

A WEAKLY-SUPERVISED DEEP NETWORK FOR DSM-AIDED VEHICLE DETECTION

Xin Wu^{1,2}, Danfeng Hong^{3,4}, Jiaojiao Tian³, Ralph Kieff⁵, Ran Tao^{1,2}

¹School of Information and Electronics, Beijing Institute of Technology (BIT), Beijing, China

²Beijing Key Laboratory of Fractional Signals and Systems, Beijing Institute of Technology (BIT), Beijing, China

³Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

⁴Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany

⁵German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Wessling, Germany .

ABSTRACT

With the breakthrough of the spatial resolution of optical remote sensing images at the sub-meter level and the explosive development of deep learning, geospatial object detection has achieved a growing interest in remote sensing community. However, labeling large training datasets in object level is still an expensive and tedious procedure. This might lead to the poor model generalization and degraded network learning ability. To this end, a weakly-supervised deep network (WSDN) is developed for geospatial object detection by applying a digital surface model (DSM)-aided auto-labeling and a pre-trained network learned from the task-independent dataset. Experimental results conducted on the stereo aerial imagery of a large camping site are performed to demonstrate that the proposed WSDN yields better detection results, with 62.78% precision and 55.13% recall.

Index Terms— Deep learning, digital surface model, geospatial object detection, optical remote sensing imagery, vehicle, weakly-supervised

1. INTRODUCTION

Recently, optical remote sensing imagery (RSI) has been paid a growing interest in many applications, such as urban mapping and monitoring [1], mineral exploration [2, 3], particularly spatial object detection [4]. Existing detection methods can be roughly categorized as follows [5]: *template matching-based*, *knowledge-based*, *object-based*, and *machine learning-based* methods. The explosive development in deep learning have made them unsurprisingly applied to object detection in RSIs [6] and have shown a stronger detection performance than the aforementioned traditional methods. Labeling training datasets [7, 8] plays an important role in object detection in optical RSIs. In addition, for objects in optical RSIs with a cluttered background, relatively small ground sampling distance (GSD), and various deformations, e.g. variabilities in viewpoint, scaling, and direction, their labeling problems have always been challenging and existing manual labeling is not only time-consuming and laborious but also inconsistent



(a) Target vehicles samples

(b) DOTA vehicles samples

Fig. 1. Some visual vehicle examples from the two datasets.

in labeling quality. It is urgent to find an efficient automatic labeling method.

This work proposes a weakly-supervised deep network (WSDN) to achieve auto-tagging and detection of vehicles objects on unlabeled target dataset (see Fig. 1(a)). Two main technical contributions of this paper are: 1) a newly developed unsupervised vehicles extraction method using aerial stereo imagery, which consists of image segmentation, e.g., superpixels, clustering and similarity measure; 2) a weakly-supervised deep network has been fine-tuned by a training set made up of three parts of the vehicles samples, namely vehicle objects in DOTA dataset [9] (see Fig. 1(b)), vehicle objects in target dataset detected by a pre-trained network learned from DOTA dataset, and a digital surface model (DSM)-aided auto-labeling vehicles objects in the target dataset.

2. METHODOLOGY

The purpose of this work is to develop a WSDN by preparing training samples automatically. The detailed framework of the WSDN is illustrated in Fig. 2, which consists of two phases, namely auto-tagging label generation (red dotted box), and detector generation (blue dotted box). The details of each phase are discussed in the following sections.

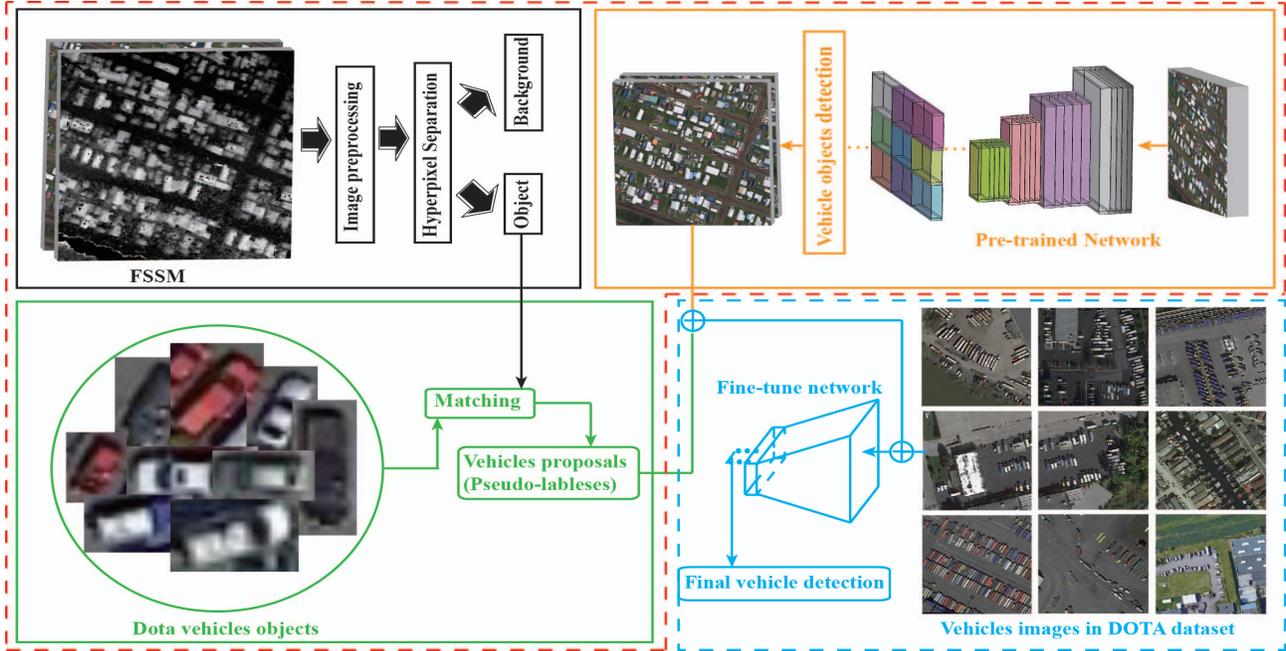


Fig. 2. The flow chart of weakly-supervised deep network for vehicle detection.

2.1. Auto-tagging Label Generation

In this section, we focus mainly on generating auto-tagging label for vehicles objects in the target dataset. To this end, a WSDN using a two-step approach that fully takes into account the similarity of the same class in different scenes and the similarity of different class objects in the same scene is proposed, in order to obtain accurate labels of vehicles in the target dataset.

Step 1. Auto-tagging label from pre-trained network: DOTA is an aerial image dataset presented in [9]. Until now it is one of the largest open-sourced tagged datasets in the field of remote sensing image object detection. Figure 1 shows one example image from DOTA and a subset from the target dataset. It shows that the target dataset has a more complex background and more types of the vehicle than DOTA, including various vehicles. If all vehicles are classified as one class, the intra-class gap is wider than in DOTA, but the inter-class gap, e.g., the white transporter and some white rectangular tents, is narrowed. Using pre-trained network from DOTA alone would be not sufficient to achieve better detection performance. Subsequently, we develop an effective unsupervised vehicle extraction approach by applying a digital surface model (DSM)-aided auto-labeling.

Step 2. Auto-tagging label from DSM: To process the target dataset with large intra-class differences and obtain the effective labels of vehicle samples, we adopt the superpixels-based image segmentation method, which consists of SLIC [10], DBSCAN clustering [11], and the similarity measure of vehicles objects. Fig 3 shows the flowchart of the auto-

tagging label generation. More specifically, they are

1) *Height information embedding:* Due to limitations in view angles, image features, such as spectrum, shape, texture, and context, can't describe the real objects in the target dataset completely. Considering that vehicles are significantly higher than the ground, the available DSM can provide very useful height information to separate vehicles from background. The DSM we used has been generated with the same approach as described in [12]. In addition, spectral information and elevation information will be fused and encoded for the separation of objects and the ground, as well as the subsequent acquisition of object location information.

2) *Image segmentation:* Existing semi-automatic or fully automatic object extraction methods are based on pixels, which can be roughly categorized by *region-based*, *line & corner-based* methods and other variants. In this part, we implement the rigid structure of the image instead of the segmentation of the pixel grid to achieve dense segmentation of the object. The main highlights of our work are fourfold.

Step 1: SLIC superpixels. Compared with the pixel-wise matching [13, 14], superpixels yield an object-based similarity measurement. SLIC iteratively updates the distance between the two cluster centers D with a certain threshold.

Step 2: DBSCAN Clustering. DBSCAN describes the proximity of samples based on the adjacency matrix generated by the SLIC. The connected samples by the maximum density derived from the density-reachable relationship is a category of our final cluster or a cluster lc .

Step 3: Object Refining. To effectively eliminate the influence of trees on object separation in the target dataset, we

use DSM-aided information again to keep the object height information below a certain threshold and define the value of each cluster as its average elevation information. Morphological profiles [15] are used to smooth the edges of the object.

Step 4: Vehicle Labels Generation. The labels used in the final network training are automatically generated with an Euclidean-based measurement between real vehicle instances from DOTA dataset and segmented objects from our dataset.

2.2. Detector Generation

A new training set consisting of the data corresponding to the two-part auto-tagging labels of vehicle samples in the target dataset generated in the previous section. In the proposed framework, the original DOTA vehicle samples, including large-vehicle and small-vehicle samples. Due to the similarity of the two vehicle classes, they are migrated to the new training set and merged into a class named as vehicle. We fine tune the pre-trained Region-based Fully Convolutional Network (R-FCN) [16] generated by DOTA dataset with 15 class objects. The backbone of this network is 101-layer Residual Net (ResNet-101) [17]. The simple steps are marked in blue, as shown in Figure 2. We have to emphatically clear that the motivation and goal of this paper apply existing labels from the task-independent dataset to achieve auto-tagging and detection of objects on target dataset, rather than greatly enhancing feature representation capabilities. Therefore, the F-RCN has been only selected as an example, the proposed WSDN could be easily adopted with other networks [18, 19].

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Aerial Imagery

This dataset used in the paper is a subset from the aerial imagery of a camping site in northern Germany [12]. It was acquired by the optical 4K camera system on the German Aerospace Center (DLR) research helicopter BO 105 (DLR (CC-BY 3.0)) at a height of 600 m above the ground. This test sites comprised an area of $1.0km \times 1.5km$. With three cameras on board, the 4K camera system is able to capture the multi-view imagery with 90% overlap along-track and 60% overlap across the track. The DSMs were then prepared using the automatic processing chain as introduced in [12].

3.2. Experimental Result

Table 1 lists the quantitative results using F-RCN trained by DOTA dataset and F-RCN fine-tuned by the target dataset with an auto-tagging label. It shows that the proposed framework yield the higher precision and recall than original F-RCN. In the experiment, although several vehicle samples in the target dataset are included in the DOTA, the pre-trained network generated by the DOTA is not robust on the target dataset due to the background of the vehicle and the semantic

information surrounding it are different. Furthermore, the RV mixes the appearance of tents and vehicles, which makes it more difficult to generate labels. In order to reduce the difficulty of auto-tagging, all types of vehicles are marked as one class, so the network finetuning is a binary classification.

Table 1. Performance comparisons of two different methods. The best result is shown in bold.

Label	Method	Precision	Recall
Auto-tagging labels	F-RCN	40.58	38.12
	WSDN	62.78	55.13

Visually, very few living tents are wrongly identified as vehicles, those with low detection scores can be removed by the pull-up threshold. There are some miss detections in cars and RV, also some false detections in white transport vehicles and rectangular tents, as shown in Fig. 4. This might be introduced by the inaccuracy of the threshold of DBSCAN clustering and the height threshold of object refinement. In addition, ambiguous edges generated in overcrowded objects can mislead the labels generation.

4. CONCLUSION AND OUTLOOK

In this work, a WSDN trained by an auto-tagging training set is proposed, which tends to use labeled task-independent dataset to explore the detection performance of an unlabeled dataset under the same network. With this deep network, especially the embedding of DSM-aided training data generation approach, it is possible to detect objects with different background complexity. Auto-tagging labels of vehicles objects in target datasets can be robustly generated using SLIC superpixels, DBSCAN clustering and similarity measure.

In the future work, we will focus on further improving the accuracy of auto-tagging labels by subdividing vehicle samples (e.g., Car, Transporter, RV, Camping trailer) and extending the proposed framework to other networks or introducing the multimodal data [20] to improve the feature representation capabilities of the objects.

5. REFERENCES

- [1] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. Zhu, "Learnable manifold alignment (leMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, 2019.
- [2] D. Hong and X. Zhu, "SULoRA: Subspace unmixing with low-rank attribute embedding for hyperspectral data analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1351–1363, 2018.
- [3] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, 2019.
- [4] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm Detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, 2019.

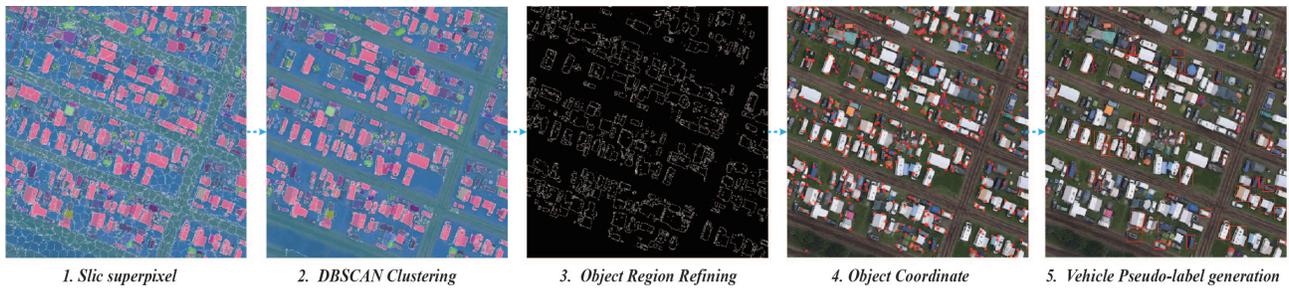


Fig. 3. The flowchart of the auto-tagging vehicle label generation.

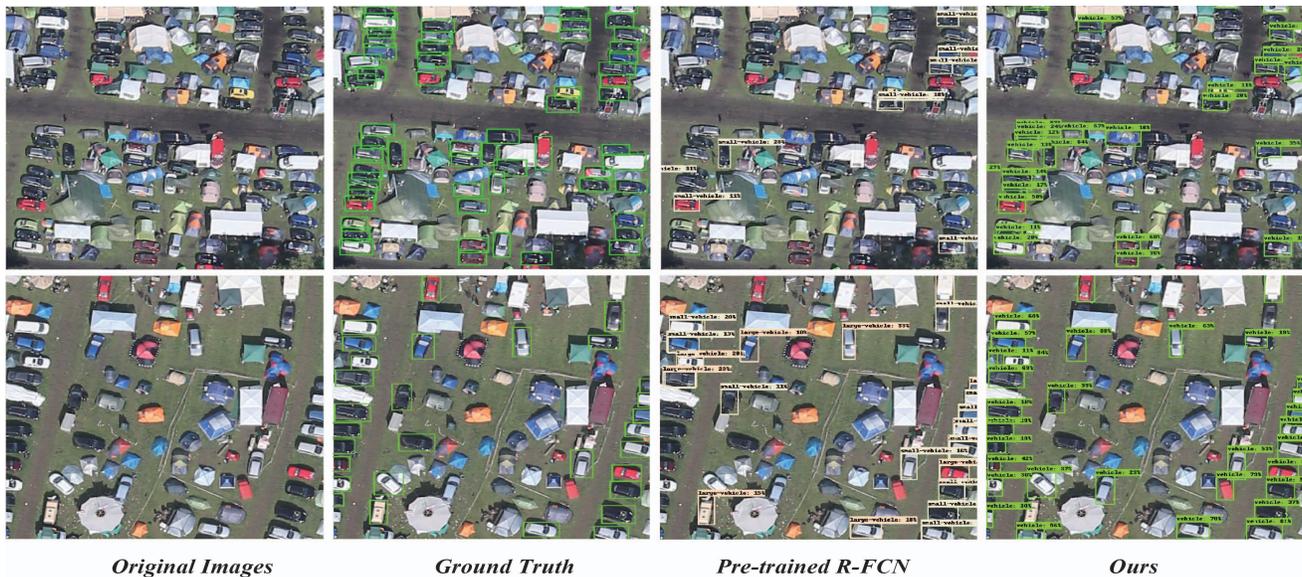


Fig. 4. Some visual vehicle results by using the pre-trained R-FCN and the proposed method on the target dataset.

[5] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.

[6] X. Wu, D. Hong, P. Ghamisi, W. Li, and R. Tao, "MsRi-CCF: Multi-scale and rotation-insensitive convolutional channel features for geospatial object detection," *Remote Sens.*, vol. 10, no. 12, pp. 1990, 2018.

[7] D. Hong, N. Yokoya, and X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, 2017.

[8] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint & progressive learning from high-dimensional data for multi-label classification," in *Proc. ECCV*, 2018, pp. 478–493.

[9] G. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proc. CVPR*, 2018.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[11] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, vol. 96, pp. 226–231.

[12] V. Gstaiger, J. Tian, R. Kiefl, and F. Kurz., "2d vs. 3d change detection using aerial imagery to support crisis management of large-scale events," *Remote Sens.*, vol. 10, no. 12, pp. 2054, 2018.

[13] D. Hong, W. Liu, J. Su, Z. Pan, and G. Wang, "A novel hierarchical approach for multispectral palmprint recognition," *Neurocomputing*, vol. 151, pp. 511–521, 2015.

[14] D. Hong, W. Liu, X. Wu, Z. Pan, and J. Su, "Robust palmprint recognition based on the fast variation veese–osher model," *Neurocomputing*, vol. 174, pp. 999–1012, 2016.

[15] J. Hu, D. Hong, Y. Wang, and X. Zhu, "A comparative review of manifold learning techniques for hyperspectral and polarimetric sar image fusion," *Remote Sens.*, vol. 11, no. 6, pp. 681, 2019.

[16] J. Dai, L. Yi, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Proc. NIPS*, 2016, pp. 379–387.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[18] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2019.

[19] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "Stfnet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Trans. Geosci. Remote Sens.*, 2019.

[20] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "Cospace: Common subspace learning from hyperspectral–multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, 2019.