IOP Conf. Series: Materials Science and Engineering **719** (2020) 012064 doi:10.1088/1757-899X/719/1/012064

Robust Vehicle Detection on Multi- Resolution Aerial Images

Ziao Wang[1], Xiaofeng Zhang[1], Xiaofei Yang[1] and Wu Xia[1]

[1]Department of Computer Science, Harbin Institute of Technology, Shenzhen, China
Corresponding author: Xiaofeng Zhang (e-mail: zhangxiaofeng@hit.edu.cn).

Abstract. Detecting vehicles in aerial images is one of the core components for intelligent transportation system. This task is challenging due to the comparably small size of the target objects, the complex background and multi-perspective views. It is particularly difficult for the real-time detection cases where only several tens of milliseconds delays are allowed. In this paper, we propose an approach to detect vehicles from aerial images with different resolutions and perspectives in an approximate real-time manner. The proposed model is robust and can detect vehicles in various detection scenarios. It can detect vehicles from an image within 47 milliseconds. We evaluate our method on a challenging data set of original aerial images over Munich and our data set collected using an unmanned aerial vehicle (UAV). The experimental results have demonstrated that our proposed method is superior to the state-of-the- art algorithms with respect to accuracy, recall and detection time.

1. Introduction

With the rapid development of UAV [1], it has been widely adopted in many application domains, such as intelligent transportation systems. UAV is equipped with automatic positioning and stability system of small high-altitude operating platform. UAV can easily collect video or images which are generally used for the object detections.

For vehicle detection problem using UAV, the traditional detection methods are not appropriate as they are trained on the car camera images. Compared with the traditional perspective of car cameras, UAV images generally cover a larger range of objects which is considered to be able to facilitate the object detections. There are a lot of problems for approaches which are based on aerial images [2], such as low detection accuracy, slow detection speed and unstable detection results when cope with multi-resolution and multi-perspective images [3]. This work is motivated by aforementioned challenging difficulties. Specifically, we propose a fast vehicle detection approach which is robust to the practical application scenes. It is also compatible with a variety of resolutions and multiple perspectives. The detection speed is very fast and can reach 21 FPS with high detection accuracy.

In order to rapidly detect vehicles, a fully convolutional neural networks (FCNN) [4] is chosen as a fast feature extractor. Then, a unified detection for labeling and localization is performed on the feature set extracted by FCNN. Multi-resolution vehicle detection scheme is then introduced to the basic vehicle detector, which makes it compatible with various resolution aerial images. Moreover, a multi-perspective vehicle detection scheme is also integrated into the proposed vehicle detectors making it to be able to work on aerial images collected from a variety of different shooting perspectives.

We evaluate the proposed approaches on two data sets to demonstrate the efficacy of the proposed vehicle detector. One data set is Munich Images adopted in [2], another data set is collected by us which contain multi-resolution and multi- perspective UAV images. In this paper, the detection accuracy rate, recall rate and detection time are chosen as the evaluation criteria of performance.

Our main contributions can be summarized as follows: 1) This approach adopts FCNN to extract

the features of images, which makes the detection speed very fast. The anchor box extracted in FCNN is utilized to localize the objects which is particularly suitable for small object detection. 2) A traffic flow detection method based on weak supervised learning is proposed to solve the problem of annotation data of aerial traffic image. 3) We designed a joint training model using the aerial image of multi-resolution and multi-perspective which can achieve a good model robustness. 4) We have collected a set of aerial images of multi-resolution and multi-perspective as shown in Fig. 1, in order to simulate the practical detection scene.

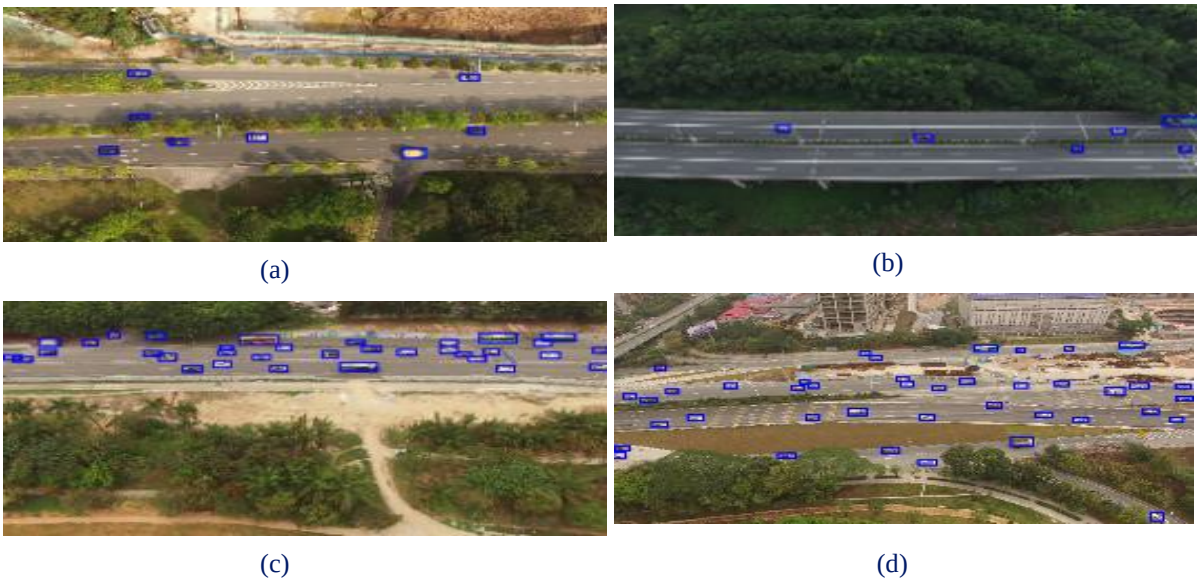


Figure 1. Aerial images with four shooting perspectives. (a) height = 90m and θ = 90o. (b) height = 120m and θ = 90o. (c) height = 90m and θ = 45o. (d) height = 120m and θ = 45o

2. Related Work

In recent years, the deep learning algorithm has achieved great success in the field of image detection and recognition. [5, 6] are state-of-the-art methods based on region proposal, the advantage of those algorithms is detection accuracy. There are some state-of-the-art methods not based on region proposal, such as [7, 8], the advantage of those algorithms is detection speed. Whether it is based on region proposal, those algorithms have achieved good results from high resolution situations, but they acquire poor results when target object is small or images are low pixel. In the literatures, Kaiming et al. proposed deep residual network of extracting highly descriptive features, [9] and it can improve the accuracy of small objects detection and recognition, but it is time consuming. It is difficult that the detection method has high detection accuracy and high detection speed simultaneously. After successful in face recognition, deep learning has been rapidly applied to many fields. However, many fields do not have public data sets as many as face recognition. In some fields, annotation data is very scarce, for example, medical domain [10] and aviation[11] domain. Therefore, more and more people study the effect of using semi-supervised [12] or weak-supervised [13,14,15] deep learning method instead of strong supervised deep learning method to train with less annotation data and achieve a large amount of annotation data to train the model results. Changyu Jiang et al. proposed a weak supervised vehicle learning algorithm for deep learning [16]. This algorithm uses image-level annotation data to train the first few layers of convolutional layers in the detection network and then uses a small amount of annotation data fine-tune the model to obtain an efficient detection model.

There exist a number of literatures about vehicle detection from aerial images. Rodney LaLonde *et al*. proposed a multi-frame multi-object detection method based on FCNN [17], the performance is superior when compared with many state-of-the-art methods. Nassim Ammour *et al*. proposed deep learning approaches to extract highly descriptive features of Vehicle Detection in UAV Images [18], and them method outperformed state-of-the-art methods in terms of accuracy and computational cost.

Yongzheng Xu et al. proposed a combination of the Viola–Jones + SVM algorithm and an HOG + SVM algorithm for vehicle detection in UAV Images [19], which uses a detector switching strategy based the different descending trends of detection speed of both algorithms to improve detection efficiency. Kang Liu *et al.* applied a fast binary detector using integral channel features in a soft-cascade structure for vehicle detection in UAV Images, and estimated the orientation and type of the vehicles [2]. All of the above algorithms are based on aerial images of the same shooting perspective, and they seldom consider the practical issues on vehicle detection using UAV images with different perspectives.

3. Our Approach

This section describes the whole process the proposed vehicle detection network. The design idea of a unified detection network is described.

3.1. Unified Detection Network

In this paper, we propose a fast vehicle detection network based on FCNN detection framework which unifies bounding box prediction and object prediction. As shown in Fig. 2, FCNN is used to directly convert image from the pixel space to the feature space, then the feature map is divided into N×N grids, each grid is responsible for predicting object of the center point falling within it. This detection network predicts whether a grid contains an object or not and predicts the possible location of objects simultaneously.

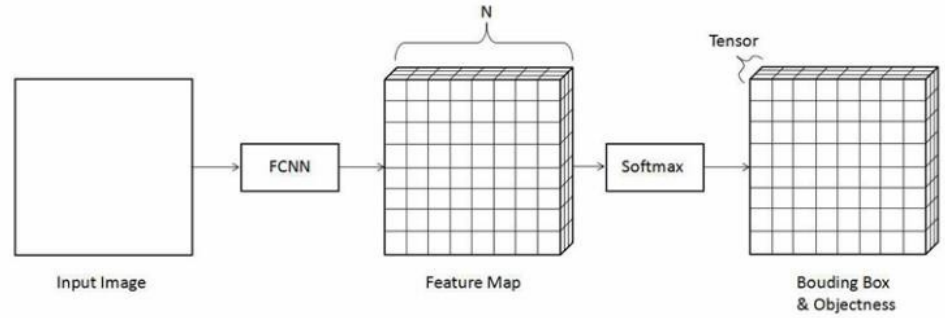


Figure 2. A Unified vehicle detection network based on FCNN. FCNN is used to extract the feature of the input image. Feature map is divided into N×N grids. Feature map is encoded by softmax. A tensor is used for object annotation and object localization of box

Let g_O and g_t denote prediction value and truth value whether grid g contains an object or not, and the range of its value is {0, 1}. The loss of g at grid g is calculated as (1).

$$\mathcal{L}_{ob}(g) = (g_o - g_t)^2 \tag{1}$$

Let $\mathcal{L}_{de}(g)$ denote detection loss of grid g. The overall loss of the network is calculated as (2).

$$\mathcal{L} = \sum_{g\epsilon G} \mathcal{L}_{ob}(g) + \sum_{g\epsilon G} g_{ot} * \mathcal{L}_{det}(g) \tag{2}$$

Where $\boldsymbol{g_{ot}} = \boldsymbol{g_o} * \boldsymbol{g}$, G denotes all grids in the feature map. The first term in (2) represents the cumulative loss when $\boldsymbol{g_o}$ is different from $\boldsymbol{g_t}$. The latter term in (2) represents the cumulative loss when $\boldsymbol{g_o}$ is same as $\boldsymbol{g_t}$.

Due to the easy-to-collect but not easily annotated nature of the aerial data of the subject, we also employ weak supervised learning as illustrated in Fig. 2 with the target of improving the vehicle detection accuracy. As shown in Fig. 3.

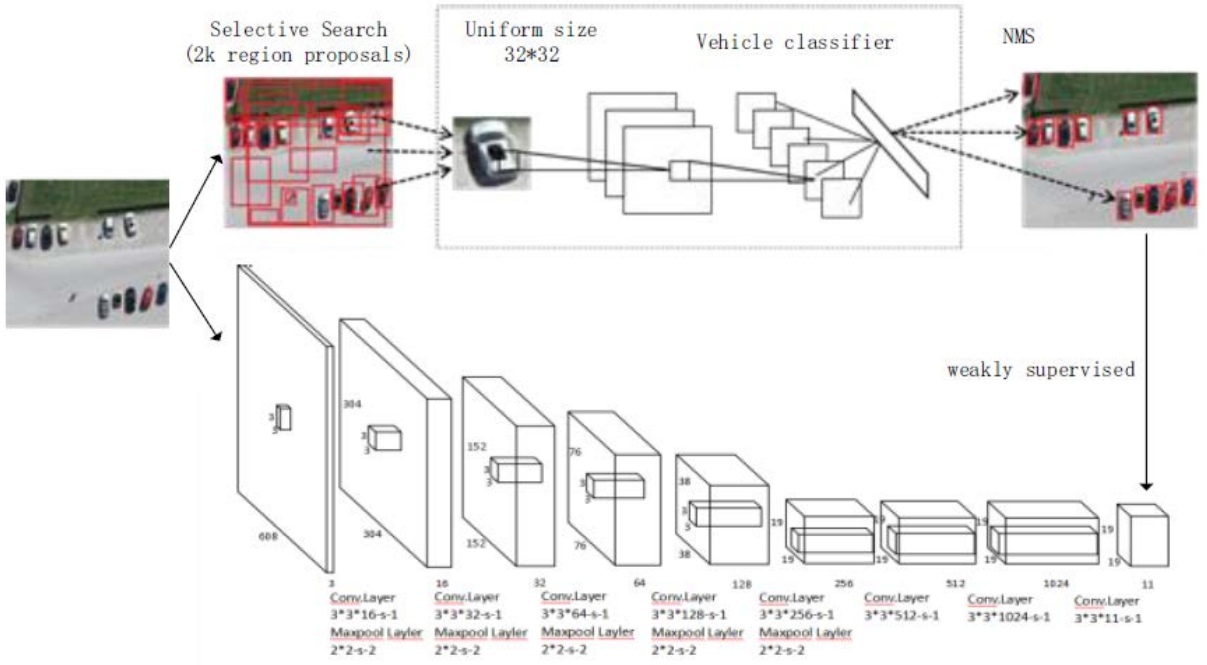


Figure 3. A vehicle detection network based on weakly supervised. The top half is used to generate weakly labeled data. The lower half will use weakly labeled data for training

Each unlabeled image will first extract several candidate boxes using the selective search [20] algorithm, and then classify these candidate boxes using a network similar to LeNet. Finally, the result of the rough annotation is filtered by the non-maximal suppression algorithm. In the rough annotation process, a network similar to LeNet is used to classify the candidate boxes which contains two convolution layers, two maximum pooling layers, two fully connected layers and a lost layer.

The class label $y^{(i)} \in \{0,1\}$ indicates that the problem is a dichotomous problem, and the loss layer uses a logistic regression to calculate the loss. Logistic regression function is given as,

$$h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)} \tag{3}$$

where θ is the model parameter and x is the input eigenvector. To train the model parameter θ, L_{cls} is denoted as,

$$L_{cls} = -\sum_{i=1}^{m}\{y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})\log\left(1-h_\theta(x^{(i)})\right)\} \tag{4}$$

where L_{cls} is classification loss and m is the number of pictures annotated in the training set.

Our network is a two-step one. After training a rough flow detection model using weak supervision, the algorithm uses a traffic detection network based on FCNN. When calculating the prediction loss, the image is treated as annotation data. Because there is some noise in the data learned by the weak supervisor used in the training process of the network, we can reduce the impact of noise on the model training by modifying the loss function. The improved network loss function is given as,

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\Gamma_{ij}^{obj}\bar{P}_{ij}\left((X_i-\bar{X}_i)^2+(Y_i-\bar{Y}_i)^2\right)$$
$$+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\Gamma_{ij}^{obj}\bar{P}_{ij}\left(\left(\sqrt{\omega_i}-\sqrt{\bar{\omega}_i}\right)^2\right.$$
$$\left.+\left(\sqrt{h_i}-\sqrt{\bar{h}_i}\right)^2\right)$$
$$+\sum_{i=0}^{S^2}\sum_{j=0}^{B}\Gamma_{ij}^{obj}\bar{P}_{ij}\left(C_i-\bar{C}_i\right)^2$$
$$+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\Gamma_{ij}^{noobj}\left(C_i-\bar{C}_i\right)^2$$
$$+\sum_{i=0}^{S^2}\Gamma_{i}^{obj}\sum_{c\in classes}\left(P_i(c)-\bar{P}_i(c)\right)^2 \quad (5)$$

Where $\bar{P}_{ij}$ represents the probability that the true box corresponding to the j-th target detection box in the i-th grid containing a car.

The loss function is mainly divided into three parts. The 1-3 items of (5) are the first part, indicating that the prediction box has the classification loss corresponding to the real position, the fourth items is the second part, indicating that the predicted box does not find the loss corresponding to the real box, and the fifth item is the third section, which indicates the classification loss of each grid. After the weak supervision and training, the first part is mainly modified, that is, when the prediction box can find the corresponding box in the annotation data, $\bar{P}_{ij}$ is the weight of the first part of the loss. This weakly supervised training method can integrate the model into the characteristics of more target vehicles, and finally use the strong supervised training method to make the model converge.

4. Experiment

We evaluate our approach on two different data sets, one is Munich Images adopted form [2], and another one is UAV image data sets collected by us including images of different heights and different perspectives. The evaluation criteria are detection time and AP (Average Precision). Rigorous experiments are performed on these two data sets to verify the efficacy of the proposed approach.

Our data set is collected from UAV Images, we use UAV to take video on five different roads with different flying height and shooting angle, a total of 40 videos are collected, accumulating up to 400 minutes. A total of 3000 images were captured from the video, which evenly includes four shooting scenarios (1.height = 90m and θ = 90o; 2.height = 120m and θ = 90o; 3.height = 90m and θ = 45o; 4.height = 120m and θ = 45o). We randomly choose 2000 images as training data set which contain 32646 cars, and the rest 1000 images as testing data set which contain 17548 cars. The image size is 800×450 pixels. The results on our UAV Images are reported in Table 1.

Table 1. Results on Our UAV Images

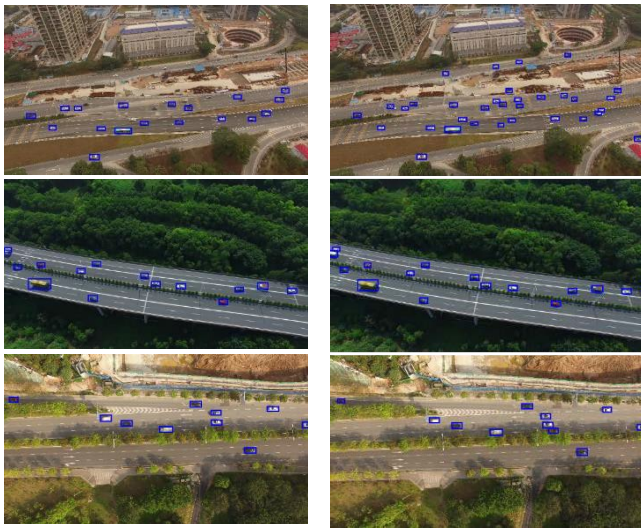
Method	Ground Truth	AP (%)	Time (ms)
Liu-Mattyu's	17548	76.1	76
Faster RCNN	17548	57.4	257
SSD	17548	63.5	39
yolo v2	17548	60.5	**29**
Ours	17548	**88.8**	47

Our method achieves the best detection accuracy on our UAV images with multi-resolutions and multi-perspectives images, which reaching high detection accuracy of 88.8% AP. Yolo v2 has the fastest detection speed with 28 milliseconds per frame, but the performance is not good in terms of detection accuracy. Our method detects an image of 800×450 pixels only took 47 milliseconds which are an approximate real-time speed. Resolution of cars is lower in our UAV Images than Munich Images, it

difficult to detect vehicle on our UAV Images. And the result of other methods proved that vehicle detection is more difficult on our UAV Images. As shown in Fig. 4, our detection results are compared with Liu-Mattyu's to verify the robustness of the proposed approach.

5. Conclusion

We proposed a robust model that can detect vehicles on aerial images of different resolutions and perspectives at approximate real-time speed. This model is very suitable for vehicles detection in aerial images with different detection scene. The method of multi-resolution and multi-perspective training is effective, it can improve the robustness of the detector. Our method outperformed state-of-the-art methods for vehicle detection on multi-resolution and multi-perspective aerial images. It is a good idea to detect the road [23] and then detect the vehicle. As future work, the performance could be further improved by using information which extracts from video frames [24] by LSTM [25], the speed of the detector would still keep fast for vehicles detection in aerial videos.



(a) (b)

Figure 4. Comparison of detection results on four different perspectives. (a) Results of Liu-Mattyu's method. (b) Results of our method. When θ close to 90o, the results of our method and Liu-Mattyu's method are almost the same. However, when θ close to 45o, the detection result of Liu-Mattyu's method is not good, and there are many cars that are not detected. But our detection results are still great on four different perspectives. It indicates that our method is more robust than Liu-Mattyu's method on multi-resolution and multi-perspective aerial images

6. Acknowledgments

This work was supported by the Guangdong Province Science and Technology Department Project under Grant NO. 2017B090901022. This paper is partially supported by the National Science Foundation of China under grant No.61872108, and Shenzhen Science and Technology Program under Grant No.JCYJ20170811153507788.

7. References

[1] Pérez A, Chamoso P, Parra V, and Sánchez A. J. Ground vehicle detection through aerial images taken by a UAV. Information Fusion (FUSION), 2014 17th International Conference on. IEEE, 2014:1-6.

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

[3] Khoshelham K, Nardinocchi C, Frontoni E, et al. Performance Evaluation of Automated Approaches to Building Detection in Multi-Source Aerial Data. Isprs Journal of Photogrammetry and Remote Sensing, 2010, 65(1):123-133.

[4] Dai J, Li Y, He K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks. Advances in neural information processing systems, 2016, 26(1):397-389.
[5] Girshick R. Fast r-cnn. Proceedings of the IEEE International Conference on Computer Vision, 2015:1440-1448.
[6] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015:91-99.
[7] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. arXiv preprint arXiv:1612.08242, 2016.
[8] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, and Berg A C. SSD: Single shot multibox detector. European Conference on Computer Vision. Springer International Publishing, 2016:21-37.
[9] He K, Zhang X, Ren S, and Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-778.
[10] Shen D, Wu G, Suk H I. Deep Learning in Medical Image Analysis. Annual Review of Biomedical Engineering, 2017, 19(1):221-248.
[11] Tao Y, Wang X, Yao B, et al. Small Moving Vehicle Detection in a Satellite Video of an Urban Area. Sensors, 2016, 16(9):1528-1542.
[12] Zhang J, Han Y, Tang J, et al. Semi-Supervised Image-to-Video Adaptation for Video Action Recognition. IEEE Transactions on Cybernetics, 2016, 47(4):960-973.
[13] Jia Z, Huang X, Chang E I, et al. Constrained Deep Weak Supervision for Histopathology Image Segmentation. IEEE Transactions on Medical Imaging, 2017, 27(7):1-16.
[14] Pathak D, Krahenbuhl P, Darrell T. Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. Proceedings of the IEEE International Conference on Computer Vision, 2015:1796-1804.
[15] Croft W B, Croft W B, Croft W B, et al. Neural Ranking Models with Weak Supervision. arXiv preprint arXiv:1704.08803, 2017.
[16] Jiang C, Zhang B. Weakly-Supervised Vehicle Detection and Classification by Convolutional Neural Network. Proceedings of the International Congress on Image and Signal Processing, 2017:570-575.
[17] LaLonde R, Zhang D, Shah M. Fully Convolutional Deep Neural Networks for Persistent Multi-Frame Multi-Object Detection in Wide Area Aerial Videos. arXiv preprint arXiv:1704.02694, 2017.
[18] Ammour N, Alhichri H, Bazi Y, Benjdira B, Alajlan N, and Zuair M. Deep Learning Approach for Car Detection in UAV Imagery. Remote Sensing, 2017, 9(4): 312-322.
[19] Xu Y, Yu G, Wang Y, Wu X, and Ma Y. A hybrid vehicle detection method based on Viola-Jones and HOG+ SVM from UAV images. Sensors, 2016, 16(8): 1325-1348.
[20] Uijlings J R, Sande K E, Gevers T, et al. Selective Search for Object Recognition. International Journal of Computer Vision, 2013, 104(2):154-171.
[21] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, and Darrell T. Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014:675-678.
[22] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 2014, 15(1):1929-1958.
[23] Cheng G, Wang Y, Xu S, et al. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(6):3322-3337.
[24] Pu F, Xie W, Cheng Y, et al. Implementation of Real-Time Vehicle Tracking in City-Scale Video Network. Cybernetics and Systems, 2016, 47(4):249-260.
[25] Greff K, Srivastava R K, Koutník J, Steunebrink B R, and Schmidhuber J. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 2016: 2222-2232.