APPLICATION OF MACHINE LEARNING FOR OBJECT DETECTION IN OBLIQUE AERIAL IMAGES

P. Zachar [1,*], Z. Kurczyński [1], W. Ostrowski [1]

[1]Warsaw University of Technology, Faculty of Geodesy and Cartography, Department of Photogrammetry, Remote Sensing and Spatial Information Systems, Warsaw, Poland
(paulina.konarzewska.dokt*, wojciech.ostrowski, zdzislaw.kurczynski)@pw.edu.pl

Commission II, WG II/6

KEY WORDS: Artificial Intelligence, Machine Learning, CNN, Oblique Aerial Imagery, Object Detection, Training Datasets

ABSTRACT:

At the time of continuous development of all technologies, deep machine learning (more precisely, convolutional neural networks), which is one of the branches of artificial intelligence (AI), has found wide application in many fields, including photogrammetry and remote sensing. One of the areas where a lot of research is conducted using these methods is the recognition of objects in aerial and satellite imagery. Through the application of deep learning algorithms and neural networks, it is possible to automate labour-intensive processes. However, while object detection in images using machine learning is popular for natural scenes and in recent years also for nadir aerial and satellite imagery, for aerial oblique imagery at the moment of this research there were relatively few publications on the subject. The challengeable task in object detection is the time-consuming generation of training datasets when access is limited or non-existent. This study proposed the methodology to automate this process with use of existing resources for transferring of references to new databases for training models for detect objects on aerial oblique images. The object detection was performed using the YOLOv3 neural network. Experiment results tested on two datasets have shown that the proposed method could realize the task of object detection in oblique aerial images.

1. INTRODUCTION

Machine learning has been widely used in the field of photogrammetry and remote sensing in recent years, especially in the area related to image processing. The development of deep learning algorithms, including convolutional neural networks (CNNs), has resulted in a large amount of current research on automating certain time-consuming processes, including object detection in images. While object detection on natural scenes using machine learning is well developed mainly due to the large number of publicly available learning sets (e.g., ImageNet (Russakovsky et al., 2015), PASCAL VOC 2012 (Everingham et al., 2015), MSCOCO (Lin et al., 2014)), for detection on aerial imagery the algorithms are still being improved. This is caused by the differences that exist between object detection in aerial images and the conventional object detection. The challenge is the variation in scale, orientation, and shape of objects on the Earth's surface, but also due to the dataset bias problem (Torralba, Efros, 2011), more specifically - the degree of generalizability across datasets is often low (Xia et al., 2018).

Over the past 20 years, several different research groups has made its publicly available Earth observation image datasets for object detection. However, in the case of aerial images, the available datasets are not as abundant as mentioned ImageNet or MSCOCO, and the variety of object class categories is poor. One of the example datasets is the TAS set (Heitz, Koller, 2008) intended for the vehicle detection from visible images. SZTAKI-INRIA dataset (Benedek et al., 2011) has been devoted to buildings detection from aerial and satellite images. NWPU VHR-10 (Cheng, Han, 2016) is a collection of images that consists of 3775 objects from 10 different classes. Other datasets used to detect cars on aerial photos are: VEDAI (Razakarivony, Jurie, 2015), UCAS-AOD (Zhu et al., 2015) and The DLR 3K Vehicle (Liu, Mattyus, 2015). For instance, RSOD (Xiao et al., 2015) and HRSC2016 (Liu et al., 2017) datasets are served to detect ships. Before the appearance of DIOR, which was used in this paper, the greatest dataset was DOTA (Xia et al., 2018) consisted of 15 categories of objects and 2806 aerial images. As it can be seen, remarkable efforts have been made to release various object detection datasets in the earth observation community (Li et al., 2020). The mentioned drawbacks related to availability, quantity and quality of the existing datasets for object detection in the Earth Observation domain motivated the development of a new dataset called DIOR (Li et al., 2020). The dataset contains 23 463 images and 192 472 instances, covering 20 object classes. DIOR dataset was applied in experiments to verify transferring reference from satellite scenes to oblique images. The detailed description of this approach can be found in Section 2.

As outlined earlier, training datasets are a particularly important consideration in learning networks for object detection and are an important first step in building a model for automatic detection and recognition in images. The input to learning the network, in addition to images, is information about the exact location of the object in the image and the class to which the object belongs. The location of an object is most often defined by using the coordinates of the bounding boxes. Despite the large availability of tools for labeling and creating references in the form of polygons surrounding objects it is still a manual process and thus is very time consuming.

* Corresponding author

One of the most challenging issue at the time of the experiments was the lack of publicly available datasets to train the network that detects objects in oblique images. Currently, papers are beginning to appear (Heo et al., 2020; Yang et al., 2021), exploiting the potential of aerial oblique images, which is not only the possibility of obtaining information about the location of the object in the terrain system, but also the use of the feature of multi-temporality. Moreover, such images present both the top and side view of an object, which is also an advantage. However, as the availability of learning sets is still low, the authors usually acquire such data themselves (Ruf et al., 2018) or look for other solutions such as fine-tuning based approaches.

To address these problems, the experiments were conducted to evaluate the possibility of transferring references between images with different characteristics in order to use existing datasets to teach the network for detection in oblique images. High-resolution oblique aerial imagery as well as ground (MSCOCO) and satellite nadir data (DIOR) sets were used for these experiments. The deep neural network model for object detection, known as YOLO (You Only Look Once) (Redmon, Farhadi, 2018), has been implemented. YOLOv3 architecture is shown in Figure 1.

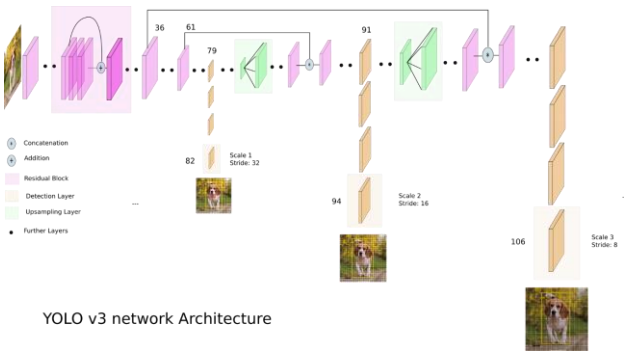

Figure 1 YOLOv3 (You Look Only Once) network architecture (source: https://towardsdatascience.com/).

The main contributions of this paper are as follows:

(1) Experiments have demonstrated the utility of using the YOLOv3 model as an object detector in high-resolution oblique aerial images.

(2) This paper presents methodology to automate the process of generation training datasets with the use of databases that are available in online resources as a starting point for creating new collections for object detection in oblique images. The reference transfer methodology includes both natural scenes (ground photos) and nadir images (derived from airborne and satellite imagery), which can be a valid training dataset for pre-learning the CNN model.

The paper is organized as follows. Section 2 describes the methodology and it is further divided into subsections about description of data, details of an implementation of the used network model, object detection with model trained on MSCOCO and DIOR dataset and finally, the baseline for object detection using the annotations obtained from detection results on oblique images. The results of the experiment and comparison of three approaches are presented in Section 3. The last Section 4 provides the final conclusions drawn from the analyses.

2. METHODOLOGY

2.1 Data Description

The study was conducted using aerial oblique imagery covering the city centre of Bordeaux, France. The learned network model on the Bordeaux data was also tested on oblique data acquired for the city of Elbląg, Poland. These two test areas differ in terms of landscape characteristics and acquired images for this areas have different ground sampling distance (GSD) values (Table 1).

Study Area	Bordeaux	Elbląg
Camera Type	Leica CityMapper	UltraCam Osprey Prime II
GSD	5 cm	10 cm

Table 1. Summary of characteristics of the used image datasets.

The experiments required prior preparation of the data to be processed by the algorithms implemented in the neural network. The images that were used in the following steps for object detection and network learning were divided into 800 x 800 pixel tiles. Due to the outlined in Section 1 limitations of the variety of object classes in the publicly available training data, it was decided to detect objects that occur in almost every training set - cars.
The methodology described in this paper may also be applied to other terrain objects. The category of cars was chosen as an example to save time in manually preparing the training dataset.

2.2 Implementation details of YOLOv3

Considering the review of available convolutional neural network (CNN) models, it was decided to use an implementation of the YOLO algorithm in the latest version available at the time of the research (v3) using Darknet53.

The network implementation was carried out using Python and compiled with OpenCV. The algorithm was also optimized with CUDA technology. The parameters of the virtual machine on which the network training and object detection processes were performed are shown in the following table (Table 2).

Parameter	**Value**
System	Ubuntu 18.04 LTS
CPU	7 cores Intel Xeon Skylake 2.3 GHz
GPU	Nvidia GeForce RTX 2080Ti 11GB
RAM	60 GB

Table 2. Parameters of the virtual machine on which the experiments were performed.

2.3 Detection with model trained on MSCOCO

As no publicly available dataset existed to train a network that detects objects for oblique aerial images, it was decided to conduct a first experiment to evaluate the feasibility of using available natural scene datasets for object detection in oblique images. The YOLOv3 algorithm previously learned on MSCOCO data was used for this first approach. The experiment was performed for two datasets, Elblag and Bordeaux. The implementation consisted of cloning the project from a

repository on GitHub and taking the trained weights for the MSCOCO classes. Although the MSCOCO set contained many more classes, the evaluation of detection was performed for the objects of interest - cars. The detector parameters assumed to detect the most confident results were: a confidence value (which defines the probability that an object was correctly detected) of 50% and a threshold value (the confidence level with which to assign bounding boxes) of 30%.

The results, obtained from the first tests, were not satisfying. The network was able to detect some of the cars, however it did not manage to detect many objects, especially when the cars were visible from the front or the back in the picture, which was noticed during the verification. The result obtained was quite expected, as cars are mapped differently in the ground photo than in the aerial oblique image. The figures below show the results of the first experiment (Figure 2).

a)

b)

Figure 2 Results of the first experiment - object detection with the YOLO algorithm learned on MSCOCO images: a) Bordeaux, b) Elbląg.

2.4 Detection with model trained on DIOR

Detection using the model trained on ground photos for both the Bordeaux area and the Elblag images did not give valid enough results. Thus, it was decided to conduct another experiment - this time the trial was to check the transfer of references from nadir satellite scenes - DIOR dataset to oblique aerial images.

While the first approach used the downloaded trained weights for prediction (with no training process), the second experiment required YOLO to be learned from scratch. Although the object of interest in this research were cars, all object categories available in the DIOR were used for learning, due to the fact that besides the number of images and objects in the training dataset, the number of object classes can also have a large impact on the results. While the algorithm is learned on a larger number of classes, it is able to distinguish objects from each other more easily.

The dataset contained a total of 23 463 images and was divided into training and validation sets in a 3 to 1 relationship. The table below (Table 3) summarizes the parameters used in learning process of the YOLO network on DIOR data.

Parameter	Value
Batches	64
Subdivisions	16
Max_batches	40000
Classes	20
Filters	75
Network size - width	800
Network size - height	800

Table 3. Parameters adjusted for the DIOR dataset used during learning YOLO network.

When analysing the results, it was noticed that the YOLO model learned on the DIOR set handled the car detection problem a little better than the YOLO learned on the MSCOCO set, but still the number of detected objects was not satisfactory. The results of the second experiment are shown in the example images below (Figure 3).

a)

b)

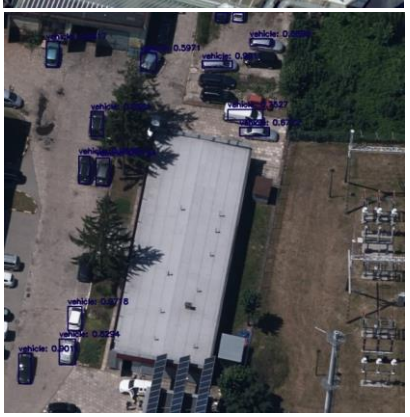

Figure 3 Results of the second experiment - object detection with the YOLO algorithm learned on DIOR images: a) Bordeaux, b) Elbląg.

On the other hand, when reviewing the images with the prediction applied, it was noticed that the detection results with two different models for the same area sometimes complement each other. In some places, where cars were not detected in one approach, the algorithm recognized them in the second, and vice

versa. The example dependence is shown in the following figure (Figure 4).

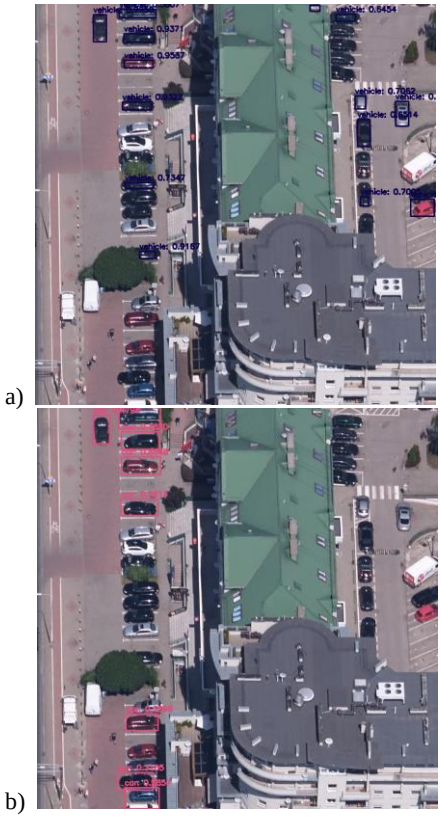

Figure 4 Example where detection results from YOLO learned on two different image sets partially complement each other: a) MSCOCO, b) DIOR – for images from Elblag dataset.

2.5 Detection with model trained on annotations obtained from detection results on oblique images

Based on the results of the second approach, it was decided to run the final experiment. From the set of oblique images from Bordeaux, 605 of images were selected as a learning set and divided into 78650 small tiles on which the model was passed twice (once trained on MSCOCO and a second time on DIOR) to predict the bounding boxes of the cars.

The results of oblique image detection from the Bordeaux area from both the network learned on the natural scenes set (MSCOCO) and the network learned on the satellite set (DIOR) were used to learn a network dedicated to car detection in oblique images. However, before proceeding to re-train the network using oblique image detection, the results from the two variants had to be combined so that would not duplicate. Furthermore, objects with frame dimension less than 20x20 pixels or any of the sides of bounding boxes had less than 15 pixels were removed from the newly created training set. The combined detection results from these two approaches became the new set for training the network for detection on the target oblique dataset.

The network training process was similar to step two with few differences in parameter settings: the number of max batches (defining the number of iterations to be performed by the algorithm) was changed to 23000, the number of classes to 1 and the number of filters to 18.

3. RESULTS

The experiments described above have yielded a YOLO network model trained in three variants:

- YOLO trained with the MSCOCO dataset
- YOLO trained with satellite images from the DIOR dataset
- YOLO trained on annotations obtained from detection results from two approaches.

In order to compare the detection results from the three different variants, a test area was selected for which the statistics were calculated. The test dataset for the Bordeaux data consisted of 10 images (2 for each direction), i.e. 1300 tiles of 800x800 resolution. In the case of the Elblag data, the set was less numerous and was used to see how the algorithm trained on photos from Bordeaux behaves in an area with different characteristics.

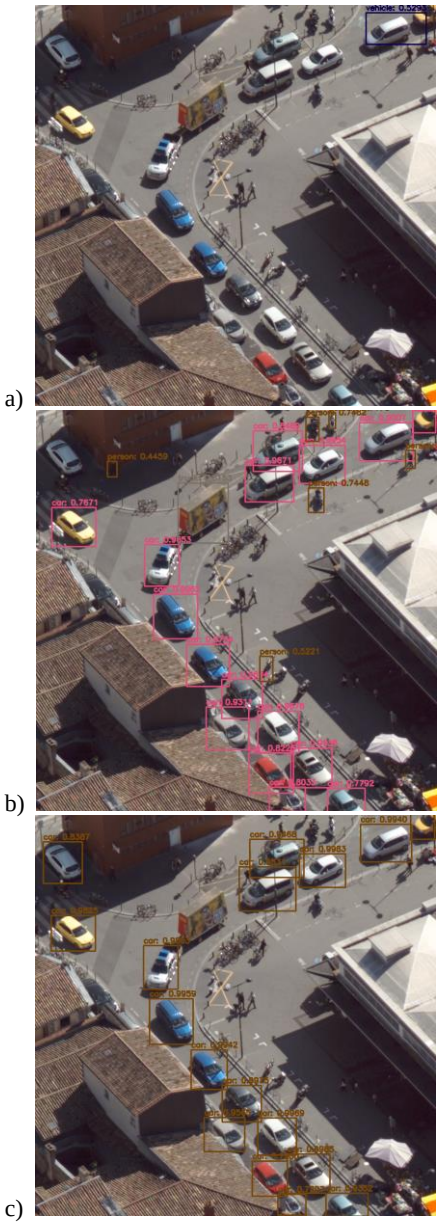

Figure 5. Detection results by YOLO network trained on three different datasets: a) DIOR, b) MSCOCO, c) a custom set consisting of detection results (annotations) on oblique images from the previous two steps (test area from Bordeaux).

As can be seen in the images showing the detection results (Figure 5), the best performance was obtained for third approach - for YOLO learned from oblique images (annotations obtained from detection results from previous approaches). For the Elbląg test area, similar conclusions were drawn, as the third model performed best with the detection task (Figure 6).

a)

b)

c)

Figure 6. Detection results by YOLO network trained on three different datasets: a) DIOR, b) MSCOCO, c) a custom set consisting of detection results (annotations) on oblique images from the previous two steps (test area from Elbląg).

Beyond the visual analysis, the paper also performed a quantitative analysis of accuracy. An important parameter to consider in object detection is the confidence with which a given object was detected in the image. Therefore, the first metric to evaluate the results was the overview and the comparison with which confidence threshold a given algorithm most often detected objects. As can be seen in the chart below (Figure 7) for the Bordeaux image set, objects with the weakest confidence were detected by the network learned on the DIOR dataset. In contrast, the YOLO model learned using detection on oblique images proved to be the best, where the prevailing detections showed high confidence (between 95% and 100% values). The algorithms behaved similarly on the Elblag data for the third model and the worst results were obtained for the model trained on MSCOCO.

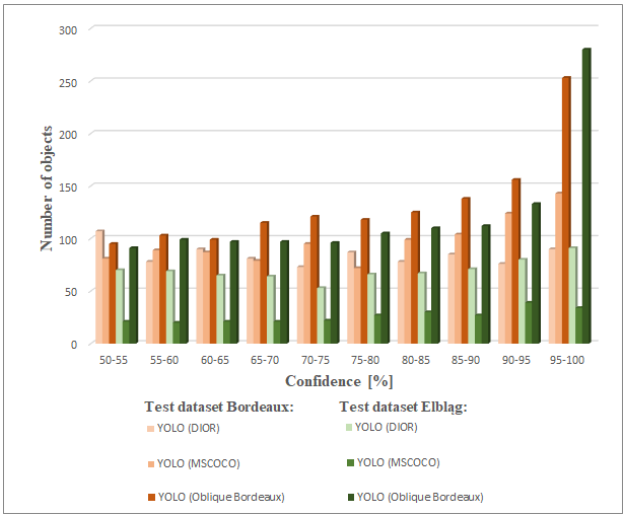

Figure 7. Plot of the dependence of the number of objects on the confidence with which were detected - for the Elblag (green colours) and for the Bordeaux (red colours) test sets.

As part of the accuracy assessment, a confusion matrix was created to show the statistics. The values of TP (true-positive), TN (true-negative), FN (false-negative), as well as accuracy (1) and recall (2), expressed in %, were included in the summary. Evaluation metrics are defined as follows:

$$Accuracy\ [\%] = \frac{TP}{TP + FN} * 100\% \quad (1)$$

$$Recall\ [\%] = \frac{TP}{TP + FP} * 100\% \quad (2)$$

	Model trained on DIOR data	Model trained on MSCOCO data	Model trained on detection results
a) Results for Bordeaux			
True Positive (TP)	845	973	1323
False Positive (FP)	89	48	54
False Negative (FN)	1270	1142	792
Accuracy [%]	39.95	46.00	**62.55**
Recall [%]	90.47	95.30	**96.08**
b) Results for Elblag			
True Positive (TP)	884	764	1128
False Positive (FP)	118	22	24
False Negative (FN)	1030	1150	786
Accuracy [%]	46.19	39.92	**58.93**
Recall [%]	88.22	97.20	**97.92**

Table 4. Summary of detection results - comparison with the ground truth a) Bordeaux and b) Elbląg.

The results of the above experiments support the conclusion that it is possible to use existing datasets (both terrestrial and satellite) that are publicly available to train network models to detect objects in oblique aerial images, and use the detection results as a learning dataset. As can be seen in the table above

(Table 4), the accuracies are highest for the YOLO model learned on oblique images, while this solution does not guarantee very high accuracies: 62.55% for Bordeaux and 58.93% for Elbląg. However, given the high recall values of the algorithm (96.08% for Bordeaux and 97.92% for Elbląg) learned on the annotations obtained from detection results, it is possible to use the methodology presented in this paper (Figure 8) to create references in a more automatic way, thus reducing the effort of manually labelling objects in the images using available tools and alleviating the cost.

The purpose of testing the model trained on the annotations obtained from detection results with oblique images from Bordeaux on test area of Elbląg city was to analyze how the model works with data in photos with a different landscape character. Similar accuracy values were obtained and it can be concluded from this that the model was not fit too closely to the training set and "overfitted".

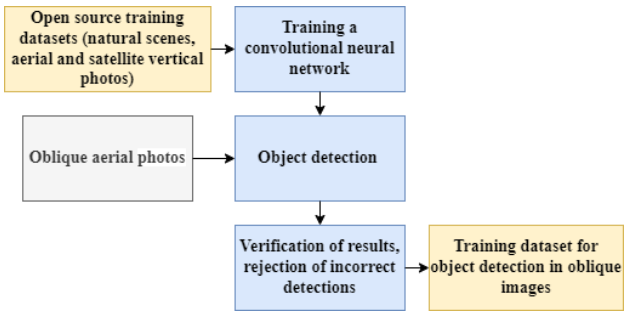

Figure 8. Flowchart of the methodology for creating a dataset for object detection in oblique images.

4. CONCLUSIONS

Machine learning methods for object detection and image classification are constantly being developed. New models are still being created and the existing neural network architectures are being improved with new versions. The experiments conducted in this research showed that the use of convolutional neural networks for object detection in images can be applied not only to natural scenes or nadir aerial and satellite imagery, for which the technique is popular, but also to high-resolution oblique aerial images.

The main research problem addressed in this paper is related to the lack of training datasets for object detection in oblique aerial photos. The experiments investigated the transferability of references available in online resources. The results showed that it is possible to apply the proposed methodology to at least partially reduce the problem related to the lack of availability of labelled training datasets. Both ground images and aerial or satellite nadir images may provide a suitable training dataset for pre-training a neural network.

The accuracy of the YOLOv3 model trained using three different approaches was evaluated. Although the accuracy values of the YOLOv3 detection results learned on oblique images were not very high (about 60%), the recall of the algorithm reached over 90%. The high value of this metric indicates that the algorithm made a small number of errors during detection. Based on this, it can be concluded that this model can be used as a semi-automatic approach to create training and test datasets on oblique images. This makes it possible to speed up the work, which currently in practice comes down to manually labelling and creating references.

Summarizing, the paper demonstrates the usefulness of using YOLOv3 neural network for object detection in aerial oblique images. Furthermore, the proposed methodology for creating a training dataset allows for semi-automation. In general, the following work allows to see the potential of using artificial intelligence systems in the field of photogrammetry and remote sensing and provides a basis for using advanced technologies to accelerate image data processing.

ACKNOWLEDGEMENTS

The experiments included in this paper were conducted as part of the project "Methodology for automation of object database creation from synchronously acquired hybrid aerial photogrammetric data". The elaboration of data from the Elblag area was performed independently from the research conducted during the project work. Data for the Elbląg area was provided by the Department of Real Estate and Geodesy, Elbląg City Hall. The authors would like to acknowledge Hexagon providing the Bordeaux CityMapper-1 dataset. The project was realized at the Faculty of Geodesy and Cartography at the Department of Photogrammetry, Remote Sensing and Spatial Information Systems, Warsaw University of Technology on the commission of the OPEGIEKA Sp. z o.o. company.

REFERENCES

Benedek, C., Descombes, X., & Zerubia, J. (2011). Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1), 33-50.

Cheng, G., & Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117, 11-28.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1), 98-136.

Heitz, G., & Koller, D. (2008, October). Learning spatial context: Using stuff to find things. In *European conference on computer vision* (pp. 30-43). Springer, Berlin, Heidelberg.

Heo, W. Y., Kim, S., Yoon, D., Jeong, J., Sung, H., 2020. Deep learning based moving object detection for oblique images without future frames. In *Automatic Target Recognition XXX* (Vol. 11394, p. 1139403). International Society for Optics and Photonics.

Kathuria A. (2018, April). What's new in YOLO v3? https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b (Access on March 25, 2022).

Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159, 296-307.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.

Liu, K., & Mattyus, G. (2015). Fast multiclass vehicle detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, 12(9), 1938-1942.

Liu, Z., Yuan, L., Weng, L., & Yang, Y. (2017, February). A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods* (Vol. 2, pp. 324-331). SciTePress.

Razakarivony, S., & Jurie, F. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34, 187-203.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ruf, B., Thiel, L., Weinmann, M., 2018. Deep cross-domain building extraction for selective depth estimation from oblique aerial imagery. arXiv preprint arXiv:1804.08302.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International journal of computer vision*, 115(3), 211-252.

Torralba, A., & Efros, A. A. (2011, June). Unbiased look at dataset bias. In *CVPR 2011* (pp. 1521-1528). IEEE.

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. DOTA: A large-scale dataset for object detection in aerial images. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3974-3983).

Xiao, Z., Liu, Q., Tang, G., & Zhai, X. (2015). Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *International Journal of Remote Sensing*, 36(2), 618-644.

Yang, C., Zhang, F., Gao, Y., Mao, Z., Li, L., Huang, X., 2021. Moving Car Recognition and Removal for 3D Urban Modelling Using Oblique Images. *Remote Sensing*, 13(17), 3458.

Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., & Jiao, J. (2015, September). Orientation robust object detection in aerial images using deep convolutional neural network. *In 2015 IEEE International Conference on Image Processing (ICIP)* (pp. 3735-3739). IEEE.