2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Weakly Supervised Vehicle Detection in Satellite Images via Multiple Instance Ranking

Yihan Sheng, Liujuan Cao*, Cheng Wang, Jonathan Li

Fujian Key Laboratory of Sensing and Computing for Smart City
School of Information Science and Engineering, Xiamen University, 361005, P. R. China
Emails:{caoliujuan,cwang,junli}@xmu.edu.cn

Abstract*—Given the difficulty in labeling sufficient amount of instances across different resolutions and imaging environment of satellite images, weakly supervised vehicle detection is with great importance for satellite images analysis and processing. To prevent such cumbersome and meticulous manual annotation, naturally we have introduced the weakly supervised detection that has recently explosively prevalent in ordinary viewing angle images. Our program merely stands in need of region-level group annotation, *i.e.*, whether this district convers vehicle(s) without plainly pointing out the coordinates of vehicles. There are two major problems are often encountered for Weakly Supervised Object Detection. One is that it is often chooses only a most expressive instance contains multiple target objects which often have a bigger probability when selecting a target block. For this problem, the number of vehicles can be estimated based on the object counting, a combinatorial selection algorithm can be used to select patch which contains at most one vehicle instance. Another problem is that precise object positioning becomes more difficult due to the lack of instance-level supervision. This problem can be optimized by a progressive learning strategy. Experiments was carried on wide-ranging remote sensing dataset and achieved better results compared to the state-of-the-art weakly supervised vehicle detection schemes.

I. Introduction

Satellite image based traffic monitoring of densely populated areas has become a research hot spot, which is especially suitable for dense urban road networks in wide areas. For instance, satellite based optical sensor like Ikonos and QuickBird nowadays can provide images with 1-meter resolution or better. Consequently, vehicle detection in satellite images has attracted ever-increasing research attention [1], [2], where export detectors from instance-level supervision, in which the common form is like a vehicle rectangles or segments, which is used explicitly [3], [4] or implicitly [5], [6] at the stage of model abstraction. The former clusters similar pixels into potential vehicles in a top-down matching manner. The latter extracts intensity or texture features surrounding the training rectangle to fit into a probabilistic model for vehicle detection. The need of extensive labels has restricted its potential applications. We are committed to the detection of vehicles using only region-level labels indicating the existence of vehicles, *i.e.*, a much easier alternative solution to label instance segments or rectangles. A multiple-instance rank learning scheme is then applied to refine and purify an instance-level classifier from these group-level weak labels. Through the probability ranking that patch to be a vehicle transformed by softmax to select the instance. Such region-level annotation allows us to get the count of objects, which is very important for the subsequent ranking. Therefore, progressive learning is carried out from these weak annatations, and finally engenders the instance-level object detectors. Figure. 1 illustrates the skeleton.

In this paper, we propose a Mutil-Instance Ranking network termed MIRN to achieve the above goals. In the offline training, following the standard Bag-of-Words approach, dense SIFT [7] or random forest features [8] are firstly extracted and then were quantized to a codebook via K-Means clustering. The Principle Component Analysis was then adopted to reduce the feature dimension. Subsequently, as inspired by [9], density map is estimated to assess the presence of vehicles. Within a given region, this map is designed to minimize the error between the ground-truth densities inferred from the user annotations and the estimated density inferred from the potential objects. Such a rough counting is latter delivered to the MIRN to gain a subtle estimation of the substantive object location. Our basic design follows the principle of fast-rcnn [10], with the difference that in weakly supervised learning we do not know the label of the instance. Therefore, we further introduce a region-level supervision and use it to get object counting. By using the probability ranking generated by softmax, we are able to get the possibility that the block belongs to the vehicle. In pose-processing, we take a progressive approach throughout the process. Fig. 1 illustrates the proposed Mutil-Instance Ranking network (MIRN) framework. The main contributions of our method are as follows.

- A user-coordinated, labor-saving strategy to uncomplicatedly aggregate ample vehicle annotations with variegated sensors and imaging conditions.
- The object counting method is used to assist the box selection in weakly supervised vehicle detection, *i.e.* one selected object proposals are expected to at most contain one unique vehicle object.
- A deep Mutil-Instance Ranking network termed MIRN to progressively learn vehicle detector, by giving the above density map estimations.

The rest of this paper is organized as follows. Section II reviews related work. Our proposed method is introduced in

*Corresponding author E-mail: caoliujuan@xmu.edu.cn

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

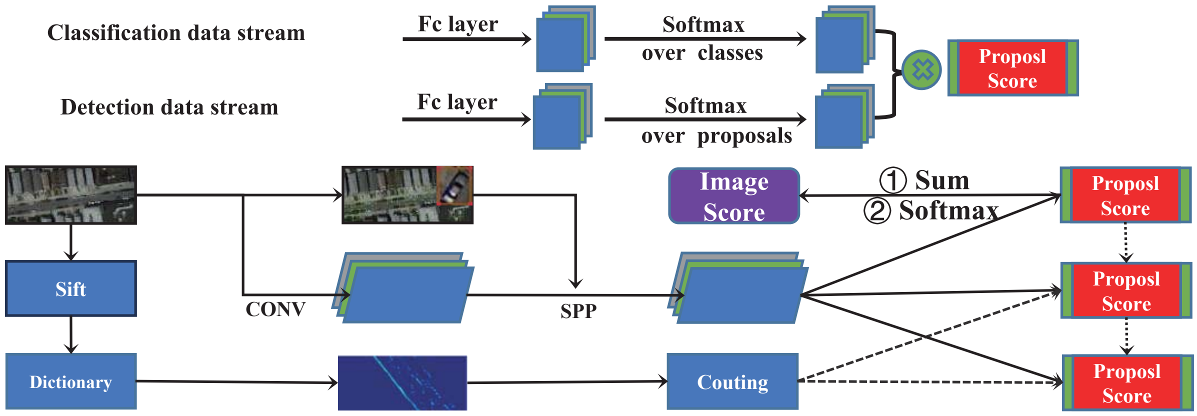


Fig. 1: The framework of the thought out multi-Instance based vehicle detection in satellite images. The key innovation lies in the usage of Multi-Instance Ranking for combinatorial region selection under the constraints of the object counting, as described in details in Sec. III.

Section III. We present the detailed evaluations in Section IV. And more detailed assessment and analysis are given in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

Vehicle detection and classification are among the core tasks of remote sensing image analysis. Chen *et al.* [11] employed hybrid deep neural networks to extract multi-scale features for scale-invariant vehicle detection in high-resolution remote sensing images. Zhao *et al.* [12] proposed a multi-level neural network structure, in which the spatial pyramid was used to capture mutil-scale spatial features for remote sensing image classification. The above methods demonstrate the strong capability of deep features for such tasks.

Multi-instance learning has been widely studied for decades. In this case, only the label of the collection(bag) is known while the labels of the individuals(instances) are unknown. The sign of a bag often depends on the bounds of the most positive pattern. If there exists one element in the set that is positive, the bag is positive, otherwise is negative [13]. The difficulty is to infer the instance-level label only using the bag-level supervision. In this case, iterative methods like mi-SVM are often adopted [14]. In addition, deep features with good representation power have also been introduced. More recently, deep model is adopted as a feature extractor [15], [16] or to train an end-to-end classifier [17], [18]. The latter strategy suits for our application scenario better, since our purpose is to find all positive patterns rather than the most positive instance. Therefore, max pooling can be skipped which only takes the most notable instance into account [13], [19]. Our task-related deep multi-instance detection model is proposed to detect multiple vehicles in the satellite images, which adopted density estimation for vehicle counting. Our inspiration partly stems from [20]. The difference is that we have adopted deep features with stronger capability. We have not adopted the widely popular form where alternately update models and labels. Instead, we adopted an online purification method similar to that used in the OICR [18]. In addition, compared to [20], we maximize the benefits of object counting. The former did not use object counting to assist the selection related to probability of target blocks. This kind of auxiliary selection helps to solve one of the most difficult problems in weakly supervised detection *i.e.* data imbalance. It is worth to note that, the classifier refinement can be further integrated into the deep network to fine-tune the model.

III. THE PROPOSED APPROACH

A. Initial Instance Classifier

In MIL, the localization of objects in image means learning from bags(regions) to classify instances(vehicle rectangles). Typically, one of the biggest challenges for multi-instance learning is to learn instance-level detector using only bag-level label. We achieve this by ranking the candidate patches and selecting the most potential ones. As shown in the Mutil-Instance Ranking block in Fig. 1, features of image are bifurcated into two streams to produce two matrices: $f^c, f^d \in R^{C\times|R|}$ by two fully-connected layers, where C indicates whether the vehicle exists in the image and $|R|$ denotes the number of proposals. Then the two matrices are assigned to different softax layer: $[\sigma(f^c)]_{ij} = \frac{e^{f^c_{ij}}}{\sum_{k=1}^{|C|} e^{f^c_{kj}}}$ and $[\sigma(f^d)]_{ij} = \frac{e^{f^d_{ij}}}{\sum_{k=1}^{|R|} e^{f^d_{ik}}}$. The former, softmax operator compares class scores for each region separately, while the latter softmax operator analyzes the scores of different regions for each class independently. Element-wise multiplication $f^R = \sigma(f^c) \odot \sigma(f^d)$ between these two matrices will produce the final proposal scores. Finally, image score of the c-th class Φ_c can be obtained by summing over all proposals, followed by a softmax normalization: $\Phi_c = \frac{\sum_{r=1}^{|R|} f^R_{cr}}{\sum_{k=1}^{C}\sum_{r=1}^{|R|} f^R_{kr}}$. It is worth to note that softmax is operated at this stage, as images certainly contain vehicle. Therefore, for each image, we increase the probability for the region that involves the vehicle, and meanwhile reduce the probability for the region that does not contain the vehicle. Assume that the image label is y_c, a collection of images

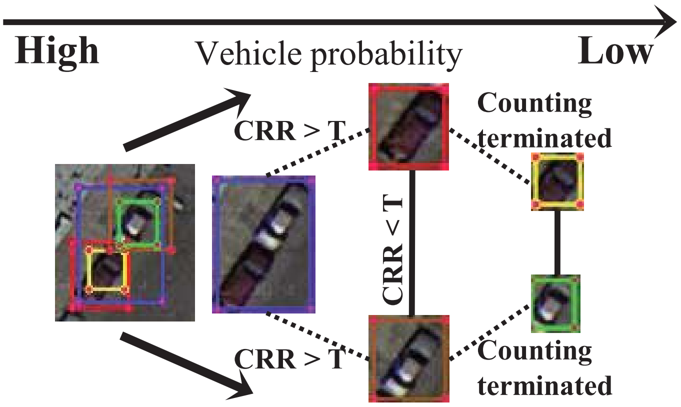


Fig. 2: To avoid the undesired case that chooses only a most expressive instance in weakly supervised detection, we apply a subset selection algorithm by aggregating the count prior to transform this region selection problem into the largest summing of constrained fixed-length subarray slice. This restrictive patch selection will tend to hold the uniqueness of the vehicle in a patch.

$x_i, i = 1, ...n$, y_c always equals to 1 in our task. A standard multi-class cross-entropy loss is adopted to train the initial instance classifier, as shown in Eq. 1,

$$\mathrm{L}_B = -\sum_{i=1}^{n}\left\{\sum_{c=1}^{C} y_{ic}\log\Phi_c + (1-y_{ic})log(1-\Phi_c)\right\}, \quad (1)$$

which can be simplified to $\mathrm{L}_B = \sum_{i=1}^{n}\sum_{c=1}^{2}\log\Phi_c$.

B. One-to-One Region Selection

In OICR [21], only one prime instance is selected, which occasionally contains multiple objects and has become a crucial problem, especially when lots of objects are close to each other [22].

To solve this region selection problem, we combine counting prior to selection strategies. Given a set of patches $\mathrm{Q} = \left\{\mathrm{Q}_1, \mathrm{Q}_2, ..., \mathrm{Q}_{|R|}\right\}$, we choose a subset Q^* with $|Q^*| = \mathrm{m}$ and m indicates the vehicle counting in each image. The probability that patch is vehicle is as large as possible, and each block can contain at most one vehicle. Patch is reserved if the spatial overlaping between it and the selected patch is below a threshold T, (as shown in Fig. 2). Then, our goal of combinational selection can be formalized as follows:

$$\begin{aligned} &\mathrm{Q}^* = \arg\max_{Q} \sum_{q_i\in Q} f_i, \\ &s.t. |\mathrm{Q}| = \mathrm{m}, CRR(q_i, q_j) < T \quad \forall q_i, q_j \in Q, i \neq j. \end{aligned} \quad (2)$$

where $CRR(q_i, q_j) = \frac{area(q_i \cap q_j)}{area(q_i)}$ is the cross-regional rate which is inclined to encourage selecting regions that contain at most one object, q_i is a patch previously selected by the greedy algorithm and q_j is a patch that is considered for selection, T is the overlapping threshold. To grab a high score of total detection scores, the algorithm tends to choose those sub-high scoring blocks based on the highest scoring block, which usually contains multiple vehicles, while lowest scoring patch contains only a small portion of the vehicle. In contrast, sub-high scoring blocks contain exactly one vehicle with the large portion of object region. More details can be found in Alg. 1

Algorithm 1 One-to-One Region Selection

```
Input: object counting $m$, area overlap threshold $T$,
object proposal R, proposal score matrices f
Output: selected proposal
Initialization: Sort(descend) R based on f by Eq. 3;
$Q^* \leftarrow \emptyset; f_{total} \leftarrow 0$;
for $i = 1 \rightarrow |R|$ do
    $\mathrm{Q} \leftarrow Q_i; F \leftarrow f_i$;
    for $j = i + 1 \rightarrow |R|$ do
        if $CRR(q_i, q_j) < T(\forall q_i \in Q)$ then
            $\mathrm{Q} \leftarrow Q \cup \{Q_i\}; F \leftarrow F + f_i$;
            if $|Q| == m \ or \ i == |\mathrm{R}|$ then
                if $F > f_{total}$ then
                    $f_{total} \leftarrow F; Q^* \leftarrow Q$
                end
                break
            end
        end
    end
end
```

C. Weakly Supervised Instance Classifier

Given the base instance classifier, since there is no ground-truth location information, we are unable to deploy bounding-box regression like faster-rcnn [10] in supervised detection. Very likely, neighbor regions with high overlaps with the selected regions will also have high scores. To find the optional surrounding region with vehicle, we form it as a bounding-box regression problem. In particular, we select multiple instances according to the spatial relation, Our inspiration also comes from the classifier updating procedure in OICR [21]. which are subsequently integrated into the basic MIRN to refine the corresponding classifiers in an end-to-end manner.

The key challenge here is how to learn the instance classifier given only region-level labels. To deal with this problem, we propose an online labelling and refinement strategy. Suppose the label vector for proposal j is $[y_{1j}^k, y_{2j}^k]^T \in R^2$, where K is the total refinement times that empirically set as 2 for fine-tunning. Suppose an image has a class label c, we first select proposal $\{j_{1c}^{k-1}, j_{2c}^{k-1}, ..., j_{mc}^{k-1}\}$ for $\{k-1\}^{th}$ times as in Alg. 1, and label it to class c. As proposals with high overlaps probably belong to the same class, we label the proposal j_{ic}^{k-1} and its adjacent proposals to class c in the $k-$th round refinement. To that effect, if proposal j has a high overlap with one of the proposals in $\{j_{1c}^{k-1}, j_{2c}^{k-1}, ..., j_{mc}^{k-1}\}$, we label proposal j to class c ($y_{cj}^k = 1$), otherwise we label proposal j to the background ($y_{(c+1)j}^k = 1$). More specially, we label proposal j to class c if the IoU between proposal j and one of the proposals in $\{j_{1c}^{k-1}, j_{2c}^{k-1}, ..., j_{mc}^{k-1}\}$ is larger than an empirical threshold I_t. Meanwhile, if there is no object

c in the image, we set $y_{cj}^{k} = 0$ for all proposals.

$$\left\{ \mathrm{Q'}_1, Q'_2, ..., Q'_{|R|} \right\} = \arg sort \left(\left\{ -f_1, -f_2, ..., -f_{|R|} \right\} \right). \tag{3}$$

Use the above assumptions, we train the refined classifier based on the following loss function:

$$L_r^k = -\frac{1}{|R|} \sum_{r=1}^{|R|} \sum_{c=1}^{2} w_r^k \, y_{cr}^{\,k} \log f_{cr}^{Rk}, \tag{4}$$

where w_r^k is the probability of instance being positive. Note that the value of w_r^k can be acquired by Alg. 2.

The above process has been summarized in Alg. 2 for better explaining the proposed MIRN algorithm. Note that, our overall network loss is designed by combining Eq. 1 and Eq. 4, into Eq. 5

$$\mathrm{L} = L_B + \sum_{k=1}^{K} L_r^k. \tag{5}$$

IV. Experiments

A. Datasets

We assess our algorithm on a wide-ranging satellite image dataset, which is composed of 80 images (979×1348) sampled from the City of New York with three (0.1m, 0.3m, 0.5m) spatial resolution levels on Google Earth. A total of 1,482 vehicle annotations are collected, which are contributed by volunteers and contain variances in imaging conditions, visual appearances, and camera perspective. In the course of the experiment, the base convolutional network we adopted is a modified VGG16 [23]. In particular, the SPP layer was arranged to be in harmony with the 1st fully-connected layer. Then, a parallel detection tributary that contains a fully connected component and a softmax component is added to the classification layer. After that, we perform element-wise product between the classification streams and the detection streams, followed by a score summing across regions. Lastly, the softmax classifier is adopted to get the class score for a given region, which are subsequently feed to a binary log-loss layer.

B. Alternative Approaches

We compare the proposed method to a series of alternative approaches, ranging from multiple instance learning to weakly supervised object detection. We list as below:

- sMIL: The sparse MIL which performs multiple instance learning by sparse coding using both set and instance kernels [24].
- stMIL: The sparse transductive MIL adds one constraint to sMIL to constrains instances within bags to be outside the margin.
- sbMIL: The Sparse balanced MIL approach seek out a balancing parameter which is very similar to the object counting indicating the fraction of positive pattern in positive bags.
- $\mathrm{MI-SVM}^{\dagger}$: A variant of MI-SVM [14] that incorporates object counting prior, The $\mathrm{MI-SVM}^{\dagger}$ no longer chooses only a positive example like MI-SVM. Instead, multiple witness or prime instance from bags are chosen with the assistance of counting prior.
- mi-SVM: An iterative scheme that employs a "train-label-retrain scheme" until label of each instance is stable.
- $\mathrm{WSDDN}^{\dagger}$: It performs simultaneously region selection and classification with weak superiversion. It uses pre-trained CNNs in a unified convolutional network [17]. The counting prior was used in the step of region selection to further improve the efficiency.

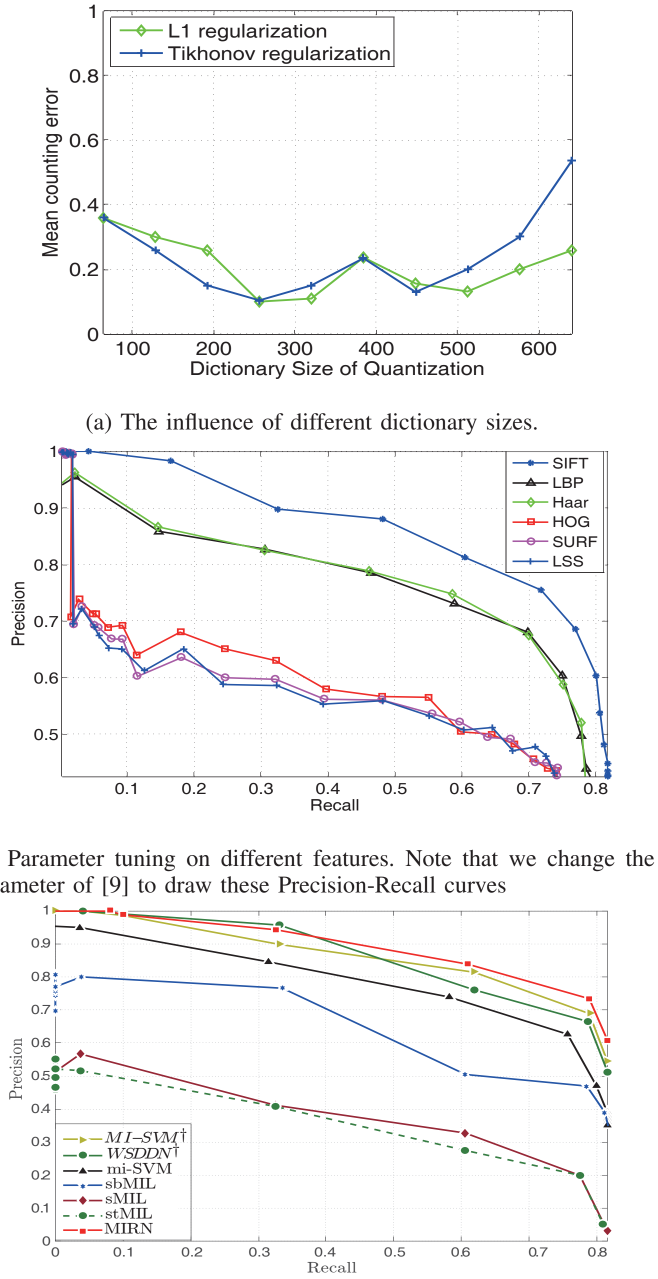
Algorithm 2 Mutil-Instance Ranking Network

```
Input: sallite image I, object counting $m$, soft refine time K
Output: proposal label vectors $\mathrm{Y}_r^k = [y_{1r}^k, ..., \mathrm{Y}_{(c+1)r}^k]^T$
        where $r \in \{1, ..., |R|\}$ and k $\in \{1, ..., K\}$.
Initialization: Feed $I$ and its proposals into the network to
        produce proposals score matrices $f^{Rk}$, $k \in \{0, ..., K-1\}$
for $k = 0 \to K-1$ do
    Initialization: Set all elements in $I = [I_1, ..., I_{|R|}]^T$ to $-\inf$.
    Set all $y_{cr}^{k+1} = 0$, $c \in \{1\}$ and $y_{(c+1)r}^{k+1} = 1$
    Choose the m top-scoring proposals $\{j_{1c}^k, j_{2c}^k, ..., j_{mc}^k\}$ by
      Alg. 1
    for $j_c^k$ $in$ $\{j_{1c}^k, j_{2c}^k, ..., j_{mc}^k\}$ do
        for $r = 1 \to |R|$ do
            Compute IoU $I'_r$ between proposal r and $j_c^k$
            if $I'_r > I_r$ then
                Set $I_r = I'_r$, $w_r^{k1} = f_{c\,j_k^c}^{Rk}$ and $flag = True$
                if $I_r > I_t$ then
                    for $\tilde{r} = 1 \to |R|$ $and$ $\tilde{r} \neq r$ do
                        if $CRR(\tilde{r}, r) \geq T$ then
                            $flag = False$
                        end
                    end
                    if $flag = True$ then
                        Set $y_{Cr}^{k+1} = 1$, $y_{(C+1)r}^{k+1} = 0$
                    end
                end
            end
        end
    end
end
```

C. Parameter Tuning

By using a 5-fold cross-validation, we have depicted and adjusted the following parameters, which have an influence on the overall performance of the proposed approach:

- The tuning of the dictionary size on counting error can be found in Figure. 3a.
- Various features applied for density estimation: In this step, we compare a set of features in density estimation,



(a) The influence of different dictionary sizes.

(b) Parameter tuning on different features. Note that we change the parameter of [9] to draw these Precision-Recall curves

(c) Quantitative comparsion to the state-of-the-art alternative methods incliding [24], [14], [17].

like HOG, SIFT, LBP *etc*, as shown in Figure. 3b, The SIFT feature exihibits better adaptability, which is selected as the feature used subsequently.

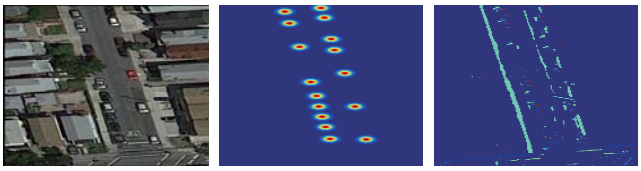


Fig. 4: Vehicle density estimation using vehicle counting. From left to right, each column respectively represents the satellite image, manual annotation for calculating Gaussian density as ground truth of density map, the estimation of the proposed density map.

D. Quantitative Analysis

Finally, we compare the proposed scheme (MIRN) with the above alternative schemes. As shown in Figure. 3c, comparing to the 6 baselines the proposed scheme has achieved better Precision-Recall curves. There is a significant performance boost by utilizing explicit or potential object counting. This is probably due to that the data distribution is hidden in the object counting, which enforces the model to contain more meaningful information. It is worth to note that, the computation costs of the deep neural network will increase explosively with the increase of image size. In addition, it is a better choice to use linear kernel in the proposed multi-instance learning methods, due to its high feature dimensions.

E. Case study

We further give a group of case study in Fig.4 and Fig. 5. In Fig. 4, we visualize the satellite image, roughly manual tagging for density estimation, and the estimated density map. It is shown in Fig. 5 that within a suitable range of variations, vehicles in different directions, colors, and intensities can be detected by using deep features. Certainly, there are also some cases of failure in positioning, especially when the density of vehicle is low, in which case it is difficult to distinguish with the vehicle and the surrounding environment. In addition, using weak supervision alone cannot provide specific location information of the instance, which leads to inaccurate bounding box detections.

V. Conclusion

In this paper, we present a weakly supervised methods to efficiently label sufficient amount of weak (region) labels towards training precise (bounding box) vehicle detectors in remote sensing images. To learn such instance-level vehicle detector from region-level supervision, a multi-instance learning ranking scheme is proposed. Furthermore, considering that the data distribution in a bag probably has a large impact on the stability of output label we further propose a novel object counting scheme to learn the robust detector more efficiently. We have conducted extensive experiments on large-scale satellite images with multiple resolutions. Our approach shows significant advantages over several supervised and weakly-supervised state-of-the-art schemes.

Acknowledgment

This work is supported by the Nature Science Foundation of China No.61772443.

References

[1] L. Cao, C. Wang, and J. Li, “Vehicle detection from highway satellite images via transfer learning,” *Information Sciences (2016) , Online First.*

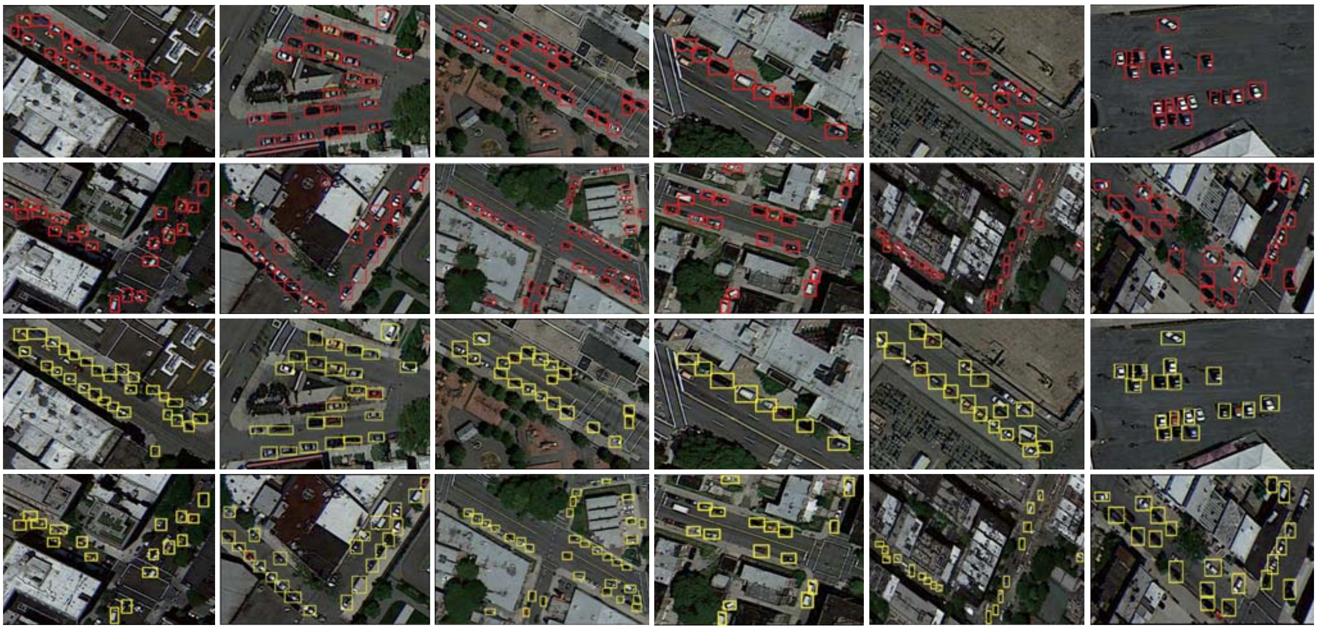


Fig. 5: The first two rows and the last two rows represent the Classification results of our derived method and WSDDN† respectively. Clearly, the proposed method is more robust in detecting ground truth vehicles with low false alarm rates. Note that both schemes do not need bounding box annotations of vehicles, and WSDDN† modified from [17] is regarded as the cutting-edge schemes in weakly supervised detection.

[2] H.-Y. Cheng, C.-C. Weng, and Y.-Y. Chen, "Vehicle detection in aerial surveillance using dynamic bayesian networks," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 2152–2159, 2012.

[3] A. C. Holt, E. Y. Seto, T. Rivard, and P. Gong, "Object-based detection and classification of vehicles from high-resolution aerial photography," *Photogrammetric Engineering & Remote Sensing*, vol. 75, no. 7, pp. 871–880, 2009.

[4] D. Lenhart, S. Hinz, J. Leitloff, and U. Stilla, "Automatic traffic monitoring based on aerial image sequences," *Pattern Recognition and Image Analysis*, vol. 18, no. 3, pp. 400–405, 2008.

[5] K. Kozempel and R. Reulke, "Fast vehicle detection and tracking in aerial image bursts," *Int. Arch. Photogramm. Remote Sens. Spat. Information Science*, vol. 38, pp. 175–180, 2009.

[6] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 6, pp. 1250–1265, 2011.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] V. Lempitsky, M. Verhoek, J. A. Noble, and A. Blake, "Random forest classification for automatic delineation of myocardium in real-time 3d echocardiography," in *International Conference on Functional Imaging and Modeling of the Heart*. Springer, 2009, pp. 447–456.

[9] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.

[10] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[11] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and remote sensing letters*, vol. 11, no. 10, pp. 1797–1801, 2014.

[12] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 113, pp. 155–165, 2016.

[13] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[14] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2003, pp. 577–584.

[15] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2017.

[16] X. Wang, Z. Zhu, C. Yao, and X. Bai, "Relaxed multiple-instance svm with application to object discovery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1224–1232.

[17] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.

[18] P. T. X. W. X. Bai and W. Liu, "Multiple instance detection network with online instance classifier refinement."

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[20] L. Cao, F. Luo, L. Chen, Y. Sheng, H. Wang, C. Wang, and R. Ji, "Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning," *Pattern Recognition*, vol. 64, pp. 417–424, 2017.

[21] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," *arXiv preprint arXiv:1704.00138*, 2017.

[22] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis, "C-wsl: Count-guided weakly supervised localization," *arXiv preprint arXiv:1711.05282*, 2017.

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

[24] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 105–112.