Journal of Theoretical and Applied Information Technology
15th March 2020. Vol.98. No 05

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A FAST PYRAMID NETWORK FOR ACCURATE LOCALIZATION OF CAR IN AERIAL IMAGES

HOANH NGUYEN
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
E-mail: nguyenhoanh@iuh.edu.vn

ABSTRACT

Although deep learning-based object detectors have achieved great success in general object detection in recent years, detecting of objects like car in aerial images is still a challenge. The main difficulty of car detection in aerial images comes from the relatively small size with multiple orientations of car in images. In addition, due to the high resolution of aerial images, the inference time of current approaches is still high. To solve these problems, this paper proposes an enhanced framework for fast and efficient car detection in aerial images. In the proposed approach, ResNet-34 architecture is adopted to create the base convolution layers. Compared with ResNet-50 and ResNet-101, ResNet-34 achieves comparable performance while being faster and simple. Then, an enhanced feature map generation module is designed to generate enhanced feature maps from input feature maps. To speed up the detection process, the detection network based on region proposal network is used to exactly locate cars in original aerial images. The detection network included region proposal networks is applied at different enhanced feature maps with different scales to detect multi-scale car in input image. Experimental results on public dataset show that the proposed approach achieves comparable performance compared with other state-of-the-art approaches.

Keywords: *Car Detection, Convolutional Neural Network, Intelligent Transportation System, Object Detection, Pyramid Network*

1. INTRODUCTION

Vision-based car detection plays an important role for a wide range of applications and is receiving significant attention in recent years. However, car detection in aerial images is still a challenging problem due to difficult environments such as the relatively small size, varying types, and variable orientation of cars. In addition, the presence of many structures such as plane, ship, and so on which appear visually similar to car, can cause false detections. Furthermore, due to the high resolution of aerial images, the processing time is limited for real-time applications, which also increases the difficulties of car detection in aerial images.

In previous studies, various algorithms had been proposed for car detection in aerial images. The most common method is based on a sliding-window search in which each image is scanned in all positions with different scales. Then, multiple handcrafted features or shallow-learning-based features associated with AdaBoost classifiers, or support vector machine classifiers, are used to examine each window for the presence of a car. These methods showed a good performance. However, these methods suffer from several limitations. First, models were trained in clumsy and slow multistage pipelines. Second, handcraft features or shallow-learning-based features influenced the representational power, as well as the effectiveness of car detection. Third, the sliding-window technique led to heavy computational costs.

In recent years, vision-based object detection methods are driven by the success of region-based convolutional neural networks (CNNs). These methods first generate object-like regions and extract highly discriminative features from each region using CNN, and then classify each region with category-specific classifiers. In addition, deeper CNN features are more powerful in representation than low-level features, which can significantly improve the performance of object detection. On the other hand, CNN-based model is faster than sliding-window-based detectors as it uses hundreds of proposed object-like regions to reduce the search space for the whole image. Enhanced network based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on region-based network such as Spatial Pyramid Pooling network (SPPnet) [11] and Fast R-CNN [12] have reduced the running time with impressive performance results. Nevertheless, the main issue of the above-mentioned models is that the human-designed region proposal generator is extremely time consuming. The recently proposed Faster R-CNN [3] presents region proposal network (RPN) and combines the RPN and Fast R-CNN into a unified network, which achieves state-of-the-art detection performance and further speed improvement. Although the Faster R-CNN model has achieved great success in general object detection, several challenges in aerial images limit its applications in car detection. First, cars in large-scale aerial images are relatively small in size with multiple orientations while Faster R-CNN has poor localization performance with small objects as the CNN feature used for classification and regression is pooled from the last convolutional feature map with lower resolution. Second, Faster R-CNN is particularly designed for extracting the bounding box of the targets without considering annotation of multiple attributes for targets. Sometimes, the attributes of a car are important for intelligent traffic management systems. Third, manual annotation is generally expensive and the available manual annotation of cars for training faster RCNN are not sufficient in number.

To tackle the above problems, this paper proposes an enhanced network based on feature pyramid networks (FPN) [1] for fast and efficient car detection in aerial images. In the proposed framework, ResNet-34 architecture is adopted to create the base convolution layers from input image. RestNet-34 is a simple and fast CNN architecture. Compared with ResNet-50 and ResNet-101, ResNet-34 achieves comparable performance while being faster and simple. Then, an enhanced feature map generation module is then designed to create enhanced feature maps from input feature maps. Each enhanced feature map is fed to the detection network, which is based on region proposal network, to exactly locate car in original aerial images. The detection network is applied at different enhanced feature maps at different scales to detect multi-scale object in input image. Experimental results on public dataset show that the proposed approach achieves comparable performance compared with other state-of-the-art approaches, while being faster.

This paper is organized as follows: an overview of previous methods is presented in Section 2. Section 3 describes detail the proposed method. Section 4 demonstrates experimental results. Finally, the conclusion is made in Section 5.

2. RELATED WORK

2.1 Deep Learning-Based Object Detection

Recently, with fast development of deep learning, many deep learning-based object detection approaches have been proposed with impressive performance [8], [9]. The most successful object detectors are region-based convolutional neural networks (CNNs), which employ region proposal algorithms to guide the search for objects, thereby avoiding a sliding window search across the whole image. Girshick et al. [10] first presented an R-CNN detector, which consists of four parts: First, object-like regions are generated by human-designed region proposal algorithms. Then, regions are resized, and highly discriminative features are extracted from each region using CNN. Finally, regions are classified by SVM and bounding boxes are refined by regression. He et al. [11] proposed SPP-net in which a spatial pyramid-based pooling layer was employed to deal with different region sizes. In order to increase the speed and accuracy of detection further, Fast R-CNN [12] trained classifiers and bounding-box regression with an end-to-end solution. However, this approach depends on time-consuming human-designed region proposal algorithms. To solve this problem, Ren et al. [3] proposed Faster R-CNN, which consists of an RPN for predicting candidate regions, and a R-CNN classifier, which achieves near real-time rates and state-of-the-art performance. To increase the inference speed and maintain the detection accuracy, SSD [13] proposed a one-stage deep learning framework for fast object detection. In SSD, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes. SSD is simple because it completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network. This makes SSD easy to train and straightforward to integrate into systems that require a detection component. You Only Look Once (YOLO) [14] solves object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation.

The above deep learning-based object detectors were successful in general object detection in natural scene images, but they have not been carefully explored in object detection in aerial images. Compared with object detection in natural scene images, detection of targets in aerial images is a more challenging task.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.2 Object Detection in Aerial Images

Objects in aerial image are usually in difficult conditions for locating such as small, heavy occlusion, numerous, and so on. To detect object in aerial images, Zhang [15] proposed an aircraft detection method based on coupled CNNs, which combines a candidate RPN to extract the image-level proposals of an aircraft and a localization network to locate the aircraft. This weakly supervised learning-based method is a promising way to alleviate the human labor cost of annotation. However, it may be problematic for direct use in vehicle detection as multiple categories of vehicles may increase the intraclass differences and many structures appearing visually similar to vehicles in the background may reduce interclass differences. Moreover, it is not a real-time approach for detection. The proposed method in [16] starts with a screening step of asphalted zones in order to restrict the areas where to detect cars and thus to reduce false alarms. Then, a feature extraction process is performed based on scalar invariant feature transform thanks to which a set of keypoints is identified in the considered image and opportunely described. In [17], filtering operations in the horizontal and vertical directions were performed to extract histogram-of-gradient features and to yield a preliminary detection of cars after the computation of a similarity measure with a catalog of cars used as reference. Chen et al. [18] presented an algorithm for vehicle detection in high-resolution aerial images through a fast-sparse representation classification method and a multiorder feature descriptor that contains information of texture, color, and high-order context. To speed up computation of sparse representation, a set of small dictionaries, instead of a large dictionary containing all training items, is used for classification. In [19], a superpixel segmentation method designed for aerial images was proposed to control the segmentation with a low breakage rate. Zhang et al. [20], a method based on the two-layer visual saliency analysis model and support vector machines was proposed for high-resolution broad-area remote-sensing images. In the first layer saliency model, a spatial-frequency visual saliency analysis algorithm based on a CIE Lab color space is introduced to reduce the interference of backgrounds and efficiently detect well-defined airport regions in broad-area remote-sensing images. In the second layer saliency model, a saliency analysis strategy based on an edge feature preserving wavelet transform and high-frequency wavelet coefficient reconstruction is proposed to complete the pre-extraction of aircraft candidates from airport regions. Konstantinidis et al. [21] proposed a building detection method that consists of two modules. The first module is a feature detector that extracts histograms of oriented gradients (HOG) and local binary patterns (LBP) from image regions. The second module consists of a set of region refinement processes that employs the output of the HOG-LBP detector in the form of detected rectangular image regions.

3. METHODOLOGY

Figure 1 illustrates the overall architecture of the proposed method. As shown in Figure 1, ResNet architecture is adopted to create convolution layers from input image. An enhanced feature map generation module is then designed to create enhanced feature map from input feature map. Each enhanced feature map is fed to the detection network, which is based on region proposal network, to exactly locate car in original image. The detection network is applied at different enhanced feature map at different scales to detect multi-scale object in input image. Details of each module will be explained in the following sections.

3.1 Feature Extraction Subnet

The feature extraction subnet is responsible for computing a convolutional feature map over an entire input image. This paper adopts the Feature Pyramid Network (FPN) [1] as the feature extraction network. FPN uses ResNets [2] as the base convolution layers. FPN augments a standard convolutional network with a top-down pathway and lateral connections so the network efficiently constructs a rich, multi-scale feature pyramid from a single resolution input image. Each level of the pyramid can be used for detecting objects at a different scale.

3.1.1 ResNet

ResNet is an efficient network proposed by Kaiming He which adopted residual learning to every few stacked layers such that the training of networks can be eased and substantially deeper than others. ResNet-50 and ResNet-101 have high precision, but they are slower than ResNet-34. For fast and efficient license plate detection, this paper adopts ResNet-34 as the based convolution layers of the FPN. Table 1 shows the architecture of the ResNet-34. The FPN chooses the output of the last layer of each stage as reference set of feature maps, which will be enriched to create the feature pyramid. The choice of last layer is natural since the deepest layer of each stage should have the strongest

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

features. More specific, let {C2, C3, C4, C5} denote the output feature map of last residual blocks for conv2, conv3, conv4, and conv5. These feature maps have strides of {4, 8, 16, 32} pixels respectively with respect to the input image. The feature map outputted by conv1 stage is not included into the pyramid due to its large memory footprint.

3.1.2 Enhanced Feature Map Generation

The FPN augments another top-down pathway besides the backbone through combining the low-resolution, semantically strong features from the top with the high-resolution, semantically weak features from the bottom. Thus, the multi-scale feature pyramid is simple and efficient to build, with rich semantic representation at all levels. Figure 2 shows the architecture of enhanced feature map generation module. As shown in Figure 2, the high-level feature (low resolution) is upsampled by the factor of 2 using the nearest neighbor upsampling method, and then it is combined with the corresponding previous feature map in the backbone by using element-wise addition. The previous feature map in the backbone would be subjected to a 1 x 1 convolution kernel to change the dimensions, which should be the same as the dimensions in the next layer. This process is repeated iteratively until the finest feature map is generated. At the beginning of the iteration, a 1 x 1 convolutional kernel is added after C5 layer to produce the coarsest feature map. Finally, a 3 x 3 convolution kernel is used on each merged map to generate the last required feature map in order to eliminate the aliasing effect of upsampling. The corresponding feature layers {C2; C3; C4; C5} are {P2; P3; P4; P5}, and the corresponding layer space dimensions are the same.

3.2 The Detection Network

Region Proposal Network (RPN) [3] is a popular sliding-window object detector. In RPN, a small subnetwork is evaluated on dense 3×3 sliding windows, on top of a single-scale convolutional feature map, performing object/background classification and bounding box regression. This is realized by a 3×3 convolutional layer followed by two sibling 1×1 convolutions for classification and regression as shown in Figure 3. The object/background criterion and bounding box regression target are defined with respect to a set of reference boxes called anchor boxes. The anchor boxes are of multiple pre-defined scales and aspect ratios in order to cover objects of different shapes. This paper attaches each RPN to each level on the feature pyramid. Because the RPN slides densely over all locations in all pyramid levels, it is not necessary to have multi-scale anchors on a specific level. Instead, this paper assigns anchors of a single scale to each level. More specific, this paper defines the anchors to have areas of $\{32^2, 64^2, 128^2, 256^2\}$ pixels on {P2, P3, P4, P5} respectively. As in [3], this paper also uses anchors of multiple aspect ratios {1:2, 1:1, 2:1} at each level. So, there are 12 anchors over the pyramid in total.

For training samples, this paper assigns training labels to the anchors based on their Intersection-over-Union (IoU) ratios with ground truth bounding boxes. More specific, an anchor is assigned as positive label if the IoU is over 0.7 with any ground truth box, and an anchor is assigned as negative label if the IoU is lower than 0.3 for all ground truth boxes. Moreover, scales of ground truth boxes are not explicitly used to assign them to the levels of the pyramid. Ground truth boxes are associated with anchors, which have been assigned to pyramid levels. Thus, no extra rules are introduced in addition to those in [3]. Notably, the parameters of the RPNs are shared across all feature pyramid levels.

3.3 Loss Function

The cross-entropy loss is a common choice to classify the foreground and background classes, which is defined as follow:

$$L_{CE}(p,q) = -(q\log(p) + (1-q)\log(1-p)) \quad (1)$$

where $q \in \{1,-1\}$ specifies the ground truth class and $q \in [0,1]$ represents for the estimated probability of the class with label $q = 1$. For convenience, this paper defines p_t as follow:

$$p_t = \begin{cases} p & \text{if } q = 1 \\ 1-p & \text{otherwise} \end{cases} \quad (2)$$

Thus, the cross-entropy loss can be rewritten as follow:

$$L_{CE}(p_t) = -\log(p_t) \quad (3)$$

The traditional cross-entropy loss function leads to extreme imbalance of the positive and negative sample ratio, and most of the negative samples are easy example. Although these easy examples are insignificant in loss, they still make a great contribution to loss because of their large number, resulting in convergence to a result that is not good enough. Focal loss [4] achieves great success on solving the problem of class imbalance during training. The focal loss is defined as follow:

$$L_{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (4)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where $\gamma \geq 0$ is tunable focusing parameter.
With focal loss, the gradient from easy examples can be automatically filtered out. In other words, the focal loss focuses training on a sparse set of hard examples and effectively discounts the effect of easy examples. Naturally, to solve this problem, this paper uses focal loss to train the proposed network.

4. EXPERIMENTAL RESULTS

In order to compare the effectiveness of the proposed method with other methods on car detection in aerial images, this paper conducts experiments on the DLR Munich dataset [5]. The proposed method is implemented on a Window system machine with Intel Core i7 8700 CPU, NVIDIA GTX 1080 GPU and 8 GB of RAM. TensorFlow is adopted for implementing deep CNN frameworks.

4.1 Dataset

DLR Munich dataset [5] a publicly available car in aerial image dataset, which was collected over the city of Munich, Germany. It contains 20 aerial images that were captured from an airplane by a Canon Eos 1Ds Mark III camera with a resolution of 5616 × 3744 pixels, 50 mm focal length. All images are in JPEG format. The optical image was taken at a height of 1000 m above ground, the ground sampling distance is approximately 13 cm. As in [5], the first ten images were used for training and the other ten images for testing. Positive training samples come from 3418 cars annotated in the training images, while the negatives are randomly picked from the background, i.e. areas without cars.

4.2 Metrics

This paper adopts widely used measures to quantitatively evaluate the performance of the proposed method on car detection in aerial image, including Precision (P), Recall (R), Average Precision (AP), and F1-score. These criteria are defined as follows:

$$P = \frac{TP}{(TP + FP)} \tag{5}$$

$$R = \frac{TP}{(TP + FN)} \tag{6}$$

$$\text{F1} - \text{score} = 2 \times \frac{(P \times R)}{(P + R)} \tag{7}$$

where TP (True Positive) represents the correct detections; FP (False Positive) represents the wrong detections; FN (False Negative) represents the number of missed detections.

Both the precision and recall metrics measure the fraction of true positive detections and correctly identified positive detections. The AP metric is measured by the area under the precision-recall curve. The higher the AP value, the better the performance, and vice versa. The F1-score combines the precision and recall metrics to a single measure to comprehensively evaluate the quality of an object detection method. Generally, a detection result is considered to be a true positive if the IoU between a detected bounding box and ground truth bounding box is greater than 0.5. Otherwise the detection is considered as a false positive. Furthermore, if several detections overlap with the same ground truth, only one detection with the highest overlap ratio is considered a true positive, and others are considered false positives.

4.3 Implementation Details

The base ResNet-34 is pre-trained on ImageNet1k. This paper uses the models released by [2]. The proposed network is trained end-to-end with stochastic gradient descent (SGD). The network is trained for 90k iterations with an initial learning rate of 0.01, which is then divided by 10 at 60k and again at 80k iterations. This paper uses horizontal image flipping as the only form of data augmentation unless otherwise noted. Weight decay of 0.0001 and momentum of 0.9 are used. As in [1], this paper includes the anchor boxes that are outside the image for training. For training loss, this paper sets γ = 2 and α_t = 0.25, which have been proved effectively in practice.

4.4 Results

To show the effectiveness of the proposed method, this paper compares the detection results of the proposed method with the results of other state-of-the-art methods on DLR Munich dataset, including AVPN [6], Faster R-CNN [3], H-RPN [7], and the method proposed by Liu et al. [5]. AVPN developed an accurate-vehicle proposal-network (AVPN) based on hyper feature map which combines hierarchical feature maps that are more accurate for small object detection. For detection network, AVPN proposed a coupled R-CNN method, which combines an AVPN and a vehicle attribute learning network to extract the vehicle's location and attributes simultaneously. Faster R-CNN proposed a combination of RPN and fast RCNN. H-RPN employed a hyper region proposal network (HRPN) to extract vehicle-like targets with a combination of hierarchical feature maps. Table 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

shows the comparison of the detection results. In Table 2, it can be observed that the proposed approach achieved the best performance in terms of recall rate. More specific, in terms of recall rate, the performance of the proposed method is improved by 9.3%, 1.58%, 9.86%, 0.3% compared with Liu [5], AVPN_large [6], Faster R-CNN [3], and H-RPN [7] respectively. This result demonstrates the effectiveness of the proposed enhanced feature map generation module for generating proposals at high recall. In terms of precision, the proposed method achieves nearly as performance as other methods. This paper proposed a simple and fast detection network, resulting in a weak precision. In future work, this paper will consider improving the precision by designing a better detection network. For the inference time, by adopting a simple and fast detection network, the proposed method achieves the best inference time. More specific, the proposed framework takes only 1.12 second to process a high-resolution image, while Faster R-CNN takes up to 3.84 seconds. This result demonstrates that the proposed method meets the requirements of real-time processing and can be applied to real-time system. Figure 4 shows examples of detection results of the proposed method on DLR Munich dataset. Figure 5 illustrates some failed detection results. As shown in Figure 4 and Figure 5, despite cars located in the shade or near the image block boundaries, the proposed approach had successfully detected most of the cars. The missing detected cars are mostly located near the boundaries with darker in color. This is because the cars located near the boundaries are easy to lose part of the information in the small-size hyper feature maps.

5. CONCLUSIONS

This paper proposes an enhanced network based on feature pyramid networks for fast and efficient car detection in aerial images. In the proposed framework, ResNet-34 architecture is adopted to create the base convolution layers from input image. Then, an enhanced feature map generation module is then designed to create enhanced feature maps from input feature maps. Each enhanced feature map is fed to the detection network, which is based on region proposal network, to exactly locate car in original aerial images. The detection network is applied at different enhanced feature maps at different scales to detect multi-scale object in input image. Experimental results on DLR Munich dataset show that the proposed approach achieves comparable performance compared with other state-of-the-art approaches, while being simpler and faster. However, the proposed method still produces some false, as well as missing detection. Extracting good car-like regions is still a critical task for accurate car detection. Hence, in future study, this paper will focus on mining hard negative samples by a bootstrapping strategy.

REFERENCES:

[1] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," In *CVPR*, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[4] Lin T, Goyal P, Girshick RB, He K, Dolla´r P (2017), "Focal loss for dense object detection," *IEEE international conference on computer vision, ICCV 2017*, Venice, Italy, 22–29 Oct 2017, pp 2999–3007.

[5] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938 1942, Sep. 2015.

[6] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3652–3664, Aug. 2017.

[7] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, p. 336, 2017.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[9] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[12] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[13] Liu, Wei, et al., "Ssd: Single shot multibox detector," *European conference on computer vision*, Springer, Cham, 2016.

[14] Z. Xiao, Q. Liu, G. Tang, and X. Zhai, "Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images," *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 618-644, 2014.

[15] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, "Efficient saliency-based object detection in remote sensing images using deep belief networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 137–141, Feb. 2016.

[16] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1635–1647, Mar. 2014.

[17] T. Moranduzzo and F. Melgani, "Detecting cars in uav images with a catalog-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6356–6367, Oct. 2014.

[18] Z. Chen et al., "Vehicle detection in high-resolution aerial images based on fast sparse representation classification and multiorder feature," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2296–2309, Aug. 2016.

[19] Z. Chen et al., "Vehicle detection in high-resolution aerial images via sparse representation and superpixels," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 103–116, Jan. 2016.

[20] L. Zhang and Y. Zhang, "Airport detection and aircraft recognition based on two-layer saliency model in high spatial resolution remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1511–1524, Apr. 2017.

[21] D. Konstantinidis, T. Stathaki, V. Argyriou, and N. Grammalidis, "Building detection using enhanced hogclbp features and region refinement processes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 888–905, Mar. 2017.

Journal of Theoretical and Applied Information Technology
15th March 2020. Vol.98. No 05

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

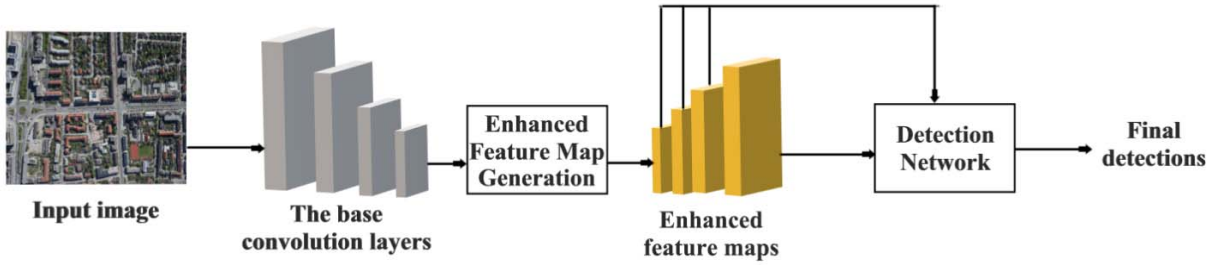

Figure 1: The Overall Framework of The Proposed Approach.

Table 1: The Architecture of ResNet-34.

Layer name	Kernel size	Output size
Conv1	$7 \times 7 \times 64$	112×112
Conv2	$\begin{bmatrix} 3 \times 3 \times 64 \\ 3 \times 3 \times 64 \end{bmatrix} \times 3$	56×56
Conv3	$\begin{bmatrix} 3 \times 3 \times 128 \\ 3 \times 3 \times 128 \end{bmatrix} \times 4$	28×28
Conv4	$\begin{bmatrix} 3 \times 3 \times 256 \\ 3 \times 3 \times 256 \end{bmatrix} \times 6$	14×14
Conv5	$\begin{bmatrix} 3 \times 3 \times 512 \\ 3 \times 3 \times 512 \end{bmatrix} \times 3$	7×7

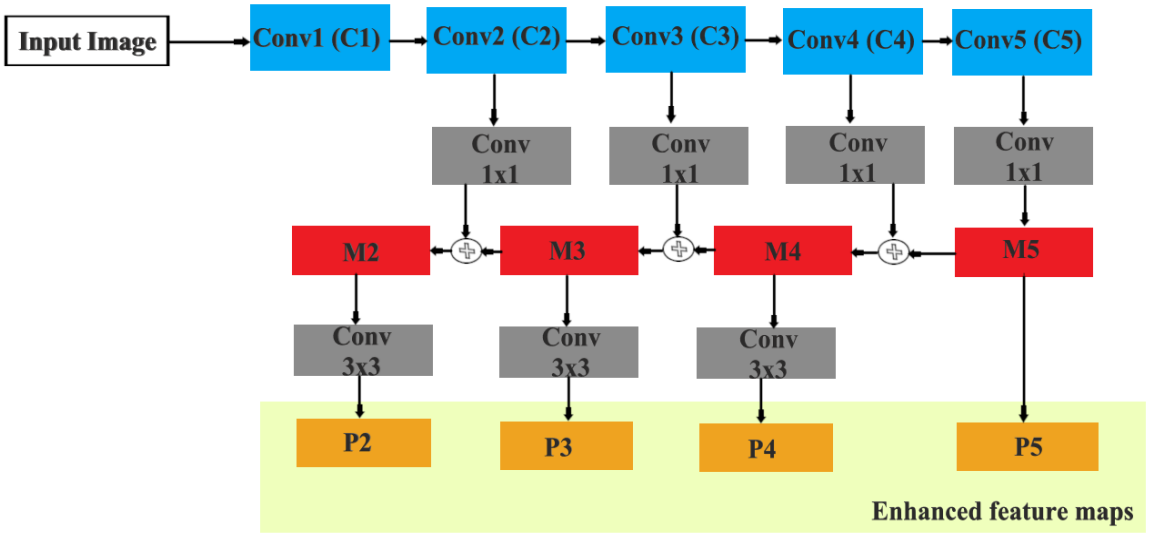

Figure 2: The Architecture of The Enhanced Feature Map Generation Module.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

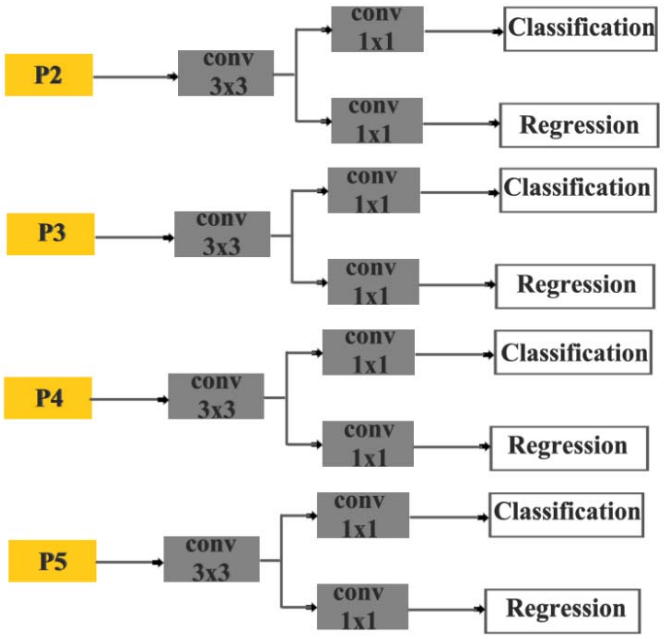

Figure 3: The Detection Network.

Table 2: Performance Comparison of Different Methods.

Methods	Recall (%)	Precision (%)	F1-score (%)	Inference Time (s)
Liu et al. [5]	69.3	86.8	77	4.4
AVPN_basic [6]	75.59	85.93	80	3.65
AVPN_large [6]	77.02	87.81	82	3.65
AVPN_basic+fast R-CNN [6]	74.73	91.98	82	4.05
Faster R-CNN [3]	68.74	88.95	78	3.84
H-RPN [7]	78.3	89.2	83	-
Proposed Method	78.6	85.3	81.8	1.12

Journal of Theoretical and Applied Information Technology
15th March 2020. Vol.98. No 05

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

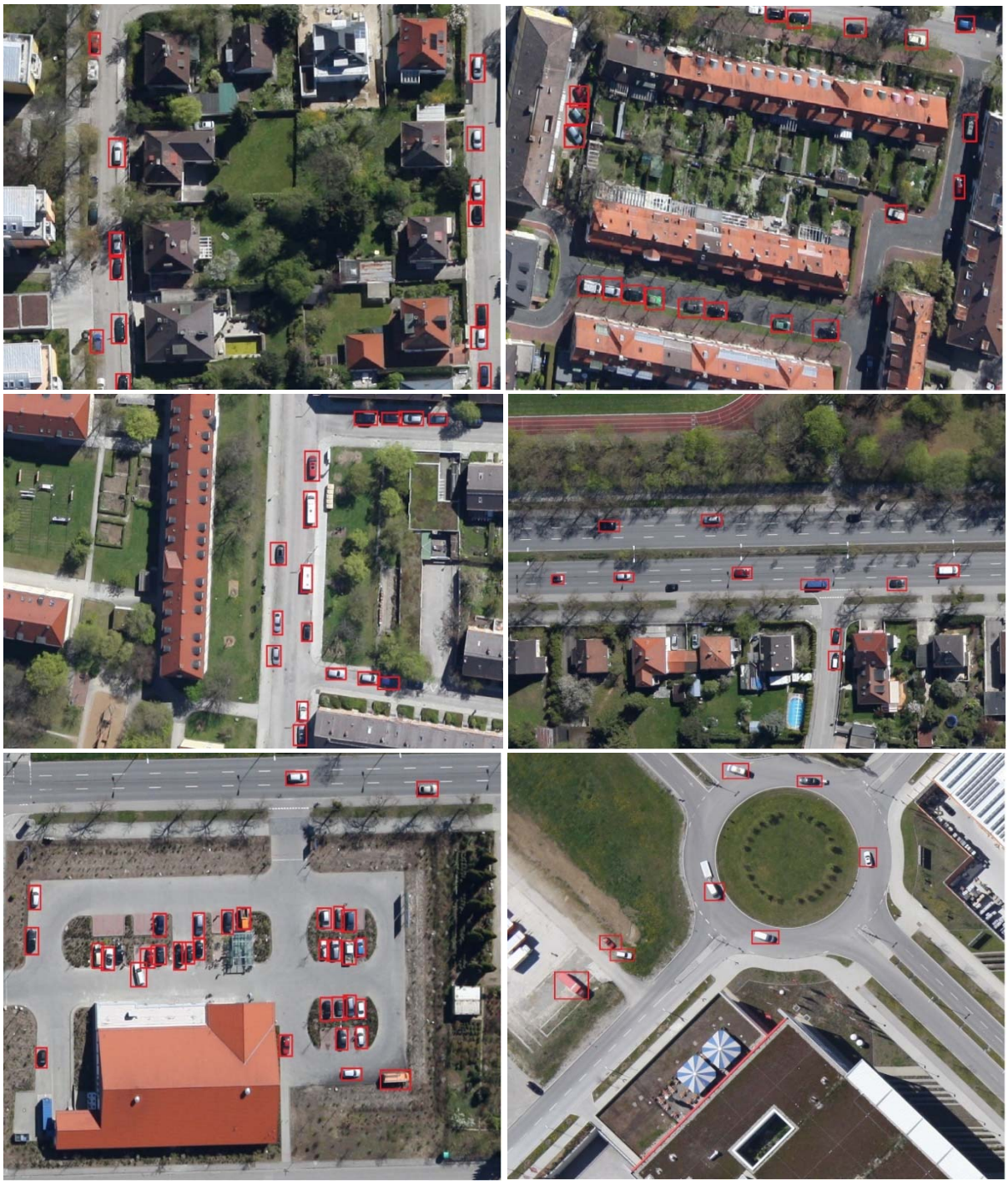

Figure 4: Detection Results on DLR Munich Dataset.

Journal of Theoretical and Applied Information Technology
15th March 2020. Vol.98. No 05

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

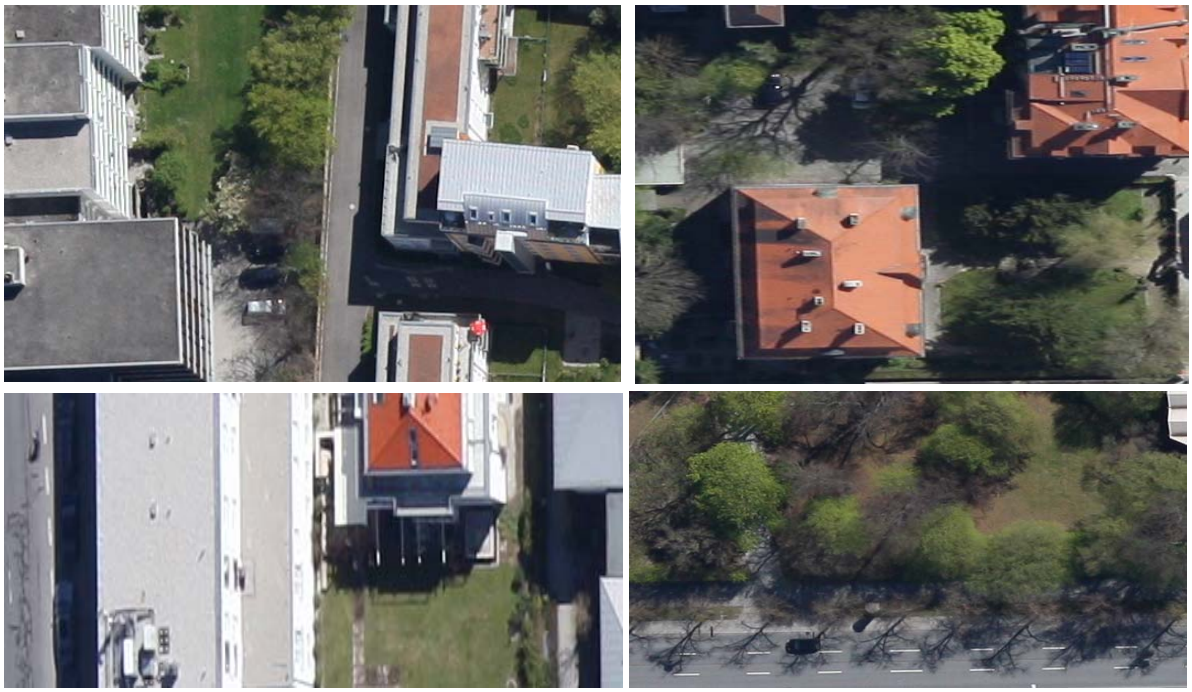

Figure 5: Undetected Cars Due to Difficult Conditions.