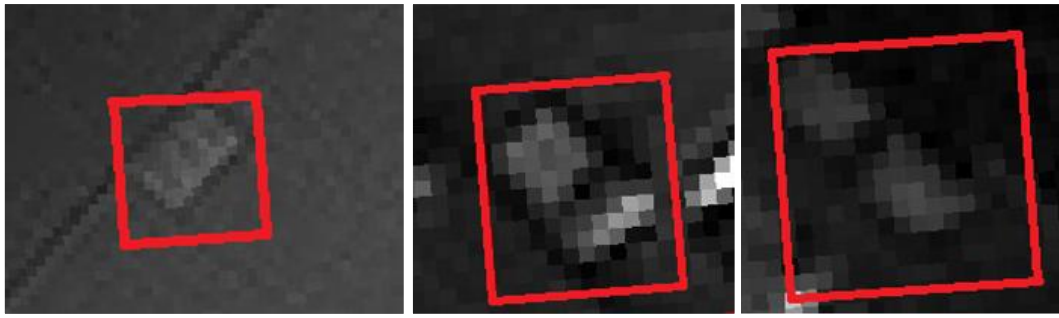


Fig. 2. Detection of false positives on building rooftops.

Most false positives correspond to rooftop structures and shadows with geometric characteristics similar to vehicles in panchromatic imagery.

Vehicle Detection Based on Lightness

The best detection results were achieved for dark vehicles, with F1-scores exceeding 70% at lower IoU thresholds, where a minimum overlap of 20% of the vehicle area was required. Light vehicles exhibited slightly lower performance. Vehicles located in shadow proved to be the most challenging class, with significantly lower F1-scores, mainly due to the limited number of training samples and the reduced contrast between vehicles and the surrounding background.

Table 3. Vehicle detection based on lightness with a training mask applied.

Type	IoU	TP	FP	FN	Precision (%)	Recall (%)	F1 score (%)
Dark	0.1	166	70	70	70.3	70.3	70.3
	0.2	166	70	70	70.3	70.3	70.3
	0.5	17	219	219	7.2	7.2	7.2
Light	0.1	181	81	71	69.1	71.8	70.4
	0.2	179	83	73	68.3	71.1	69.6
	0.5	29	233	233	11.1	11.5	11.3
in shadow	0.1	21	20	20	51.2	52.5	51.9
	0.2	21	20	19	51.2	52.5	51.9
	0.5	2	39	38	4.9	5.0	4.9

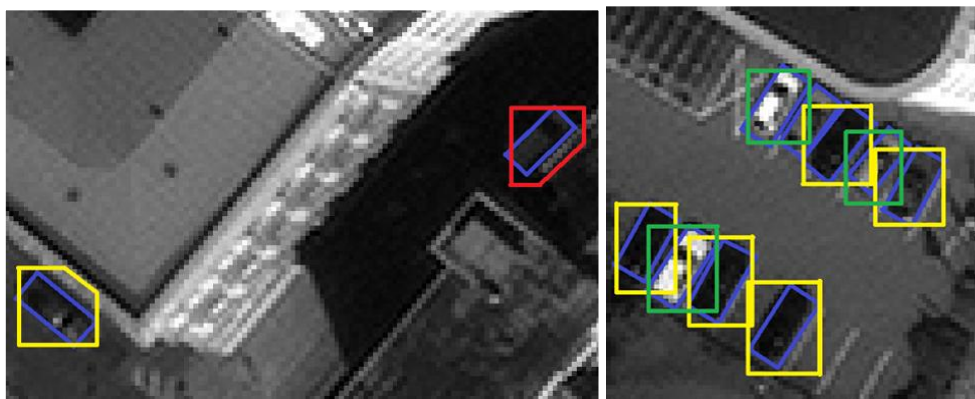


Fig. 3. Vehicle detection by lightness; blue – reference data; red – vehicles in shadow; yellow – dark vehicles; green – light vehicles.

Vehicle Detection by Vehicle Type

In vehicle type classification, passenger cars achieved F1-scores close to 70% at lower IoU thresholds when training masks were applied. Detection performance for vans was

considerably lower, with F1-scores below 40%, primarily due to the limited number of available training samples. Public transport vehicles were present in the test area but could not be reliably detected, highlighting the need for a substantially larger and more balanced training dataset. This result is largely explained by the extremely limited number of annotated bus samples available for training.

Table 4. Vehicle detection by vehicle type with a training mask applied.

Type	IoU	TP	FP	FN	Precision (%)	Recall (%)	F1 score (%)
Car	0.1	361	157	158	69.7	69.6	69.6
	0.2	352	166	167	68.0	67.8	67.9
	0.5	65	453	454	12.5	12.5	12.5
Van	0.1	3	6	4	33.3	42.9	37.5
	0.2	4	6	4	33.3	42.9	27.5
	0.5	0	9	7	0	0	0

Conclusion

This study demonstrates that the Faster R-CNN model implemented within the ArcGIS Online environment can successfully detect vehicles in complex urban environments using high-resolution panchromatic satellite imagery, achieving F1-scores of approximately 70% at lower IoU thresholds. The results highlight the importance of selecting an appropriate backbone architecture, optimizing training parameters, and applying data augmentation techniques. Spatial masks proved to be a key component in reducing false positive detections and improving overall detection accuracy.

The main limitation of the proposed approach lies in the imbalance of class representation within the training data, particularly for vehicles in shadow and less common vehicle types. Future research should therefore focus on expanding and balancing the training dataset, evaluating additional detection architectures, and integrating complementary spectral and spatial information to further enhance model robustness in complex urban environments.

References

- [1] Q. Tan, J. Ling, J. Hu, X. Qin, and J. Hu, "Vehicle Detection in High Resolution Satellite Remote Sensing Images Based on Deep Learning," *IEEE Access*, vol. 8, pp. 153394–153402, 2020, doi: 10.1109/ACCESS.2020.3017894.
- [2] S. Zambanini, A.-M. Loghin, N. Pfeifer, E. M. Soley, and R. Sablatnig, "Detection of Parking Cars in Stereo Satellite Images," *Remote Sens.*, vol. 12, no. 13, Art. no. 13, Jan. 2020, doi: 10.3390/rs12132170.
- [3] J. Kumar, T. Huan, X. Li, and Y. Yuan, "Panchromatic and Multispectral Remote Sensing Image Fusion Using Particle Swarm Optimization of Convolutional Neural Network for Effective Comparison of Bucolic and Farming Region," 2018.
- [4] M. Maity, S. Banerjee, and S. Sinha Chaudhuri, "Faster R-CNN and YOLO based Vehicle detection: A Survey," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, pp. 1442–1447. doi: 10.1109/ICCMC51019.2021.9418274.
- [5] S. Shahrabadi, T. Adão, E. Peres, R. Morais, L. G. Magalhães, and V. Alves, "Automatic Optimization of Deep Learning Training through Feature-Aware-Based Dataset Splitting," *Algorithms*, vol. 17, no. 3, p. 106, Mar. 2024, doi: 10.3390/a17030106.
- [6] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.
- [7] V. Lalitha and B. Latha, "A review on remote sensing imagery augmentation using deep learning," *Mater. Today Proc.*, vol. 62, pp. 4772–4778, Jan. 2022, doi: 10.1016/j.matpr.2022.03.341.
- [8] O. Adediji, P. Owoade, O. Ajayi, and O. Arowolo, "Image Augmentation for Satellite Images," Jul. 29, 2022, *arXiv*: arXiv:2207.14580. doi: 10.48550/arXiv.2207.14580.

- [9] M. a. A. Ghaffar, A. McKinstry, T. Maul, and T. T. Vu, “Data augmentation approaches for satellite image super-resolution,” *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. IV-2-W7, pp. 47–54, Sep. 2019, doi: 10.5194/isprs-annals-IV-2-W7-47-2019.
- [10] L. Derczynski, “Complementarity, F-score, and NLP Evaluation,” in *Proc. of 10th International Conference on Language Resources and Evaluation (LREC’16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 261–266. Accessed: Mar. 28, 2024. [Online]. Available: <https://aclanthology.org/L16-1040>