# Ensemble of Two-Stage Regression Based Detectors for Accurate Vehicle Detection in Traffic Surveillance Data

Lars Sommer[2,1] Oliver Acatay[1] Arne Schumann[1] Jürgen Beyerer[1,2]

[1]Fraunhofer IOSB
Fraunhoferstrasse 1
76131 Karlsruhe, Germany

[2]Vision and Fusion Lab
Karlsruhe Institute of Technology KIT
c/o Technologiefabrik, Haid-und-Neu-Str. 7
76131 Karlsruhe, Germany

## Abstract

*The growing amount of traffic surveillance data results in an increased need for automatic detection systems to analyze the data. For this purpose, deep learning based detection frameworks like Faster R-CNN and SSD have been employed in recent years. Though the detection accuracy is clearly improved compared to conventional detection methods, there exists large potential for further improvements especially in case of adverse weather conditions. In this paper, we employ the RefineDet detection framework as it combines advantages of several detection frameworks including Faster R-CNN and SSD. We use an ensemble of two detectors with different base networks to generate detections that are more robust. For this, SENets – the winner of the ImageNet2017 classification challenge – are used in addition to ResNet-50. To account for small vehicles in the background and strong variation in vehicle scale, we apply multi-scale testing. Our proposed detector achieves top-performing results on the UA-DETRAC dataset especially in case of rainy and nighttime scenarios.*

## 1. Introduction

Growing cities and traffic densities have led to an increased demand for surveillance systems capable of automatic traffic monitoring and traffic analysis. Such systems are generally based on images or videos captured by surveillance cameras. Due to the huge amount of data to analyze, automatic detection systems are required. However, detecting all vehicles robustly and efficiently is challenging due to various object scales, different object types, various perspectives, occlusion, varying daytimes and different weather conditions as depicted in Figure 1.

Figure 1. Example images of the UA-DETRAC dataset [21] that possesses various challenges such as varying daytimes, different weather conditions, various perspectives, occlusions, different object types and object scales.

In recent years, object detection has achieved significant improvements with the advances in deep learning. Deep learning based detection frameworks such as Faster R-CNN [19] and SSD [14] have been employed for different domains including traffic surveillance. Extensions of these detection frameworks, *e.g.* exploiting geometric proposals to encode the scene layout of a static camera [1], have been proposed to improve the detection accuracy. However, there is still potential for improvement especially in case of adverse weather conditions [15].

In this paper, we propose an ensemble of two deep learning based detectors for more robust vehicle detection. For this purpose, we employ RefineDet [22] as detection framework, which combines advantages of both Faster R-CNN and SSD. As the base network has a large impact on the detection accuracy, we use SENets [12], which achieved the first rank in the ImageNet2017 classification challenge, besides ResNet-50 [11] as base network. To account for the characteristics of traffic surveillance data, *e.g.* small

978-1-5386-9294-3/18/$31.00 ©2018

vehicles in the background and strongly varying vehicle scales, we apply multi-scale testing as proposed in [22]. We clearly outperform the leading object detectors on the UA-DETRAC dataset [21] especially in case of rainy and nighttime scenarios. We further analyze the impact of our proposed detector on the tracking performance of the IoU Tracker [2] that won IWT4S Challenge on Advanced Traffic Monitoring 2017 [15]. To reduce the number of fragmented tracks and ID switches, we combine the IoU Tracker with a Kalman filter.

The remainder of this paper is organized as follows. In Section 2, we give an overview about deep learning based detection methods and vehicle detection in traffic surveillance data. In Section 3, we describe our proposed ensemble of vehicle detectors in detail. The performed experiments and results are discussed in Section 4. We conclude in Section 5.

## 2. Related Work

In recent years, a large variety of deep learning based detection frameworks has been proposed in literature [23]. In general, these detection frameworks can be distinguished into region proposal based frameworks and regression/classification based frameworks [23]. Region proposal based frameworks like R-CNN [9], Fast R-CNN [8], Faster R-CNN [19], R-FCN [4], FPN [13] and Mask R-CNN [10] typically comprise two stages. In an initial stage, candidate regions so called region proposals are generated and then classified in a subsequent classification stage. Regression/classification based frameworks such as MultiBox [6], YOLO [16], YOLOv2 [17], YOLOv3 [18], SSD [14] and DSSD [7] perform classification and localization in a single stage. RefineDet [22] comprises the advantages of regression/classification based frameworks and region proposal based frameworks. It consists of two modules that are both regression based. The first module identifies and removes negative anchors to reduce search space and adjusts the locations and sizes of anchors used for initialization of the subsequent object detection module. The second module predicts multi-class labels and further refines the regression.

Deep learning based detection frameworks have been employed for different domains including traffic surveillance. Several modifications such as employing novel network architectures as base network have been proposed to improve the detection performance [15]. To better learn respective features for each vehicle class, multiple sub-classes of vehicles are used instead of a single vehicle class [15]. To further improve the detection accuracy, several extensions have been proposed. For instance, Wang *et al*. [20] extend Faster R-CNN by applying an additional fine-tuning network that refines the regression and class prediction results. Amin *et al*. [1] extend Faster R-CNN by exploiting geometric proposals to encode the scene layout of a static camera. The GP-FRCNN achieves the first rank in IWT4S Challenge on Advanced Traffic Monitoring 2017 [15].

## 3. Methodology

In the following section, we describe our proposed ensemble of two deep learning based detectors. At first, we introduce the fundamental principle of RefineDet, which is used as base detection framework. Then, we present the network architectures employed as base network. Finally, we introduce multi-scale testing and the combination of both detectors.

### 3.1. RefineDet

RefineDet [22] is comprised of two modules: the *Anchor Refinement Module* (ARM) and the *Object Detection Module* (ODM). Both modules are connected via so called called *Transfer Connection Blocks* (TCBs).

The functional principle of the ARM is similar to the original SSD. The output of multiple convolutional layers is used as feature maps. To predict class scores and bounding box offsets for a fixed set of anchor boxes, a set of convolutional filters is applied on each feature map. In contrast to SSD, the ARM differentiates only between object and non-object (background). The ARM learns to coarsely adjust the anchor location and size. Furthermore, anchors that are well classified as background are filtered out, which results in a reduced search space for the subsequent ODM.

The TCBs are used to transfer features from the ARM to the ODM. Furthermore, the TCBs are used to integrate large-scale context by adding high-level features to the transferred features to improve the detection accuracy. For this, deconvolutional layers are applied to up-sample features of deeper layers.

The ODM takes the refined anchors as input to further improve the regression and to predict multi-class labels. For this, a set of convolutional filters is applied to the features maps of the ODM, which are outputs of the TCBs. Different to SSD, the refined anchors of the ARM are used as reference for the bounding box regression instead of the predefined set of anchor boxes.

### 3.2. Base Networks

The employed base network is substantial for the detection performance as the output of multiple convolutional layers is used as feature maps. For our proposed ensemble of two detectors, we consider two different network architectures: ResNet-50 [11] and SE-ResNeXt-50 [12]. ResNet-50 belongs to the residual networks, which are used by default as base network for RefineDet.

SE-ResNeXt-50 belongs to the recently proposed SENets that generalize well across challenging datasets [12]. Thus, these networks seem promising for

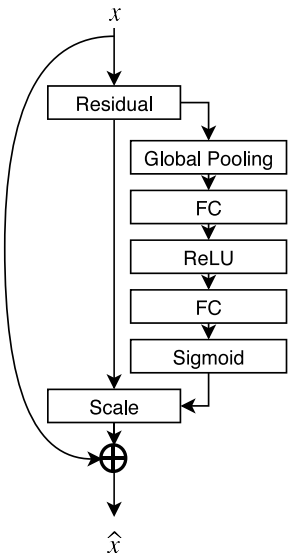


Figure 2. Squeeze-and-Excitation module [12].

traffic surveillance data with strongly varying weather and lightning conditions. *Squeeze-and-Excitation* (SE) blocks are directly applied beyond standard convolutions or in case of residual networks used as non-identity branch of a residual module as depicted in Figure 2. By explicitly modeling the interdependencies between channels, SE blocks adaptively recalibrate feature responses so that informative features can be emphasized and less useful features can be suppressed.

### 3.3. Multi-Scale Testing and Ensemble of Detectors

To account for small vehicles in the background and strongly varying vehicle scales, we employ multi-scale testing [22]. Each input image is scaled to multiple aspect ratios. Detection is performed on each scaled image as well as a vertically flipped version of it. Detections overlapping with one another by an IoU of $0.45$ or higher are combined via averaging. The combined detections are used as final detections. The same combination of detections is done in a second step between the detection results of the two detectors with different base networks.

## 4. Experimental Results

In this section, we evaluate the detection performance of our proposed ensemble of two detectors. First, we introduce the UA-DETRAC dataset that is used for our experiments. We analyze the impact of the employed base network, multi-scale testing and the use of an ensemble of detectors. Finally, we compare our detector to state-of-the-art detectors and analyze the impact on tracking results.

For our experiments, we use average precision (AP) as evaluation metric as given by the UA-DETRAC Evaluation Protocol [21]. A detection is considered as true positive, if the correct class label is predicted and the Intersection over Union (IoU) between the ground truth (GT) annotation and the predicted bounding box is above 0.7.

| Class | Number of Vehicles | |
|---|---|---|
| | Train | Validation |
| All | 147.1k | 65.8k |
| Car | 124.4k | 54.9k |
| Van | 14.1k | 5.8k |
| Bus | 7.7k | 4.8k |
| Other | 886 | 196 |

Table 1. Class distribution of the train and validation set.

### 4.1. UA-DETRAC Dataset

The UA-DETRAC dataset [15] consists of 60 video sequences for training and 40 video sequences for testing. The sequences are recorded at 25 frames per second. The train and test sequences comprise 83,791 and 56,340 frames respectively with an image resolution of 960×540 pixels. Annotations are available for the train sequences. The annotations include four classes: *car*, *bus*, *van* and *other*. As depicted in Figure 1, the dataset comprises videos taken at various weather conditions and daytimes with large variations in object type, scale and pose.

For our preliminary experiments, we split the training sequences into a training and validation set. The validation set comprises following sequences: MVI_20065, MVI_40162, MVI_40211, MVI_40963 and MVI_63563. The sequences of the validation set are selected such that different scenarios are covered. The remaining sequences are used for training. To reduce the required memory during training, we use image tiles of size 576×576 pixels. Neighboring tiles overlap by 184 pixels. Furthermore, we consider only every 5th frame due to redundancy of subsequent frames and we black out the ignore regions. Table 1 shows the class distribution of our train and validation set.

### 4.2. Implementation Details

For our experiments, we use the publicly available RefineDet framework[1]. All trainings are performed on a single NVIDIA Tesla P40 GPU. All models are trained end-to-end for 70k iterations with an initial learning rate of 0.01. The learning rate is decreased by a factor of 10 after 40k, 50k and 60k iterations. The *batch_size* is set to 3 and the *iter_size* is set to 6. Weights pre-trained on ImageNet are used for initialization of both base networks. In case of using ResNet-50 as base network, *res3d*, *res4f*, *res5c* and *res6* are used as feature maps for the ARM. In case of using SE-ResNeXt-50 as base network, *conv3_4*, *conv4_6*, *conv5_3* and *conv6* are used as feature maps for the ARM. The minimum anchor sizes are set to 32, 64, 128, and 256, and the maximum anchor sizes are set to 48, 96, 192, and 384.

[1] https://github.com/sfzhang15/RefineDet

| Base Network | AP (in %) |
|---|---|
| ResNet-50 | 93.52 |
| ResNet-50+ | 93.94 |
| SE-ResNeXt-50 | 94.12 |
| SE-ResNeXt-50+ | 94.62 |
| Ensemble | **94.95** |
| Ensemble+ | 94.44 |

Table 2. Average precision (AP) on the validation set @ IoU 0.7. + indicates multi-scale testing.

| Base Network | mAP | car | bus | van | other |
|---|---|---|---|---|---|
| ResNet-50 | 82.87 | 89.11 | 92.21 | 80.65 | 69.49 |
| ResNet-50+ | 83.14 | 91.17 | 85.77 | 82.07 | 73.54 |
| SE-ResNeXt-50 | 82.57 | 89.80 | **95.02** | 80.86 | 64.59 |
| SE-ResNeXt-50+ | 83.44 | **91.22** | 91.24 | 82.41 | 68.88 |
| Ensemble | 84.95 | 91.01 | 94.93 | 82.50 | 71.35 |
| Ensemble+ | **85.58** | 91.13 | 92.66 | **83.63** | **74.88** |

Table 3. AP (in %) classes car, bus, van and other as well as the mean AP on the validation set @ IoU 0.7. + indicates multi-scale testing.

| Base Network | mAP | car | bus | van | other |
|---|---|---|---|---|---|
| ResNet-50 | 89.00 | 97.02 | 97.44 | 83.43 | 78.12 |
| SE-ResNeXt-50 | 88.43 | 97.05 | 97.56 | 84.46 | 74.64 |
| Ensemble | **90.30** | **97.45** | **97.76** | **85.37** | **80.61** |

Table 4. AP (in %) classes car, bus, van and other as well as the mean AP on the validation set @ IoU 0.5.

### 4.3. Preliminary Results

Table 2 shows the AP for both base networks and for the ensemble on our validation set. Following the UA-DETRAC Evaluation Protocol [21], all vehicle categories are summarized into one single class. Using SE-ResNeXt-50 as base network shows better results compared to ResNet-50. The best results are achieved for the ensemble of both detectors due to less false negative detections. We further analyzed the impact of multi-scale testing. For this purpose, we rescaled the images by factor 0.5, 0.75, 1.5 and 1.75, respectively. The AP is clearly improved in case of the single detectors due to a decreased number of false negative detections caused by small vehicles. However, the AP is worse in case of the ensemble of both detectors.

To analyze the AP in more detail, we evaluate each detector for multiple vehicle classes. Table 3 shows the AP for the classes *car*, *bus*, *van* and *other* as well as the mAP. SE-ResNeXt-50 achieves better AP for the classes *car*, *bus* and *van* while ResNet-50 achieves the best AP for class *other*. However, class *other* consists of considerably less object instances compared to the other classes as given in Table 1. The best AP without multi-scale testing for the classes *car*, *van* and *other* is achieved for the ensemble of both detectors. Employing multi-scale testing results in better AP for all detectors and all classes except for class *bus*. The worse AP in case of class *bus* is caused by tiling of the input image as large vehicles (mainly buses) can result in split detections, which is emphasized by the tiling. The AP in case of single classes is clearly worse compared to the AP for all classes summarized into a single class as more false positives and false negatives are caused by incorrect class predictions.

We further analyze the localization quality of the single detectors and the ensemble of both detectors. For this, we consider a detection as true positive, if the IoU between the GT annotation and the predicted box is above 0.5. As shown in Table 4, the AP is clearly higher for all detectors and all classes especially for class car that contains a huge number of small object instances. The IoU criterion is more sensitive to small object instances as small deviations between predicted box and GT annotation are more likely to result in missed detections. Thus, the number of missed detections in case of small and only partially visible object instances is clearly reduced. The high AP values in case of the less strict IoU criterion indicate that the localization accuracy has potential for further improvements.

### 4.4. Comparison to Leaderboard

Table 5 shows the comparison of our proposed ensemble with the current leader of the UA-DETRAC detection challenge [2] and the top-performing detectors of the IWT4S Challenge on Advanced Traffic Monitoring 2017 [15]. Our proposed ensemble of two detectors with multi-scale testing clearly outperforms all other detectors on the complete test set (overall). Furthermore, using the full images for training and testing results in clearly improved AP as the number of split detections due to the tiling is reduced. Compared to GP-FRCNN [1] – the winner of the IWT4S Challenge on Advanced Traffic Monitoring 2017 – the AP is improved by 8.78 percentage points. The UA-DETRAC dataset comprises sequences with three levels of difficulty. We clearly improve the AP for sequences that are rated as hard. The dataset further comprises four subsets for different weather conditions. Our proposed approach achieves the best AP in case of rainy scenarios and during night. Table 5 further gives the number of frames per second. For this, we performed testing on two NVIDIA Titan X GPU. The number of frames per second is less compared to the other detectors due to the multi-scale testing. In case of single-scale testing, the number of frames per second is more than doubled compared to SSD_VDIG and HAVD, which are the current leader of the UA-DETRAC challenge.

[2] Results taken from `http://detrac-db.rit.albany.edu/DetRet`, 09/04/2018

| Method | Overall | Easy | Medium | Hard | Cloudy | Night | Rainy | Sunny | Speed |
|---|---|---|---|---|---|---|---|---|---|
| SSD_VDIG | 82.68 | 94.60 | 89.71 | 70.65 | **89.81** | 83.02 | 73.35 | 88.11 | 2 |
| HAVD | 80.51 | 94.48 | 86.13 | 69.02 | 87.28 | 82.30 | 69.37 | 89.71 | 2.1 |
| DCN | 79.85 | 93.85 | 85.07 | 69.00 | 85.55 | 82.38 | 68.95 | 89.08 | 8 |
| GP-FRCNN [1] | 76.57 | 91.79 | 80.85 | 66.05 | 81.23 | 77.20 | 68.59 | 85.16 | 4 |
| EB [20] | 67.99 | 87.77 | 73.03 | 54.74 | 75.13 | 71.80 | 52.99 | 82.04 | **11** |
| Ensemble | 81.80 | 94.42 | 86.00 | 72.37 | 87.14 | 81.76 | 74.07 | 88.97 | 5 |
| Ensemble+ | 83.45 | 94.63 | 87.95 | 74.31 | 87.73 | 85.30 | 75.96 | 89.27 | 1 |
| Full Image Ensemble | 83.74 | **96.03** | 88.67 | 73.88 | 88.61 | 84.05 | 76.48 | 89.79 | 5 |
| Full Image Ensemble+ | **85.35** | 95.80 | **89.84** | **76.64** | 89.67 | **86.59** | **78.17** | **90.49** | 1 |

Table 5. Comparison to current leader of the UA-DETRAC detection challenge and to the top-performing detectors of the IWT4S Challenge on Advanced Traffic Monitoring 2017.

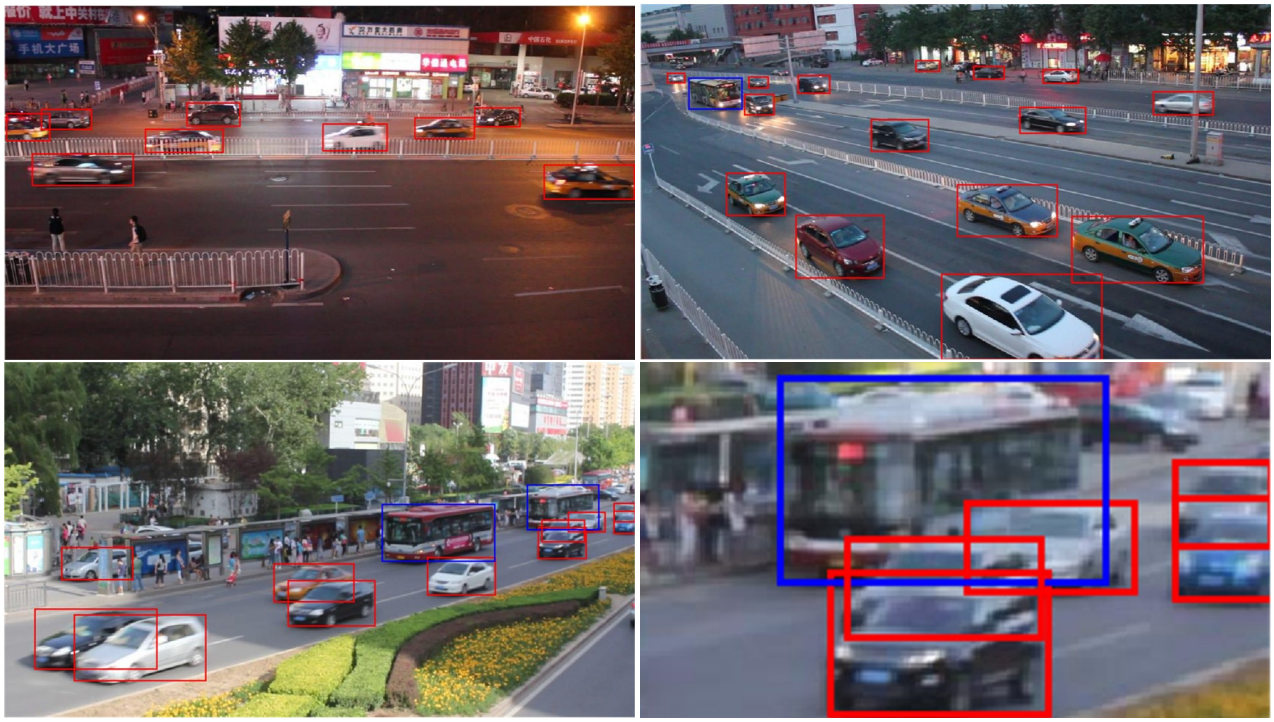

Figure 3. Qualitative detection results for class car (red boxes) and class bus (blue boxes) on the test set. The zoomed in region (bottom right) shows that even strongly occluded vehicles in dense traffic are correctly detected.

### 4.5. Qualitative Results

Figure 3 illustrates qualitative detection results for class car (red boxes) and class bus (blue boxes) on the test set. The predicted bounding boxes are tight around the corresponding vehicles and correctly classified for challenging scenarios with high traffic density or with poor lightning conditions, *e.g.* during night. Even strongly occluded vehicles are correctly detected as shown in the zoomed in region (bottom right).

### 4.6. Tracking Results

We further analyze the impact of our proposed detector on tracking results. For this, we use the IoU-Tracker (IOUT) [2] that won the IWT4S Challenge on Advanced Traffic Monitoring 2017 and a combination of the IOUT with a Kalman filter. We adopt the parameter settings given in [2]. For comparison, we use three object detection algorithms: ACF [5], R-CNN [9] and CompactACT [3][3]. We apply the UA-DETRAC Evaluation Protocol for evaluation. Table 6 shows the tracking results on the validation sequences. Using our proposed detector results in clearly improved PR-MOTA , PR-MOTP, PR-MT and PR-ML, whereas the number of ID switches (PR-IDS) is not reduced. Combining the IOUT with a Kalman filter further improves the tracking results as the number of fragmented tracks and ID switches are considerably reduced.

[3]Detections for all three detection methods on the UA-DETRAC dataset are taken from `http://detrac-db.rit.albany.edu/Tracking`, 20/03/2018

| Method | PR-MOTA | PR-MOTP | PR-MT | PR-ML | PR-IDS | PR-Frag | PR-FP | PR-FN | FPS |
|---|---|---|---|---|---|---|---|---|---|
| IOUT+CompACT [3] | 23.47 | 42.56 | 14.60 | 20.20 | 241.96 | 252.21 | **734.21** | 14384.28 | **1373** |
| IOUT+ACF [5] | 24.67 | 44.16 | 15.62 | 20.99 | 208.65 | 243.40 | 1120.48 | 15299.35 | 1355 |
| IOUT+R-CNN [9] | 26.14 | 45.03 | 18.45 | 19.38 | 489.30 | 507.70 | 1129.40 | 15048.48 | 1172 |
| IOUT+Ours | 39.31 | **52.96** | 42.57 | 8.03 | 300.97 | 251.81 | 4625.64 | 7072.76 | 1188 |
| KF-IOU+R-CNN | 28.12 | 43.54 | 20.00 | 18.84 | 91.81 | 162.73 | 1145.23 | 14333.89 | 183 |
| KF-IOU+Ours | **40.08** | 52.85 | **42.95** | **7.82** | **90.46** | **61.81** | 4684.23 | **6797.57** | 153 |

Table 6. Tracking results on the five validation sequences. We employ the publicly available source code of the IOUT [2]. The detections of CompACT, ACF and R-CNN are provided by the UA-DETRAC challenge.

## 5. Conclusion

In this paper, we have proposed an ensemble of two detectors for vehicle detection in traffic surveillance data. We employ RefineDet as detection framework as it combines advantages of both SSD and Faster R-CNN. In addition to ResNet-50, we employ the recently proposed SE-ResNeXt-50 as base network as it shows superior generalization abilities. We show that the ensemble of both detectors outperform the single detectors as less vehicles are missed. We achieve top performing results on the UA-DETRAC dataset. Current state-of-the-art detectors for traffic surveillance are outperformed especially for hard, nighttime and rainy scenarios.

## References


[1] S. Amin and F. Galasso. Geometric proposals for faster r-cnn. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.

[2] E. Bochinski, V. Eiselein, and T. Sikora. High-speed tracking-by-detection without using image information. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.

[3] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015.

[4] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[5] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014.

[6] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *CVPR*, 2014.

[7] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[8] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[12] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. 2018.

[13] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[15] S. Lyu, M.-C. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. Del Coco, et al. Ua-detrac 2017: Report of avss2017 & iwt4s challenge on advanced traffic monitoring. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–7. IEEE, 2017.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[17] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017.

[18] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[20] L. Wang, Y. Lu, H. Wang, Y. Zheng, H. Ye, and X. Xue. Evolving boxes for fast vehicle detection. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 1135–1140. IEEE, 2017.

[21] L. Wen, D. Du, Z. Cai, Z. Lei, M. Chang, H. Qi, J. Lim, M. Yang, and S. Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *arXiv CoRR*, abs/1511.04136, 2015.

[22] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018.

[23] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. Object detection with deep learning: A review. *arXiv preprint arXiv:1807.05511*, 2018.