Deep learning based multi-category object detection in aerial images

Lars W. Sommer[a,b], Tobias Schuchert[b], and Jürgen Beyerer[b,a]

[a]Vision and Fusion Lab, Karlsruhe Institute of Technology KIT, Adenauerring 4, 76131 Karlsruhe, Germany
[b]Fraunhofer IOSB, Fraunhoferstrasse 1, 76131 Karlsruhe, Germany

ABSTRACT

Multi-category object detection in aerial images is an important task for many applications such as surveillance, tracking or search and rescue tasks. In recent years, deep learning approaches using features extracted by convolutional neural networks (CNN) significantly improved the detection accuracy on detection benchmark datasets compared to traditional approaches based on hand-crafted features as used for object detection in aerial images. However, these approaches are not transferable one to one on aerial images as the used network architectures have an insufficient resolution of feature maps for handling small instances. This consequently results in poor localization accuracy or missed detections as the network architectures are explored and optimized for datasets that considerably differ from aerial images in particular in object size and image fraction occupied by an object. In this work, we propose a deep neural network derived from the Faster R-CNN approach for multi-category object detection in aerial images. We show how the detection accuracy can be improved by replacing the network architecture by an architecture especially designed for handling small object sizes. Furthermore, we investigate the impact of different parameters of the detection framework on the detection accuracy for small objects. Finally, we demonstrate the suitability of our network for object detection in aerial images by comparing our network to traditional baseline approaches and deep learning based approaches on the publicly available DLR 3K Munich Vehicle Aerial Image Dataset that comprises multiple object classes such as car, van, truck, bus and camper.

Keywords: deep learning, multi-category object detection, aerial imagery

1. INTRODUCTION

Detecting objects in aerial images and classifying these objects into different categories is an important task for many applications such as surveillance, tracking or search and rescue tasks. In general, object detection in aerial images is based on hand-crafted features and a classifier or cascade of classifiers within a sliding window approach [1]. In recent years, deep learning based detection frameworks, *e.g.* Fast R-CNN [2] and Faster R-CNN [3], using features extracted by convolutional neural networks (CNN) significantly improved the detection accuracy on detection benchmark datasets compared to traditional approaches based on hand-crafted features [2, 3]. However, these detection benchmark datasets, *e.g.* Pascal VOC 2007 [4] differ considerably from aerial images. For instance, images of these datasets generally contain only one or a few objects that are typically centered and occupy a high fraction of the image [5] whereas aerial images can contain multiple objects that are smaller and randomly located.

In this paper, we evaluate the applicability of Faster R-CNN [3], which achieves top performing results on the detection benchmark datasets, for multi-category detection in aerial images. As this detection framework is optimized for the detection benchmark datasets, it is not transferable one to one on aerial images. To account for the characteristics of the aerial images, in particular the size of occurring objects ($\sim$15$\times$30 pixels); we adapt the parameters of the Region Proposal Network (RPN) used to generate a set of candidate regions that are classified in the subsequent Fast R-CNN module. We further propose our own network as the originally used network architecture exhibits an insufficient resolution of the applied feature map for handling small instances [6].

Further author information: {lars.sommer, tobias.schuchert, juergen.beyerer}@iosb.fraunhofer.de

The insufficient resolution results in poor localization accuracy or missed detections [7]. We further analyze the impact of the adapted RPN and Fast R-CNN module on the detection performance. For our experiments, we use the publicly available DLR 3K Munich Vehicle Aerial Image Dataset [8] that comprises multiple object classes, *e.g.* car and truck. To account for the sparse number of annotated training data as in case of category truck, we investigate the impact of data augmentation. Finally, we compare the adapted Faster R-CNN to the original approach and state-of-the-art vehicle detection methods from literature.

The remainder of this paper is organized as follows: Section 2 gives an overview of related work. In Section 3, we introduce the functional principle of Faster R-CNN and describe our performed adaptions to aerial images. Experimental results are given in Section 4. Finally, we conclude in Section 5.

2. RELATED WORK

In literature, there exists a broad variety of detection methods for detecting vehicles in aerial images [1]. These methods are generally based on hand-crafted features and a classifier or cascade of classifiers within a sliding window approach. For this, several different feature and classifier combinations such as HOG and SVM [9, 10] or ICF and AdaBoost [8, 11] have been proposed. Recently, several authors applied CNNs to classify candidate regions within a sliding window approach to overcome drawbacks of hand-crafted features. However, the computation of convolutional features for every candidate window is expensive [2].

Recently, object detection frameworks such as Fast R-CNN [2] and Faster R-CNN [3] clearly reduced the computational costs by sharing convolutional features across region proposals instead of computing the convolutional features for each candidate window separately. In [6], the authors exhibit the potential and limitations of Faster R-CNN for medium sized objects as in case of pedestrian detection. In [7], the authors demonstrate the applicability of Fast R-CNN and Faster R-CNN for vehicle detection in aerial images. Adaptions to account for the characteristics of the aerial images have been discussed in detail. The main improvement in performance was achieved by increasing the feature map resolution. However, only one vehicle category was considered. In the following, we extend the detection task to multiple categories.

3. DETECTION METHOD

In the following section, we introduce the Faster R-CNN detection method, which achieves top performing results on common detection benchmark datasets such as Pascal VOC [4]. Then, we describe adaptions performed to account for the characteristics of aerial images. Therefore, we propose our own network, which is optimized for handling small instances such as vehicles in aerial images.

3.1 Faster R-CNN

Faster R-CNN is composed of two modules: a deep learning based object proposals method called Region Proposal Network (RPN) and the Fast R-CNN object detector proposed by [2]. The RPN and Fast R-CNN object detector are merged into a single network by sharing their convolutional layers. First, the RPN generates a set of candidate regions. For this, a small network is shifted in a sliding window approach over the output of the last shared convolutional layer, which is used as feature map. Then, the output of the network is fed into a classification layer and bounding box regression layer. The classification layer provides a confidence value used to rank the proposals whereas the output of the bounding box regression layer is the corresponding coordinates. For this, anchor boxes centered at each sliding window location are used as bounding box reference. The region proposals with the highest confidence values are used as input for the Fast R-CNN object detector. Then, the region of each region proposals is projected on the feature map, which is the output of the last shared convolutional layer. The corresponding features are extracted by a region of interest pooling layer to a feature vector of fixed length. Each feature vector is fed into a sequence of fully connected layers followed by two sibling output layers, which provide classification scores and the refined coordinates of the corresponding bounding box.

Table 1. Basic structure of the proposed CNN to handle small object sizes.

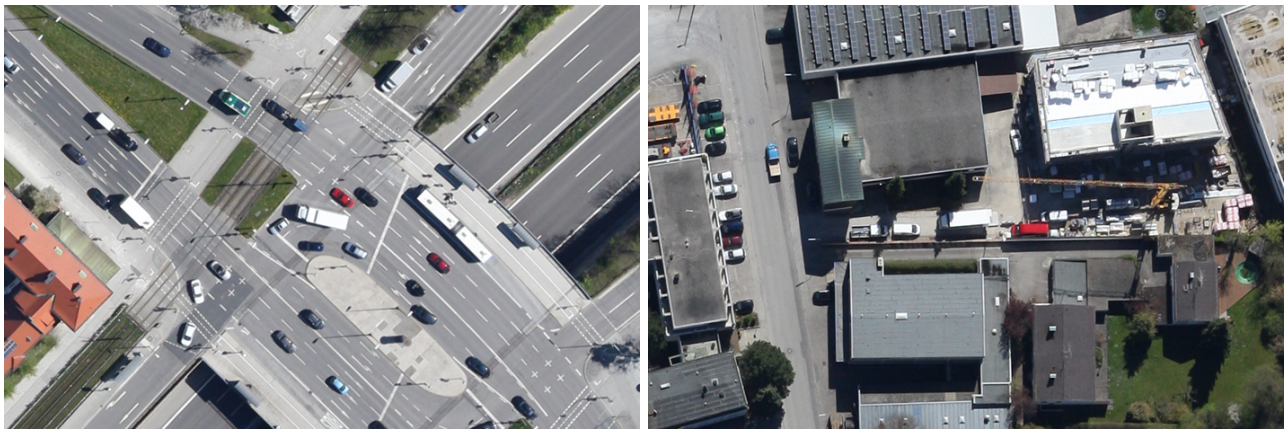
input size: 32x32			
type	**kernel size**	**stride, pad**	**data size**
conv	3×3×128	1, 1	32×32×128
relu			32×32×128
pool	3×3	2,0	16×16×128
conv	3×3×256	1, 1	16×16×256
relu			16×16×256
pool	3×3	2, 0	8×8×256
conv	3×3×384	1, 1	8×8×384
relu			8×8×384
conv	3×3×512	1, 1	8×8×512
relu			8×8×512
pool	3×3	2, 0	4×4×512
fc			1024
relu			1024
dropout			1024
fc			512
relu			512
fc			2

3.2 Adaption to Aerial Imagery

As shown in [7], the performance of Faster R-CNN mainly depends on the localization quality of the proposals generated by the RPN and the resolution of the feature map. The original Faster R-CNN is based on the publicly available VGG-16 net [12] pre-trained on the ImageNet dataset [13]. The stride for the VGG-16 net on the last convolutional layer, which is used as feature map, is 16 [3]. Thus, the resolution of the feature map is only 1/16 of the input image and is consequently insufficient for object sizes between 15 and 30 pixels [7]. To overcome this shortcoming, we replace the VGG-16 net by a network especially designed for handling small object sizes. The proposed network is inspired by [14] and given in Table 1. The proposed network takes input images of size 32×32 pixels, which is roughly equivalent to the size of objects present in aerial images like the DLR 3K Munich Vehicle Aerial Image Dataset. All relevant network parameters including number of layers, number of filters per layer, kernel size and dropout were optimized for classification. For this, about 30,000 positive and 100,000 negative samples have been considered as training data. Analogous to the original Faster R-CNN, we replace the pooling layer after the last convolutional layer by a region of interest pooling layer used to convert the features inside any valid region of interest into a small feature map with a fixed spatial extent as required for the following fully connected layers. In this case, the stride on the last convolutional layer is 4 and the resolution of the feature map is 1/4 of the input image. As the localization quality of the generated proposals not only depends on the feature map resolution, we set the anchor base size to 7 and the scale parameter to 4 as proposed in [7].

3.3 Implementation Details

We adopt the alternating training as proposed [3]. To account for smaller object sizes, we reduce the minimal height and width of region proposals considered for classification to 4 pixels. As the aerial images can contain considerably more objects per image compared to detection benchmark datasets, we increase the number of top scoring boxes to keep before and after applying Non-Maximum Suppression (NMS) to the region proposals to 20,000 and 5,000, respectively. The number of object classes ist set to 3 (background, car and truck). We further set the lower overlap threshold value used to consider region proposals as background to 0.0. Otherwise only region proposals that slightly overlap with Ground Truth (GT) objects and consequently only a minor part of the background would be considered for negative samples. To avoid re-scaling of the images, we set the scale of an images shortest side to 624 pixels. Other hyper-parameters are adopted from [3]. For testing, we set the minimal height and width of region proposals considered for classification to 4 pixels as well. The number of top



(a) (b)

Figure 1. Illustrative examples of the DLR 3K Munich Vehicle Aerial Image Dataset [8] showing varying sceneries such as a busy intersection (a) or industrial areas (b).

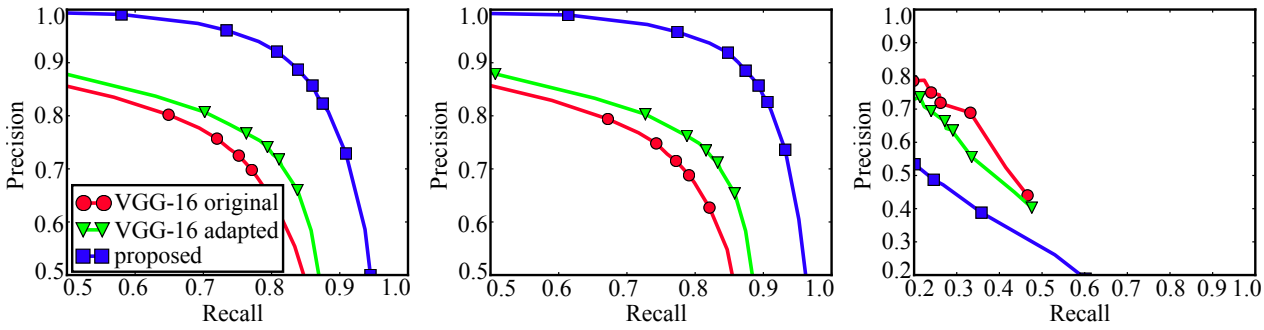
scoring boxes to keep before and after applying NMS to the region proposals is set to 10,000 and 2,000 as fewer proposals would result in worse performance [7].

4. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of Faster R-CNN for aerial images. Therefore, all experiments are performed on the publicly available DLR 3K Munich Vehicle Aerial Image Dataset [8], which comprises 20 images with a resolution of 5616×3744 pixels and a ground sampling distance of approximately 13 cm. Illustrative image sections of the DLR 3K Munich Vehicle Aerial Image Dataset are given in Figure 1. Each image comprises about 450 vehicles on average. For our experiments, we consider two classes (car and truck). Further classes, e.g. bus, are not considered due to the very limited number of annotated objects. We further divided each image into tiles of size 936×624 pixels to speed up the training and we aligned the GT at image edges as required for Faster R-CNN. We use the first 50% of the images as training data and the rest for testing. For testing, we performed data augmentation to increase the number of occurring trucks. For this, we rotated images that include a truck by angles of 30, 60 and 90 degree, respectively. To evaluate the detection performance, we use Precision-Recall curves, Average Precision (AP) and mean Average Precision (mAP) computed as defined in [4]. GT objects are accepted as recalled if the PASCAL-overlap criterion [4] is satisfied.

4.1 Adaption to Aerial Images

In this section, we compare the performance of the original Faster R-CNN and the performance of the Faster R-CNN adapted to the characteristics of aerial images as described in Section 3.2. Figure 2 (a) shows Precision-Recall curves for all vehicles. The performance of the original approach (red curve) is clearly improved by adapting the RPN (green curve) to generate proposals that are more likely to fit the size of present objects. For this, we reduce the initial anchor base size to 2. Thus, anchor boxes that are used as bounding box reference (see Section 3.1) are in the size of present objects. The performance is further clearly improved by replacing the initially used network by our proposed network (blue curve) and consequently increasing the feature map resolution as described in Section 3.2. In this case, the size of the used anchor boxes are already adapted to the size of present objects. Reasons for the improved performance due to the performed adaptions are described in detail in [7]. Poor localization quality of region proposals due to inappropriate anchor boxes as well as the insufficient feature map resolution in case of using the VGG-16 net result in an inaccurate localization of the detections. Figure 2 (b) shows the Precision-Recall curves for the category car. Both adaptions result in clearly improved precision and recall values. Figure 2 (c) shows the Precision-Recall curves for the category truck. In contrast to the category car, the performance is not improved by performing the proposed adaptions. However, the performance is poor for all different approaches. Reason for that is the small number of trucks available in the training data.



(a) all vehicles (b) car (c) truck

Figure 2. Precision-Recall curves for all vehicles (a), category car (b) and category truck (c). The performance for all vehicles and category car is clearly improved by the proposed adaptions especially by increasing the feature map resolution. However, the performance for category truck is poor due to the small number of trucks available in the training data.

4.2 Data Augmentation

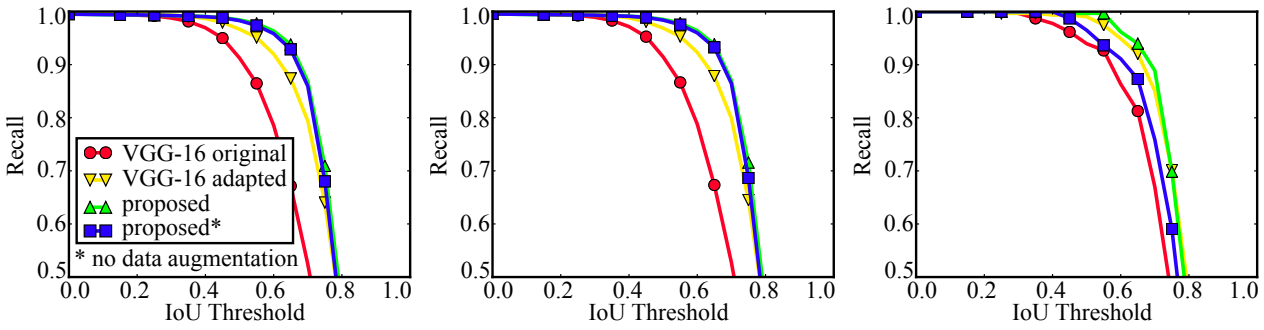
To account for the small number of trucks, we perform data augmentation. Therefore, we rotate images that include a truck by angles of 15, 30, 45, 60, 75, 90, 105, 120, 135, 150 and 165 degree. The original training data contains only about 100 annotated trucks. Thus, several types, scales and orientations are not or only insufficiently represented. Consequently, several trucks of the test data can not be detected and classified correctly. The data augmentation results in more than 1,000 annotated trucks and a high number of trucks is available for each orientation. Table 2 shows the AP and the mAP for each method before and after data augmentation. The APs before data augmentation correspond to the Precision-Recall curves shown in Figure 2 (b) and (c), respectively, whereas the mAP is the mean value for all AP. The AP of category car is clearly improved by the performed adaptions especially for the increased feature map resolution. The AP of category truck is poor for the original Faster R-CNN and for the adapted Faster R-CNN. The AP of both categories and the mAP are clearly improved by using the augmented training data. The AP of category car is improved as well as the number of cars (with different orientations) is clearly enhanced by the performed data augmentation. The most improvement is achieved for category truck as the training data comprises considerably more trucks for all different orientations. However, the AP is still worse compared to category car. Reason for that are discussed in Section 4.4.

Table 2. Comparison of the Average Precision (AP) before and after data augmentation. The AP is clearly improved for all methods and object categories especially for category truck that posseses the fewest annotations.

Method	Car	Truck	mAP
VGG-16	0.588	0.306	0.447
VGG-16 adapted	0.650	0.275	0.462
proposed Net	0.916	0.253	0.585
VGG-16 with data augmentation	0.659	0.698	0.679
VGG-16 adapted with data augmentation	0.719	0.698	0.709
proposed Net with data augmentation	**0.936**	**0.713**	**0.825**

4.3 Analysis of the Faster R-CNN Modules

In the following section, we analyze the impact of the RPN and the Fast R-CNN detector on the detection performance. To evaluate the performance of the RPN, we use Recall-IoU curves. For this, we vary the Intersection over Union (IoU) threshold used to accept a GT object as covered by a region proposal. Figure 3 shows the Recall-IoU curves for all vehicles (a), category car (b) and category truck (c). The recall values of the original



(a) all vehicles (b) car (c) truck

Figure 3. Recall-IoU curves for all vehicles (a), category car (b) and category truck (c). The recall values are clearly improved by the performed adaptions and by data augmentation especially for category truck.

RPN (red curve) are clearly improved for all vehicles and both categories by adapting the anchor box sizes (yellow curve) as described in Section 3.2. Anchor boxes used as reference in the size of present objects result in proposals, which exhibit better overlap with GT objects. Increasing the feature map resolution (green curve) further improves the recall values for all categories. The recall values for category car are only slightly improved by data augmentation, whereas the recall values for category truck are clearly improved. The recall values of category truck are slightly higher compared to category car. This indicates that the region proposals are not the reason for the worse detection performance of category truck.

Table 3. Confusion matrices of category classification with and without data augmentation. The classification accuracy for category truck is clearly improved with more training data.

no data augmentation			with data augmentation		
A/P*	Car	Truck	A/P*	Car	Truck
Car	4517	3	Car	4630	12
Truck	25	33	Truck	7	202

* Actual class / Predicted class

To analyze the Fast R-CNN module, we use confusion matrices of category classification. Table 3 shows the confusion matrices for Faster R-CNN using our proposed network with and without data augmentation. Therefore, only detections that satisfy the PASCAL-overlap criterion are considered. The given values are achieved for a classification threshold value of 0.8, which shows the best f-score values for both cases. The classification accuracy of category car is close to one with and without data augmentation, whereas the classification accuracy of category truck is considerably improved. Furthermore, the number of missed detections is clearly reduced for both categories. However, the number of false positive detections (85) and false negative detections (111) are still high for category truck. Thus, the classification between category truck and the background has to be improved. In the following section, we show some qualitative results, to verify this assumption.

Both recall (87,4%) and precision rate (91,3%) are clearly improved for all vehicles compared to literature. Liu et al. [8] achieve a recall rate of 69.3% and a precision rate of 86.8% on DLR 3K by applying ICF and an AdaBoost classifier in a soft cascade structure. However, recall and precision rates for the particular categories are not provided.

4.4 Qualitative Results

Figure 4 shows true positive detections exemplarily for both categories. For this, we use our proposed network and the augmented training data. The detections for category car (red boxes) and category truck (dark blue boxes) show good overlap with the corresponding GT (green and light blue boxes, respectively) for varying backgrounds and sceneries as well as for varying density of present objects. Figure 5 shows false positive examples for category truck (dark blue boxes). The main reason for false positive detections are objects with similar shapes such as trailers. The high number of trailers detected as trucks is caused by the training samples. Regions of interest,



Figure 4. Detection examples for category car (red boxes) and category truck (dark blue boxes) and corresponding Ground Truth for category car (green boxes) and category truck (light blue) for Faster R-CNN using our proposed network and with data augmentation.

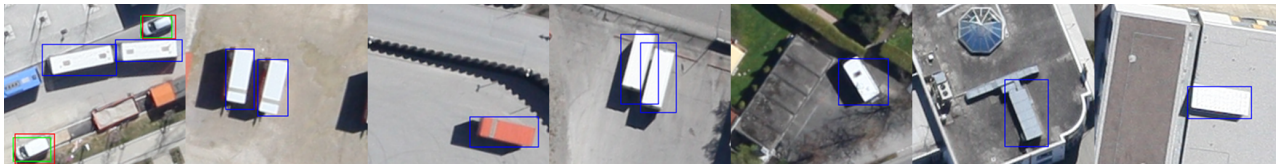


Figure 5. False positive examples for category truck (dark blue boxes).

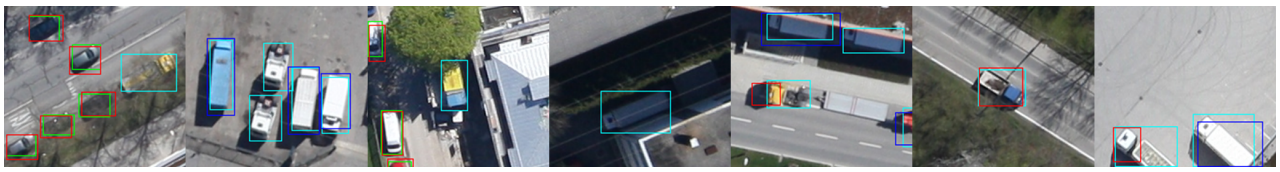


Figure 6. False negative examples for category truck (light blue boxes).

which have an overlap of 0.5 with a GT object, are considered as positive sample. Thus, several positive examples only contain the trailer of the truck. To reduce the number of such positive samples, the overlap used to consider objects as positive samples should be increased. Figure 6 shows false negative examples for category truck (dark blue boxes) as well as trucks that are classified as car. One main reason for false negative detections are types of trucks, which are not present in the training data, and therefore not detected in the test data (trucks without trailer in second column). Furthermore, trucks, which are partially occluded by trees or in the shadow, are often missed. Thus, more training data with a greater variety of trucks is required to reduce the number of false negative detections.

5. CONCLUSIONS

In this paper, we have demonstrated the applicability of Faster R-CNN for multi-category object detection in aerial images. We adapted relevant parameters of the detection approach to account for the characteristics of the aerial images. The main improvement was achieved by replacing the initially used VGG-16 net by our proposed network to increase the resolution of the feature map to account for the lower spatial resolution of the aerial images. Due to the small number of annotated trucks, we performed data augmentation, which considerably improved the detection performance for category truck. However, the variety of present trucks is limited and should be increased to account for different types of trucks. In future work, we will investigate the impact of the threshold value used to consider regions of interest as positive samples to reduce the number of false positive detections for category truck. Furthermore, we will investigate the applicability of further detection frameworks, *i.e.* SSD, for multi-category object detection in aerial images.

REFERENCES

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

[2] Girshick, R., "Fast r-cnn," in [*Proceedings of the IEEE International Conference on Computer Vision*], 1440–1448 (2015).

[3] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," in [*Advances in neural information processing systems*], 91–99 (2015).

[4] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *International journal of computer vision* **88**(2), 303–338 (2010).

[5] Lenc, K. and Vedaldi, A., "R-cnn minus r," *arXiv preprint arXiv:1506.06981* (2015).

[6] Zhang, L., Lin, L., Liang, X., and He, K., "Is faster r-cnn doing well for pedestrian detection?," in [*ECCV*], 443–457, Springer (2016).

[7] Sommer, L. W., Schuchert, T., and Beyerer, J., "Fast deep vehicle detection in aerial images," in [*Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*], IEEE (2017 accepted, In press).

[8] Liu, K. and Mattyus, G., "Fast multiclass vehicle detection on aerial images," *Geoscience and Remote Sensing Letters, IEEE* **PP**(99), 1–5 (2015).

[9] Tuermer, S., Kurz, F., Reinartz, P., and Stilla, U., "Airborne vehicle detection in dense urban areas using hog features and disparity maps," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **6**(6), 2327–2337 (2013).

[10] Moranduzzo, T. and Melgani, F., "Detecting cars in uav images with a catalog-based approach," *IEEE Transactions on Geoscience and Remote Sensing* **52**(10), 6356–6367 (2014).

[11] Teutsch, M. and Krüger, W., "Robust and fast detection of moving vehicles in aerial videos using sliding windows," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*], 26–34 (2015).

[12] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[13] https://people.eecs.berkeley.edu/~rbg/faster-rcnn-data/ (2015).

[14] Herrmann, C., Willersinn, D., and Beyerer, J., "Low-Resolution Convolutional Neural Networks for Video Face Recognition," in [*Advanced Video and Signal Based Surveillance*], IEEE (2016).