A Comprehensive Study on Object Proposals Methods for Vehicle Detection in Aerial Images

Lars Wilko Sommer[1,2], Tobias Schuchert[2] and Jürgen Beyerer[2,1]

[1]Vision and Fusion Lab
Karlsruhe Institute of Technology KIT
Adenauerring 4, 76131 Karlsruhe, Germany

[2]Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation IOSB
Fraunhoferstrasse 1, 76131 Karlsruhe, Germany

{lars.sommer|tobias.schuchert|juergen.beyerer}@iosb.fraunhofer.de

***Abstract*—Detecting vehicles in aerial images is an important task in many applications such as traffic monitoring or screening of large areas. In general, vehicle detection in aerial images is performed by applying classifiers or a cascade of classifiers within a sliding window algorithm. However, detecting vehicles in a real-time system is limited by the huge number of windows to classify, especially in case of varying object scales, aspect ratios or object orientations. To reduce the high number of windows, we propose to apply so called object proposals methods. In recent years, several object proposals methods have been proposed for generating candidate windows in detection frameworks. However, aerial images differ considerably from datasets that are typically used for exploring such methods. To examine the applicability of such methods for aerial images, we evaluate 11 state-of-the-art object proposals methods on the publicly available DLR 3K Munich Vehicle Aerial Image Dataset. First, we manually modified the provided ground truth data to enable comparison to the generated object proposals. To compensate for the differing characteristics of the aerial images, we adapted seven methods by examining different parameter settings and extensions for each method separately. Finally, we demonstrate the potential of such methods for a detection framework for aerial images as significantly fewer candidate windows are generated in comparison to sliding window.**

1. Introduction

Detecting vehicles in aerial images is an important processing step for applications such as traffic monitoring or screening of large areas as used for surveillance, tracking or rescue tasks. These applications share the need for accurate detection of all relevant objects, e.g. vehicles inside the camera's field of view before the scene can be analyzed and interpreted. To reduce the work load of image analysts, an automatic detection of candidate objects is required.

In general, classifiers or a cascade of classifiers within a sliding window approach are applied to detect vehicles in aerial images [14], [19]. A search window is shifted in horizontal and vertical direction across the entire image. At each window position, appearance features are calculated and a classifier returns a confidence value for the occurrence of a vehicle. However, sliding window approaches are computationally expensive due to the huge number of windows to classify, especially in case of different object scales, aspect ratios or orientations [8].

In aerial images, several approaches are proposed to overcome this challenge such as reducing the search space or applying a cascade of weak classifiers. In recent years, an alternative approach was proposed to reduce the number of candidate windows. So called object proposals methods are used to generate a set of regions that are likely to contain an object [1], [20]. Several different object proposals methods have achieved impressive results on datasets such as Pascal VOC 2007 and ImageNet for a significantly reduced number of candidate windows [8]. However, these datasets differ considerably from aerial images as images contain only one or few objects that are typically centered and occupy a high fraction of the entire image [13]. In contrast, aerial images are often larger and can contain multiple objects that are smaller and more randomly located.

In this paper, we examine the applicability of proposed methods for generating candidate windows in aerial images. Therefore, we evaluate 11 object proposals methods on the publicly available DLR 3K Munich Vehicle Aerial Image Dataset [14], [19]. The dataset consists of 20 images with a ground sampling distance of approximately 13 cm. Each image contains roughly 500 vehicles on average. We first modify the provided ground truth (GT) to enable comparison between ground truth and generated proposals. The differing characteristics of the aerial images are compensated by adapting seven methods that are most promising for generating object proposals in aerial images. Therefore, we examine different parameter settings and extensions for each method separately. The adaptations that show high impact on the performance are discussed in detail. Finally, we demonstrate the potential of such methods for a detection framework for aerial images as significantly fewer candidate windows are generated in comparison to sliding window. Thus, our contribution is threefold:

- The applicability of object proposals methods for aerial images with small objects is examined.
- We systematically analyze the impact of several parameters to adapt these methods to aerial images.
- Finally, we demonstrate the potential of object proposals methods for aerial images by comparison to baseline approaches.

The remainder of this paper is organized as follows: related work is discussed in Section 2. A detailed overview of object proposals methods is given in Section 3. The evaluation results are presented in Section 4. We conclude in Section 5.

2. Related Work

In literature, a broad variety of vehicle detection methods have been proposed. In general, vehicle detection is performed by applying classifiers or a cascade of classifiers within a sliding window algorithm. Therefore, several different feature and classifier combinations have been proposed. We limit our discussion to recently proposed approaches which aim to reduce the search space. Tuermer et al. [19] apply a sliding window approach with HOG features and SVM classifiers to find stationary and moving vehicles. To reduce the search space and consequently the number of candidate windows, they propose to incorporate road maps. Leitloff et al. [12] reduce the search space by using a road database. Vehicle detection is only performed along the roads in a certain direction. However, such approaches are limited by the availability of road maps and typically require georeferenced images [14]. Furthermore, vehicles offside the road that may of interest can be missed. Teutsch et al. [18] propose Integral Channel Features and an AdaBoost classifier to detect moving vehicles. The search space is reduced by using motion vectors. However, image sequences are required. Furthermore, static objects, e.g. parked vehicles are missed.

In recent years, several object proposals methods have been proposed to reduce the number of candidate windows. Selective Search is the mostly used method for generating object proposals in detection frameworks such as R-CNN [7] and Fast R-CNN [6]. The method greedily merges regions together based on the similarity between neighboring regions. Good detection results are achieved on Pascal VOC 2007 and ImageNet [8], [20]. However, these datasets differ significantly from aerial images as described above.

A comprehensive survey on state-of-the-art object proposals methods is provided by Hosang et al. [8]. They compare 12 different object proposals methods to baseline approaches like sliding window. Another survey is given by Chavali et al. [3]. They perform different experiments to analyze the category independence of object proposals methods. Both surveys focus on datasets that are used to explore and optimize such methods and not on datasets similar to aerial images. To the best of our knowledge, there exists no literature about the applicability of object proposals methods for significantly differing datasets such as in case of aerial images.

Method	Approach
Edge Boxes [21]	Window Scoring
Endres [4]	Grouping
GOP [10]	Grouping
LPO [11]	Grouping
MCG [2]	Grouping
Objectness [1]	Window Scoring
Rahtu [16]	Window Scoring
Randomized Prim's [15]	Grouping
Rantalankila [17]	Grouping
Rigor [9]	Grouping
Selective Search [20]	Grouping

TABLE 1. EVALUATED OBJECT PROPOSALS METHODS.

3. Object Proposals Methods

In recent years, several object proposals methods have been proposed for generating candidate windows in detection frameworks. In general, these methods can be categorized into *grouping methods* and *window scoring methods* [3], [8]. A detailed overview of the object proposals methods proposed in literature is given in the following subsections. An overview of the object proposals methods evaluated in the context of this paper is listed in Table 1.

3.1. Grouping Methods

Grouping methods are typically comprised of initial segmentation of the image followed by grouping segments. According to the applied segmentation and merging strategy, the grouping methods can be further distinguished into three types as proposed by Hosang et al. [8]. The most common approach is to simply group initial segments based on a diverse set of cues. Alternative approaches are based on solving multiple graph cut problems with diverse seeds or applying edge contours.

- Selective Search [20] is broadly used as a proposals method in detection frameworks such as R-CNN and Fast R-CNN [6], [7]. Initial segments are generated by the segmentation approach proposed by Felzenszwalb et al. [5]. Then, segments are greedily merged together based on the similarity between neighboring segments. The similarity is calculated based on simple features like colour or texture histograms.
- Randomized Prim's [15] merges segments by connecting subgraphs of a weighted connectivity graph of the initial segmentation. Edge weights representing the probability that segments are from the same object are based on similar features as in [20], whereby the feature weights are learned. Starting from a random segment, neighboring segments are added to a subgraph if the according edge weight is high until a stopping criterion is reached.
- Rantalankila [17] combines greedy merging of superpixels and solving several graph cut problems on a superpixel graph. The greedy merging is similar to [20], however, different features are used.

- Endres [4] applies an occlusion boundary algorithm that outputs four segmentations as well as the probability of occlusion and of foreground/background label for each boundary in the segmentation. Segments are used as seeds for a graph cut problem. The probabilities are used to compute the probability that an other segment belongs to the same object as the seed.
- Rigor [9] computes segments by using graph min-cuts from multiple seeds as applied in [4]. However, the computation cost is considerably reduced by reusing a single residual graph for starting the parametric min-cuts for all different seeds.
- GOP (Geodesic Object Proposals) [10] computes a boundary probability map that is used to produce superpixel segmentation. Heuristic seed placement is replaced by a learning-based approach. Foreground and background masks are generated for each seed and used as input for the geodesic distance transform. The geodesic distance transform specifies an image region that is used as object proposals.
- LPO (Learning to Propose Objects) [11] computes superpixels similar to [10]. An ensemble of binary segmentation models is trained and used to generate a diverse set of foreground and background masks. Foreground regions are used as object proposals.
- MCG (Multiscale Combinatorial Grouping) [2] performs hierarchical segmentation based on detected contours at different image scales. Multiscale segments are combined to object proposals based on edge strength.

3.2. Window Scoring Methods

Window scoring methods are based on an initial set of candidate windows. Candidate windows can be generated by a sliding window approach or random sampling. Then a score is calculated for each candidate window and used to rank or filter these windows. Several cues have been proposed to score candidate windows.

- Edge Boxes [21] uses an edge-based scoring function to score candidate windows generated by a sliding window for various scales and aspect ratios. The edge-based score measures the number of edge groups within a candidate window, whereby edge groups centered in the candidate window are weighted low.
- Objectness [1] uses different cues including multi-scale saliency, colour contrast, edge density and superpixels straddling to score each candidate windows. In case of considering multi-scale saliency (default setting) that is based on the residual of the FFT, the candidate windows are determined by the scale used for calculating the FFT. Otherwise a uniform or dense distribution is used to generate candidate windows.
- Rahtu [16] generates an initial set of candidate windows by superpixel segmentation, a prior distribution learnt from training data and randomly sampled windows. In addition to superpixels straddling as proposed in [1], three features based on boundary information and window symmetry are used to score the candidate windows.

Figure 1. Image section of the DLR 3K Munich Vehicle Aerial Image Dataset. Annotated vehicles are highlighted by green bounding boxes.

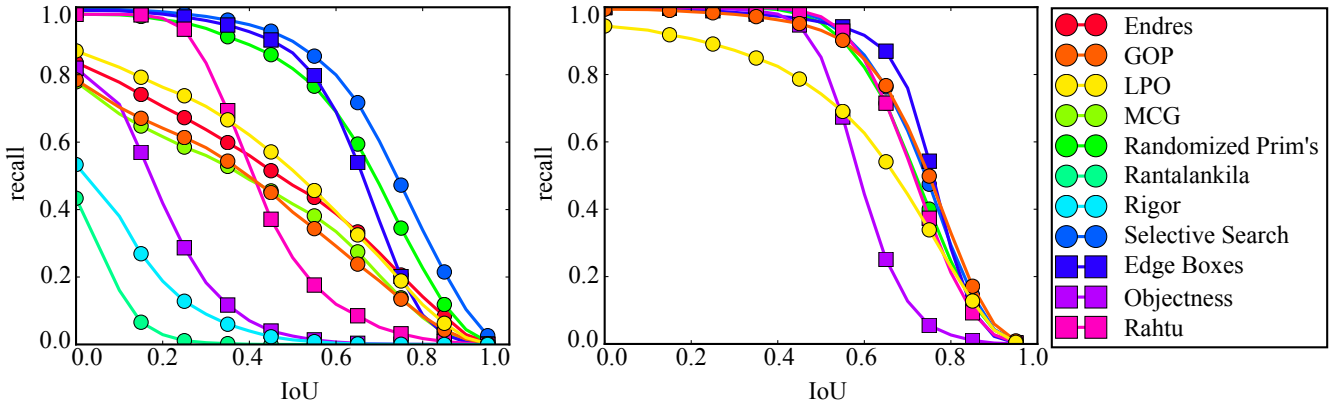
4. Experimental Results

In this section, we examine the applicability of 11 object proposals methods for generating object proposals in aerial images. Therefore, we only consider methods for which source code is publicly available. The performance of the selected methods is evaluated on the publicly available DLR 3K Munich Vehicle Aerial Image Dataset. The dataset comprises 20 images with annotated ground truth (GT). Each image has a resolution of 5616×3744 pixels and a ground sampling distance of approximately 13 cm. The annotated GT objects are classified into different vehicle classes, e.g. car, truck, bus. For our experiments, we divided each image into tiles of size 936×624 pixels and we aligned all ground truth boxes at the image edges as illustrated in Figure 1. Aligned ground truth boxes are used to enable the comparison to both grouping and window scoring methods as window scoring methods generate proposals that are aligned at image edges. The first 10 images are used as training data as training of parameters is required for several methods. The other 10 images are used for testing the performance of each method.

Metrics that are typically applied for evaluating the performance and accuracy of object proposals methods, are functions of intersection over union (IoU) between generated object proposals and GT annotations [3]. IoU (also known as Jaccard Index) is given by

$$IoU = \frac{A_{proposal} \cap A_{GT}}{A_{proposal} \cup A_{GT}}, \quad (1)$$

where $A_{proposal}$ and A_{GT} are the area of the proposed bounding boxes and the ground truth bounding boxes, respectively. For our experiments, we use two metrics: Recall as a function of IoU and recall as a function of the number



(a) Original methods

(b) Adapted methods

Figure 2. Comparison of recall values between original methods (a) and methods adapted to aerial images (b).

of proposals. GT boxes are considered as recalled, if their highest IoU value is greater than a threshold. Recall-IoU curves are generated by varying this threshold value. As different object proposals methods generate a different number of proposals, we plot the recall as a function of the number of proposals for a fixed IoU threshold. Threshold values at or above 0.5 are considered as relevant for subsequent classification [6], [7].

4.1. Original vs. Adapted Methods

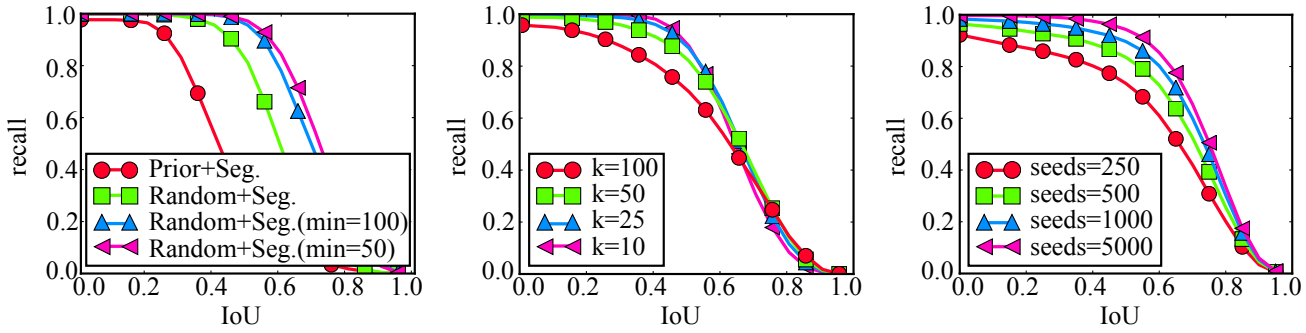
Recall-IoU curves of all original object proposals methods are shown in Figure 2 (a). Therefore, we applied the original algorithms and parameters as proposed in the corresponding literature. However, more proposals are considered in case of window scoring methods as aerial images are larger compared to images of datasets that are used to explore these methods. The best recall values are achieved for the Selective Search method. The Selective Search method exhibits recall values close to 1 for small threshold values. However, the recall value decreases with increasing threshold values. For example, the recall value for a minimal IoU of 0.5 is less than 0.9 so that more than 10% of all GT objects are not considered as recalled. The other object proposals methods exhibit worse recall values and consequently more GT objects are not considered as recalled. Hence, the original algorithms are not applicable to generate candidate windows in aerial images as too many GT objects are not recalled. Reason for this is that the methods are developed and optimized for datasets that considerably differ from aerial images as described above.

In the following, we adapted seven object proposals methods to take the characteristics of the aerial images into account. The approaches Endres, Rantalankila and Rigor are not considered as solving graph cuts with different seeds are computational expensive [4, 8]. The computational cost would increase significantly for aerial images as more seeds are required due to small objects and larger images. MCG is not considered as the computation time is higher compared to Rantalankila and Rigor [8]. To adapt the algorithms to aerial images, we systematically analyzed the impact of parameters as described below. Recall-IoU curves of the adapted methods are given in Figure 2 (b). All adapted object proposals methods exhibit considerably improved performance. The best recall values for thresholds below 0.55 are achieved for the method proposed by Rahtu that slightly outperforms Selective Search and Randomized Prim's. All methods exhibit recall values close to 1 for small threshold values except LPO. We expect that the segmentation models used for generating foreground and background masks are reason for the low recall values. As the provided segmentation models are trained on the Pascal VOC 2012 dataset that considerably differs from aerial images.

4.2. Adaptation to Aerial Images

All adapted methods show clearly improved recall values as shown in Figure 2 (b). As the fundamental functions differ for different methods, we analyzed the impact of parameters for each method separately. We examined amongst others the impact of the segmentation and its corresponding parameters, the impact of candidate sizes, the impact of applied features or cues as well as post-processing strategies such as Non-Maximum Suppression. Parameters that have been trained on Pascal VOC 2007 are trained on the first 10 images of the dataset. In the following, we focus on parameters that exhibit the most impact as a detailed survey on all examined parameters is not possible in the context of this paper.

In case of window scoring methods, the main improvement is achieved by adapting the window sizes as illustrated in Figure 3 (a) for Rahtu. The candidate windows are originally generated by superpixel segmentation and category independent window prior from training data. However, the prior distribution is not applicable for aerial images as objects can randomly occur. Instead, we apply randomly sampled windows. Further improvement is achieved by reducing the minimal size for superpixels as generated superpixels are



(a) Various initial windows (b) Varying segment sizes (c) Various number of seeds

Figure 3. Adaptation of parameters that show the most impact such as various initial windows for Rahtu (a), varying segmentation sizes for Selective Search (b) and various number of seeds for GOP (c).

used as candidate windows. The impact of the minimal size of superpixels on the window score is in contrast minimal. The recall values for Objectness are clearly improved by discarding the multi-scale saliency cue so that candidate windows are defined by a uniform or dense distribution. Adapting the minimal and maximal dimensions of candidate windows to the object sizes further improves the performance. Adapting the window sizes is also applied for Edge Boxes, whereas other adaptations show minor impact on its performance.

The performance of grouping methods is generally improved by adapting the size of the initial segments as exemplarily shown in Figure 3 (b) for Selective Search. The size of initial segments is adjustable by varying the threshold parameter k as described in [5]. Smaller values for k result in smaller segments. Further improvement is achieved by reducing the minimal dimensions for accepting segments as proposals. Randomized Prim's is based on the same segmentation approach. Thus, similar adaptation results in improved performance. The recall values are further improved by increasing the approximated number of final proposals as more subgraphs and consequently more proposals are generated. The recall values of LPO are mainly improved by adapting the initial segmentation as well. We increase the number of computed superpixels by an order of magnitude, whereas the impact of further adjustable parameters is minimal.

The main improvement for GOP is achieved by increasing the number of seeds used to generate foreground and background masks as shown in Figure 3 (c). More seeds result in better recall values as the chance that objects are covered by seeds increases. However, computation costs increase significantly with the number of seeds. Hence, GOP is not suited for aerial images with small objects.

The impact of parameters that are learnt on the first 10 images is small compared to the impact of candidate sizes and segmentation, respectively.

4.3. Comparison to Sliding Window

Object proposals aim to reduce the number of candidate windows to classify. To show how the number of candidate windows is reduced, we compare the adapted Selective Search and the adapted Rahtu method to a baseline approach. Therefore, we consider a sliding window approach. We use three different window scales specified by the mean window dimensions of the training data to consider different object sizes and aspect ratios. Figure 4 shows the corresponding recall values as a function of the number of proposals for a fixed IoU threshold of 0.5. In case of the sliding window approach, we adapted the step size based on the actual number of windows.

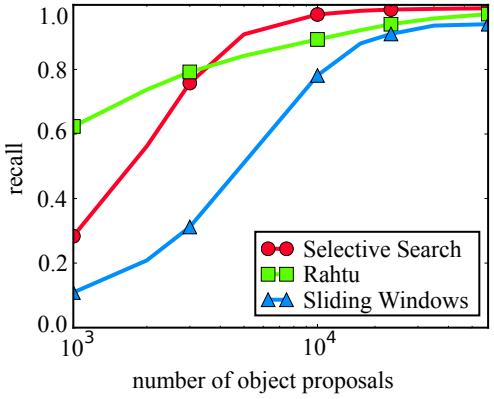


Figure 4. Selective Search and Rahtu show considerably improved recall values with respect to the number of candidate windows.

Both object proposals methods clearly outperform sliding window for all numbers of proposals. Rahtu exhibits high recall values for a small number of proposals as proposals that are more likely to contain an object are better ranked. Selective Search achieves recall values close to 1 for comparable fewer proposals. Reason for that is the better localization accuracy of proposals generated by grouping methods [8]. In case of the sliding window approach, more than 5% of all GT objects are not considered as recalled even for 50,000 candidate windows. Hence, more different scales and aspect ratios and consequently a higher number of candidate windows are necessary to achieve recall values close to 1, even though the used scales are adapted to the mean dimensions of the objects in the training data.

5. Conclusion

In this paper, we have examined the applicability of different object proposals methods for aerial images. We have shown that the original methods are not applicable as these methods are developed and optimized for datasets that considerably differ from aerial images. The performance of seven object proposals methods was significantly improved by adapting these methods to the characteristics of the aerial images. Therefore, we examined different parameter settings and extensions for each method separately. The adaptations that show high impact on the performance were discussed for each method. Selective Search, Rahtu and Randomized Prim's seem to be suited for aerial images as these methods achieve recall values close to 1 for an IoU threshold of 0.5 so that almost all objects are recalled. Finally, we have demonstrated the potential of applying object proposals methods in a detection framework for aerial images as significantly fewer candidate windows are generated in comparison to sliding window. In future work, we will integrate the adapted object proposals methods into a detection framework to further evaluate the quality of generated object proposals. Based on the results, we will implement an object proposals method specifically for aerial images that takes the advantages of different object proposals methods into account.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 34(11):21892202, 2012.

[2] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 328335, 2014

[3] N. Chavali, H. Agrawal, A. Mahendru, and D. Batra. Objectproposal evaluation protocol is gameable. *arXiv preprint arXiv:1505.05836*, 2015.

[4] I. Endres and D. Hoiem. Category independent object proposals. *European Conference on Computer Vision*, pages 575588. Springer, 2010.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167181, 2004.

[6] R. Girshick. Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, pages 14401448, 2015.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580587, 2014.

[8] J. Hosang, R. Benenson, P. Doll ar, and B. Schiele. What makes for effective detection proposals? IEEE transactions on pattern analysis and machine intelligence, 38(4):814830, 2016.

[9] A. Humayun, F. Li, and J. M. Rehg. Rigor: Reusing inference in graph cuts for generating object regions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336343, 2014.

[10] P. Krähenbühl and V. Koltun. Geodesic object proposals. *European Conference on Computer Vision*, pages 725739. Springer, 2014.

[11] P. Krähenbühl and V. Koltun. Learning to propose objects. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15741582. IEEE, 2015.

[12] J. Leitloff, D. Rosenbaum, F. Kurz, O. Meynberg, and P. Reinartz. An operational system for estimating road traffic information from aerial images. *Remote Sensing*, vol. 6, no. 11, pp. 11315-11341, 2014.

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

[14] K. Liu and G. Mattyus. Fast multiclass vehicle detection on aerial images. *Geoscience and Remote Sensing Letters, IEEE*, PP(99):15, 2015.

[15] S. Manen, M. Guillaumin, and L. Van Gool. Prime object proposals with randomized prim's algorithm. *Proceedings of the IEEE International Conference on Computer Vision*, pages 25362543, 2013.

[16] E. Rahtu, J. Kannala, and M. Blaschko. Learning a category independent object detection cascade. *Proceedings of the IEEE International Conference on Computer Vision*, pages 10521059. IEEE, 2011.

[17] P. Rantalankila, J. Kannala, and E. Rahtu. Generating object segmentation proposals using global and local search. *In CVPR*, pages 24172424, 2014.

[18] M. Teutsch and W. Kruger. Robust and fast detection of moving vehicles in aerial videos using sliding windows. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2634, 2015.

[19] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla. 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):23272337, 2013.

[20] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154171, 2013.

[21] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. *European Conference on Computer Vision*, pages 391405. Springer, 2014.