Electronic Letters on Computer Vision and Image Analysis 19(3):98-117, 2020

Scale-Invariant Mask R-CNN for Pedestrian Detection

Ujwalla Gawande* , Kamal Hajari* and Yogesh Golhar+

* *Associate Professor, Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India*
* *Research Scholar, Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India*
+ *Assistant Professor, Department of Computer Science and Engineering, G. H. R. I. E. T, Nagpur, India*

Received 2nd of August 2020; accepted 19th of September 2020

Abstract

Pedestrian detection is a challenging and active research area in computer vision. Recognizing pedestrians helps in various utility applications such as event detection in overcrowded areas, gender, and gait classification, etc. In this domain, the most recent research is based on instance segmentation using Mask R-CNN. Most of the pedestrian detection method uses a feature of different body portions for identifying a person. This feature-based approach is not efficient enough to differentiate pedestrians in real-time, where the background changing. In this paper, a combined approach of scale-invariant feature map generation for detecting a small pedestrian and Mask R-CNN has been proposed for multiple pedestrian detection to overcome this drawback. The new database was created by recording the behavior of the student at the prominent places of the engineering institute. This database is comparatively new for pedestrian detection in the academic environment. The proposed Scale-invariant Mask R-CNN has been tested on the newly created database and has been compared with the pedestrian benchmark databases. The experimental result shows significant performance improvement in pedestrian detection as compared to the existing approaches of pedestrian detection and instance segmentation. Finally, we conclude and investigate the directions for future research.

Key Words: Convolutional Neural Network, Instance Segmentation, Pedestrian Detection, Mask R-CNN.

1 Introduction

Pedestrian or human detection is an imperative and integral process of the intelligent video surveillance system. It is an important step for perceiving semantic information in the environment. Pedestrian detection helps in providing fundamental information in a wide range of applications such as elder human surveillance, patients surveillance in hospitals, prisoners surveillance in a prison, human surveillance in ATM and Banks, driverless cars, crowd event monitoring, pedestrian counting, road event detection, pedestrian traffic survey in intelligent transportation, etc. However, because of the diverse change in human pose and appearance pedestrian detection accuracy gets affected. Pedestrian detection in high-density areas (railway station, shopping malls, airports, etc.), non-uniformed illumination, complex background, pedestrian deformation, occlusion, shadow, etc. continue to be an unsolved problem in this area.

Correspondence to: <ujwallgawande@yahoo.co.in>

Recommended for acceptance by <Angel D. Sappa>
https://doi.org/10.5565/rev/elcvia.1278
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

In recent years pedestrian detection demand has been increased. Many researchers focused is on the enhancement of pedestrian detection performance using deep neural network frameworks. Figure 1 shows the timeline chart of the evaluation of deep learning neural networks for pedestrian detection. The state-of-the-art convolutional deep neural networks, semantic and instance segmentation frameworks, such as Mask R-CNN [6][7], You Only Look Once (YOLO) [8], Region-based Fully Convolutional Network (R-FCN) [9], Single Shot MultiBox Detector (SSD) [10], Fully Convolutional Network (FCN) [11], Faster R-CNN [12][13], Fast R-CNN [14], Deep Convolutional Generative Adversarial Network (DCGAN) [15], Residual Neural Network (ResNet) [16], GoogLeNet [17], Visual Geometry Group (VGG Net) [18], ZFNet [19], AlexNet [20], Deep Belief Network (DBN) [21], LeNet [22], etc. are used for the pedestrian detection.

These frameworks significantly outperform the traditional Support Vector Machine (SVM) [23], AdaBoost [24], Probabilistic Neural network (PNN) [25], Radial basis Neural Network (RBN) [26], Artificial neural network (ANN) [27], etc. pedestrian detection approaches.

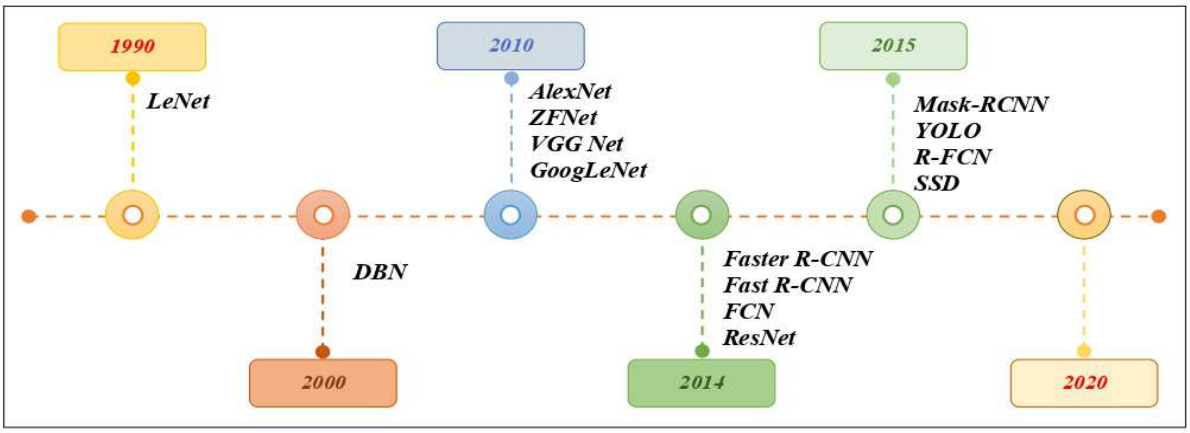

Figure 1: A timeline chart of the evaluation of deep neural network frameworks for pedestrian detection.

The first CNN architecture LeNet has been introduced by LeCun et al. [22] in 1998. It includes only two convolutional layers, a pooling layer along with the backpropagation network for training. LeNet uses the $tanh$ activation function for deciding whether an input layer neuron should be activated or not by computing the cumulative sum of weights and bias. It uses the Mean Squared Error (MSE) function for estimating the average square difference between the computed parameter and the actual parameter. As it consists of two convolutional layers that indicate it was used for the small training database and having less processing and computational capabilities. LeNet has trained on the Modified National Institute of Standards and Technology (MNIST) database with 50,000 images divided into 10 categories. It was successfully used commercially for object detection and handwritten signature detection at that time. It had an error rate of 26.2%.

Next, AlexNet has been introduced by Krizhevsky et al. [20] in 2012. It was very similar to LeNet. AlexNet uses the ReLU activation along with Cross entropy loss function. It consists of five convolutional layers with a similar combination of pooling and backpropagation networks as LeNet. As AlexNet is having more convolution layers it was trained on the ImageNet large database consists of more than 1 million images from 1000 categories. AlexNet object detection accuracy is more than LeNet and used for different types of object detection purposes. Its error rate reduces to 15.4%. The modified version of AlexNet was ZFNet introduced by Zeiler et al. [19] in 2013. ZFNet uses 7 x 7 filters in the first convolutional layer instead of 11 x 11 as used in AlexNet. The low resolutions features are extracted by reducing the smaller size filter in convolution layers. It achieves an error rate of 11.2%.

Liu et al. [21] introduced, unsupervised Deep Neural Network architecture in 2009. DBN consists of multiple layers along with multiple feature detectors or hidden units. These hidden units help to establish the correlation between input data. The greedy learning algorithm is employed for the training of DBN. It uses

the layer by layer architecture in a top-down manner. In the training phase, adjacent layers weight depends on each other called generative weight. The disadvantage of this network is the fact that it increases the run-time complexity. The pre-trained model does not adapt to the newly input complex data. However, this architecture pattern is popular as much as AlexNet, VGGNet.

Simonyan et al. [18] introduced VGGNet in 2014, which reduces the error rate to 7.2%. It expanded the number of convolutional layers to 19 layers. The filter size reduces 16 times was limited to 3 x 3. A similar architecture pattern has been used in the next architecture i.e. GoogLeNet and ResNet. Vanhoucke et al. [17] [18] introduce the Google Inception network with VGGNet in 2014. It achieves the error rate of 6.7% marginally better than the VGGNet at the cost of complex architecture design patterns compares to VGGNet. It uses the average pooling and design inception module for processing input data with multiple filters in parallel. The inception module uses four convolution layer filters as 1 x 1, 3 x 3, 5 x 5, and 3 x 3 with max Pooling. In the max-pooling process, maximum elements regions are selected from the region of the feature map mapped by the filters, which results in a feature map with prominent features. Inception module design helps to convolution layer numerical computation from 854 million to 358 million.

ResNet was introduced by He et al. [16] in 2015. It achieves an error rate of 3.57%. ResNet consists of 152 convolutional layers. In this architecture, the model parameters are adjusted according to input data in a specific direction, called gradient descent, due to this updation process weight was not modified in the backpropagation algorithm for training. Instead of the inception model ResNet uses the forward and backward passes of the backpropagation algorithm. In the forward pass, the few convolution layers skip and input data is processed from input to output without error. Backward passes the error gradient data return back from the output layer to the input layer of the network. This feature makes ResNet ultra-deep architecture without affecting the performance. Later frameworks of the deep neural networks are employed to the object detection and classification task using semantic segmentation. In semantic segmentation, each pixel of the image has been divided into an object class. FCN was the first deep neural network architecture used for the semantic segmentation introduced by Jonathan et al. [11] in 2015. In previous CNN architecture, an input image is going through the convolutional layers and fully connected layers. The output is the predicted class label for the input image. In FCN fully connected layers are converted into 1 x 1 convolution layers. After, the convolution output image size is reduced compared to the input image. Fusing steps are performed after max-pooling which is similar to previous techniques such as AlexNet, GoogLeNet, and VGGNet. In this process, multiple convolution model outputs have been fused for a more accurate prediction. Next, the convolved image is upsampled or it is also known as transpose convolution to get the image size larger in the output. FCN architecture includes both the process i.e. convolution and upsampling. The upsampled processed output image is the pixel-wise segmented image. The problem with these approaches is the selection of a region before convolution. The object in an image may have different spatial locations and aspect ratio. The huge region needs to be select to get the object of interest in an image. Therefore, an algorithm like YOLO, R-CNN has been developed to solve the problem of selection of regions that consist of an object of interest.

R-CNN has been proposed by Ross Girshick et al. [28] where a selective search algorithm is used for finding the Region of Interest (RoI) in an image, called region proposals. Selective search algorithm extract 2000 regions from the image. In this algorithm, the first step is to generate sub-segmented candidate regions. Each segment size and color features have been obtained here. Regions with similar information have been combined to obtain the larger region. The greedy algorithm is then is subsequently applied recursively to generate the region of interest. CNN features are extracted from the region of the 4096 dimensions. Then SVM has been applied to classify the availability of the object in the region. The classification results are represented by the bounding boxes in the output. The problem with R-CNN is that 1) it required a huge amount of time to train the 2000 region proposals in the image. 2) it cannot be used for real-time applications. Because it takes 47 seconds for testing the image. 3) the selective search algorithm is fixed, not adaptive, so sometimes generate bad regions proposals for the complex image. The drawbacks of R-CNN are solved in Fast R-CNN.

Ross Girshick et al. [14] again proposed the enhanced version of R-CNN i.e. Fast R-CNN. It is similar to the R-CNN but instead of inputting the region proposals to CNN, this architectures input image is feed into CNN

to generate a convolved feature map. These feature maps region of proposals have been identified by using RoI pooling. In this process, max pooling on the input feature map has been performed that generates a fixed size 7 x 7 square region. This region feeds into the fully connected Softmax layer to predict the class of the object. Results have been represented using the bounding box. The Fast R-CNN is significantly efficient and faster than the R-CNN because there is no need for 2000 region proposal training to classify the object. Both the architecture uses the selective search to find region proposals using a selective search algorithm. This process is slow and time-consuming that affects the performance of the network.

Shaoqing Ren et al. [12][13] introduced a Faster R-CNN to solve the Fast R-CNN speed issue by eliminating the selective search algorithm. In this architecture, the network itself learns the region proposals instead of using a selective search algorithm to identify region proposals. A separate network has been used for the identification region proposals in the convolved feature map. Hence, it is faster than the Fast R-CNN and it can be used in the real-time applications for object detection. In all the previous versions of the R-CNN family, regions are utilized to the localization of objects in the image.

Joseph Redmon et al. [8] introduced a new architecture i.e. YOLO. In this, a single convolutional layer network predicts the bounding box and class label for these boxes. The input image has been divided into the S x S grid and each grid, m bounding boxes have been generated. For each of the bounding boxes, the convolutional network generates the class probability and bounding box location values which are used for locating the object within the image. YOLO processed 45 frames per second which are faster than the other object detection algorithm. The problem with YOLO is that it does not detect small objects within the image, because of the spatial plane coordinate location constraints of the algorithm.

The semantic segmentation algorithms classify the object at the pixel level. Recent architecture, i.e. Mask-RCNN has been used for instance segmentation. The process of identifying each instance of the individual object within the image and locating each instance pixel is called as instance segmentation. However, the instance segmentation is difficult, as it needs the accurate localization and detection of all the available moving and non-moving objects in an image. Hence, it combines object detection and semantic segmentation techniques. Here object detection, object classification, and its representation are done using bounding boxes. Whereas, in the semantic segmentation, each pixel is classified into a meaningful group, without distinguishing the object instances. Figure 2 shows examples of instance segmentation.

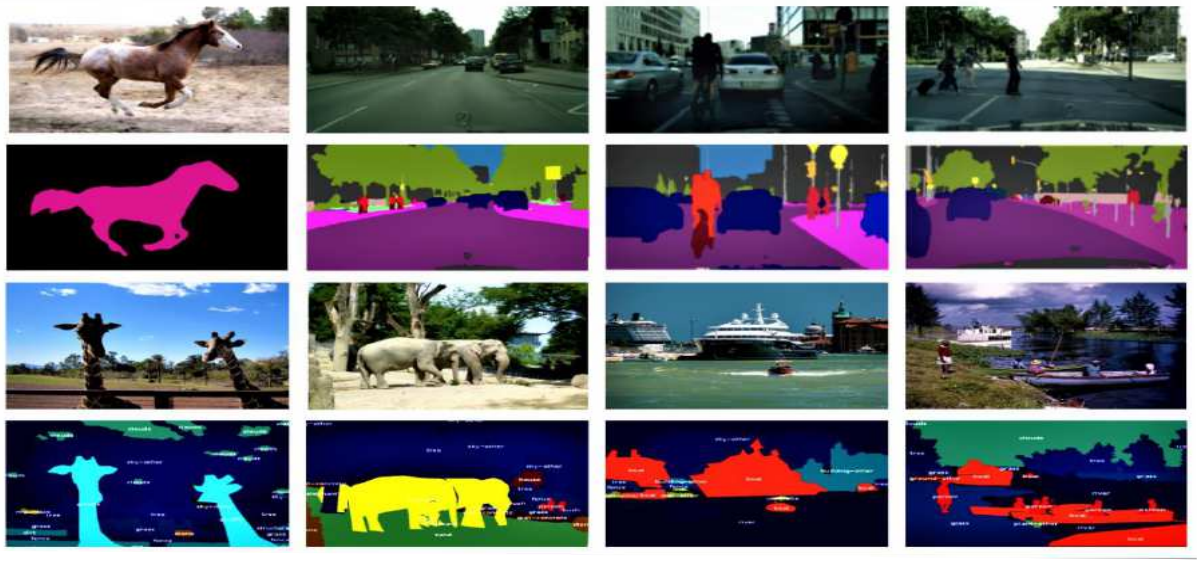

Figure 2: Example of instance segmentation. First and third row from left to right (original sample images of the COCO database [3]. Second row from left to right (Semantic segmentation using Waterfall Atrous Spatial Pooling Architecture [29]). Fourth row from left to right (Instance segmentation using Mask-CNN [6][7]).

Mask R-CNN has been introduced by He. Kaiming et al. [6][7] in 2017, for detection and instance segmentation of the object in an image. It is an extended version of Faster R-CNN [12][13], Fast R-CNN [14], and FCN [11] deep learning frameworks. Mask R-CNN has been applied to single and multiple object detection, classification, and semantic segmentation respectively. Here, RoI has been identified using a selective search mechanism, and then high-level features are extracted for classification using the SVM. The motivation for the proposed work is shown in Figure 3. Pedestrian instances in Caltech [1], INRIA [2], MS COCO [3], ETH [4], and KITTI [5] database frequently have small size. Localizing these small instances are a challenging tasks due to issues like 1) hazy appearance, 2) blurred and unclear boundary, 3) overlapped pedestrian instances, 4) small and large size instances having different characteristics, etc.

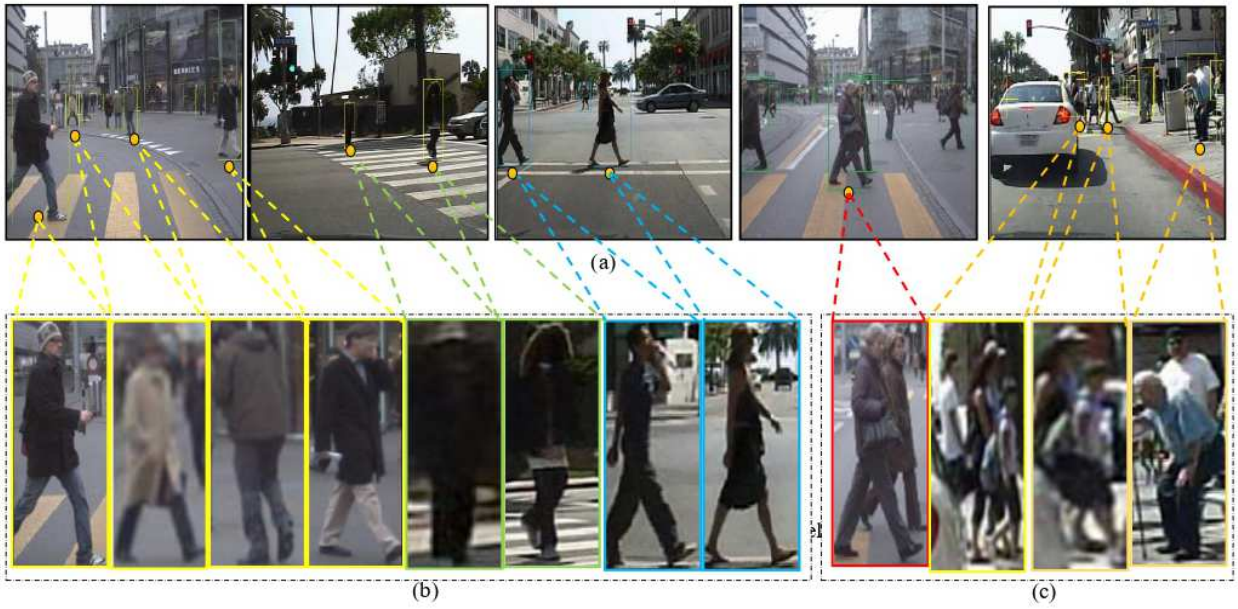

Figure 3: Illustration of the motivation of the proposed Scale-invariant Mask R-CNN (SIM R-CNN) model. (a) shows some example pedestrian images of Caltech [1], ETH [2] database. (b) demonstrate large and small size pedestrian instances visual appearance are significantly different. (c) occluded, blurred, unclear, the hazy appearance of pedestrian instance make the small size pedestrians instances difficult to detect.

Ross Girshick et al. introduced Fast R-CNN [14] to address the problem of the scale-variance problem by scaling the input image using brute-force data at the cost of time and computational complexity. Yunchao Gong et al. [30] presented a single model with different multi-scale filters on all the objects of various sizes. However, due to the heterogeneous size of instances, these approaches are computationally complex and require more time to detect the object of different sizes. The state-of-the-art approach for handing scale-variance approach named as Scale-Aware Fast R-CNN (SAF R-CNN) framework has been introduced by Jianan Li et al. [31]. In this approach, the divide and conquer strategy has been adopted with the Fast R-CNN to resolve the scale variance problem.

The SAF R-CNN uses a large-size and small-size sub-network for each instance. The confidence score has been generated for each instance for the detection of the object. The scale-aware weighted layer assigns the higher weight in large-size sub-network and lowers size weight in small sub-network for robust detection of the instances at various scales. The limitation of this approach is 1) irrespective of the size of instances in the proposal, both the sub-network computes the confidence score for each instance, which increases the time for the identification of the object. 2) the small and large size instances required independent training and testing. The proposed SIM Mask R-CNN addresses the existing method issues by using the scale-invariant feature map generation. The key idea of Scale-invariant feature map generation is illustrated in Figure 4.

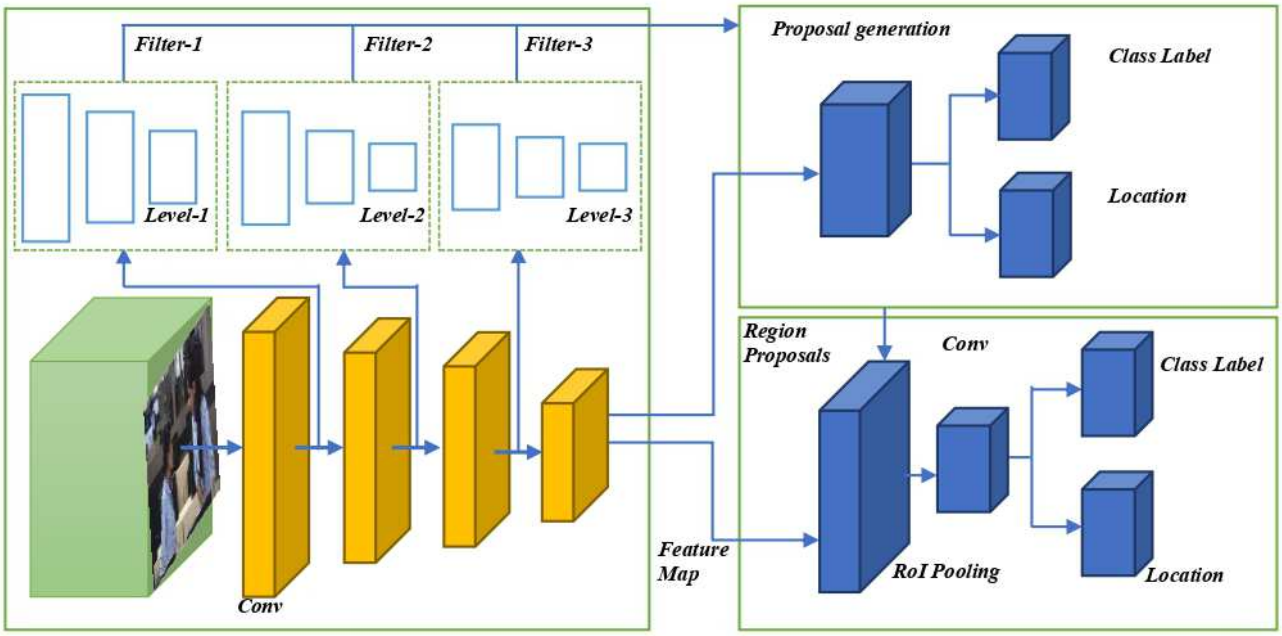

Figure 4: Illustration of the proposed SIM Mask R-CNN. A multiple scaled image is filtered with multiple pre-defined boxes filters separately to detect instances of different sizes. The final output is obtained by fusing the results of two-stage detectors and the scale-invariant feature map according to the object proposal size.

Motivated by the above idea, a novel SIM R-CNN framework has been proposed, which is built on the Faster R-CNN pipeline [12]. The proposed SIM R-CNN integrates a scale-invariant feature map with the two-stage backbone network into a unified architecture. As shown in Figure 4, given an input image with object proposals in it, the SIM R-CNN first passes the raw image through the bottom shared convolutional layers to extract its whole feature maps. Taking these feature maps and the different size filtered confidence score, combinedly generates the final detection results, defined over the proposal size. The conclusive results can always be boosted by the scale-invariant feature map and two-stage backbone network, appropriate for the current input of certain scales. Therefore, SIM R-CNN can achieve outperforming detection performance in a wide range of input scales. Moreover, since SIM R-CNN shares convolutional features for the whole image with different object proposals, it is very efficient in terms of both, training and testing time.

Mask R-CNN has the advantage of predicting the binary mask for each specific detected object which was not available in Faster-RCNN and FCN. The objects are identified and classified in these networks, but the segmentation mask at the pixel level is not generated and individual objects are not distinguished accurately. Mask R-CNN gains a considerable amount of attention in terms of object detection accuracy. In general, it divides into two-parts as training and testing. The Mask R-CNN be trained on the MS COCO dataset using a pre-defined weighted model. Its testing part involved two major steps i.e. proposal region generation and object classification respectively. The proposal region generation step generates a RoIs, which may or may not contain the desired object. In the classification step, each RoIs are classified either as an object or the background. However, despite its outstanding performance in terms of object detection accuracy, Mask R-CNN is computationally costly for the new image as input to the system. The proposal region generation procedure requires a time to generate the RoIs. Again, the smallest objects are not classified efficiently due to the unavailability of the scaling factor for the objects. These two drawbacks restrict the use of the Mask R-CNN in a real-world application for object detection in video surveillance.

The main objective of this research paper is to propose a method to reduce the computational cost of the pre-trained Mask-RCNN in the testing phase and to detect the small size object efficiently. The key thought of

the proposed method is to modify the version of Region Proposal Network (RPN), to generate a feature map with a scale-invariant feature map that will provide additional object scale information. In particular, feature maps of the network can be utilized to locate the object in an image. Hence, proposal regions for a new input image could be generated with scale aware information in the trained network, which results in efficient small object detection in less computational cost. The detection accuracy increases due to the availability of object information at multiple scales. Joint feature maps of multiple resolutions are extracted from different layers of the two networks and are used for pedestrian detection, which results in a low false-positive rate. Experimental results show improvements in detection rates due to the scale-invariant feature map.

The rest of the research paper is organized as follows. In section 2, related work in terms of pedestrian detection and semantic segmentation is described with the limitations of each method. Sections 3, present a proposed pedestrian detection framework. Section 4, presents a new pedestrian database in an academic environment and the details of experimental results and comparative analysis of the proposed method with related methods. The final section concludes with a future research direction.

2 Related Work

In literature various deep learning-based pedestrian detection methods were developed for improving the pedestrian detection accuracy [31], [32], [33], [34], [35], [36]. However, pedestrian detectors still suffer from various issues caused by a complex background in the image, scales of pedestrians in the image, pedestrian occlusion, illumination variations, etc. These issues are partially addressed, which significantly affect the performance of pedestrian detection. Most of the pedestrian detectors use the Histogram of Oriented Gradient (HOG) [2]. In this approach gradient information is used for detecting the object. The limitation of this approach is 1) partially occluded pedestrians are not detected. 2) computing HOG features are time-consuming. Hence it is not suitable for a real-time video surveillance system. Further, Xiaoyu Wanget al. [37] presented a combined approach of Local Binary Pattern (LBP) and HOG to handle partial occlusion of the pedestrian. In this texture descriptor and gradient, descriptors have been used together to detect the pedestrian in a video. This method achieves a pedestrian detection rate of 91.3% with False Positives Per Window (FPPW) = 10^{-6}, 94.7% with FPPW = 10^{-5}, and 97.9% with FPPW = 10^{-4} on the INRIA dataset.

Piotr Dollar et al. [38] proposed a combined approach of HOG descriptors with LUV (where L is used for the luminance, whereas U and V represent a chromaticity value of color in an image). Color features called as Image Channel Feature (ICF). The ICF detector has a faster computational speed compare to the HOG descriptor, as it uses the integral images over the feature channels. In this method feature pyramid generated for the nearby gradient at multiple scale results in fast feature computation in real-time. The sliding windows scheme and the hand-crafted feature-based method proposed, partially address the issues of pedestrian detection.

Hence, the researcher moves towards the region-based deep learning approaches to overcome the existing hand-crafted based system issues [39][40]. Sliding window schemes are computationally costly and require more time to detect the pedestrian. However, the region-based method uses the pedestrian candidate region, which is small in size than that of the sliding window. Detection and classification of the pedestrian have been performed by focusing the region proposals to be more time and cost-efficient. Nathan Silberman et al. [41] presented an instance segmentation using a coverage loss method for object segmentation purposes in indoor scenes. In this approach, for each image in the dataset CNN and scale-invariant feature transform (SIFT) features have been extracted. Then Ultrametric Contour Map (UCM) has been used for boundary detection. The connected components are represented in the tree structure. The tree proposal method restricts the search space of instance segmentation. Each pixel assigns a label as per the tree structure. At last, a combined approach of SVM has been used for instance segmentation. This method gives coverage upper bound (CUB) score of 64.1%. Using CNN features, it achieves a weighted coverage score of 87.4%. Coverage upper bound metric measures similarity measure between the segmented region. The comparison has been performed between instance segments and human annotators segments. The limitations of this approach are 1) due to tree structure

representation, non-neighboring regions are not combined to form a segment when the objects in a scene are occluded. 2) coverage loss function does not solve the problem of false positive. 3) object detection and instance segmentation process are slow. 4) poor performance without depth information of the scene.

Ziyu Zhangin [42] presented an instance segmentation with deep densely connected Markov Random Fields (MRF). The architecture combines patch-based CNN prediction and global MRF reasoning. In this approach, input image instances have been extracted using different size patches (270 x 432, 180 x 288, and 120 x 192). CNN has been used for assigning instance labels to each extracted patch. MRF consist of the unidirectional graphical model. Each vertex represents an instance label of each pixel. This method has several drawbacks such as 1) it is suitable only for a single type of object. 2) interconnected object consideration fails in case occluded object. 3) object detection fails in the case of small size object. This method achieves Average false Positive (AvgFP) of 83.9% and Average false negative (AvgFN) of 0.375%. If a predicted instance does not overlap with any ground truth instance, then it is a false positive instance. Similarly, if a ground-truth instance does not overlap with any prediction, then it as a false negative instance. X. Wang et al. [43] presented a Relief R2-CNN (R^2-CNN) for pedestrian extraction in the real-time environment. In this approach main focus was on the faster RoI generation from the convolutional features of trained CNN. Here the input normalized feature maps have been used for generating the RoI. A normalized feature map has been generated by dividing each feature map element by its maximum feature map value. The object detection and classification are similar to R-CNN. This approach can be useful in person re-identification. Limitations of this approach are 1) the R^2-CNN is not tested for the real-time application. 2) the classification task requires more time because R^2-CNN needs repetitive fine-tuning of object localization.

Kelong Wang et al. [32] proposes a unified joint detection framework for pedestrian and cyclist pedestrian detection. The method uses Fast R-CNN. In this approach multilayer feature fusion, and multitarget candidate region has been designed to improve detection accuracy and to solve the problems of frequent false detection and missed detection in pedestrian and cyclist target. The experiments have been conducted on the cyclist database developed in Beijings urban traffic environment. The limitation of this method is 1) the joint detection framework is not tested in a real-time environment. 2) improvement in the target detection rate of pedestrians and cyclic pedestrians could be improved and needs the verification of the target detection method in intelligent driven vehicles. Further, Miran Pobar et al. [31] proposed a combined approach of Mask R-CNN and an Optical flow-based method to determine the active football player's pedestrian in the scene.

The pedestrian detection accuracy is 85%. The limitation of this method is 1) non-active football player pedestrians have been classified as active pedestrians. 2) segmentation is manual hence, not suitable for the real-time application. 3) for a long duration, video sequence computational time is more, and the miss-classification rate increases uniformly. Jianan Li et al. [26] proposed a Scale aware Fast R-CNN (SAF R-CNN) framework for pedestrian detection. In this framework, two sub-networks are used to detect different scale pedestrian. The main disadvantage of this framework is that the training and testing time increase due to the involvement of a two-sub network and each input image processed using this two-network results in higher computation cost. The proposed SIM R-CNN framework overcomes this issue by processing different scale images filtered with different scale filtered size masks and cumulative confidence scores increase the accuracy by detecting the pedestrian at a different scale. The next section describes the proposed SIM R-CNN briefly.

3 Scale-Invariant Mask R-CNN (SIM R-CNN)

3.1 Overview of Proposed Framework

The novel scale-invariant feature map generation algorithm has been introduced in RPN, for multiscale feature map generation for different scale pedestrian instances in the image. The proposed framework is divided into three stages viz. 1) scale-invariant feature map generation. 2) region proposals generation. 3) extraction of a detected pedestrian object from the image. Each of these stages is described in brief as follows. Input to the mask R-CNN is the raw image and it generates the object surrounded by a bounding box, class label, and

pixel-level mask.

The Mask-RCNN is used for pixel-level mask generation as shown in Figure 5. The process of pedestrian detection is divided into three stages 1) the RPN generates and identifies the proposals for regions, 2) scale-invariant feature map generation algorithm generates the multi-scale feature map. 3) bounding box regression, class prediction and binary mask prediction is based on the proposal of regions extracted in stages 1 and 2. Both stages are merged using a backbone structure. Each of this stage is described in brief as following:

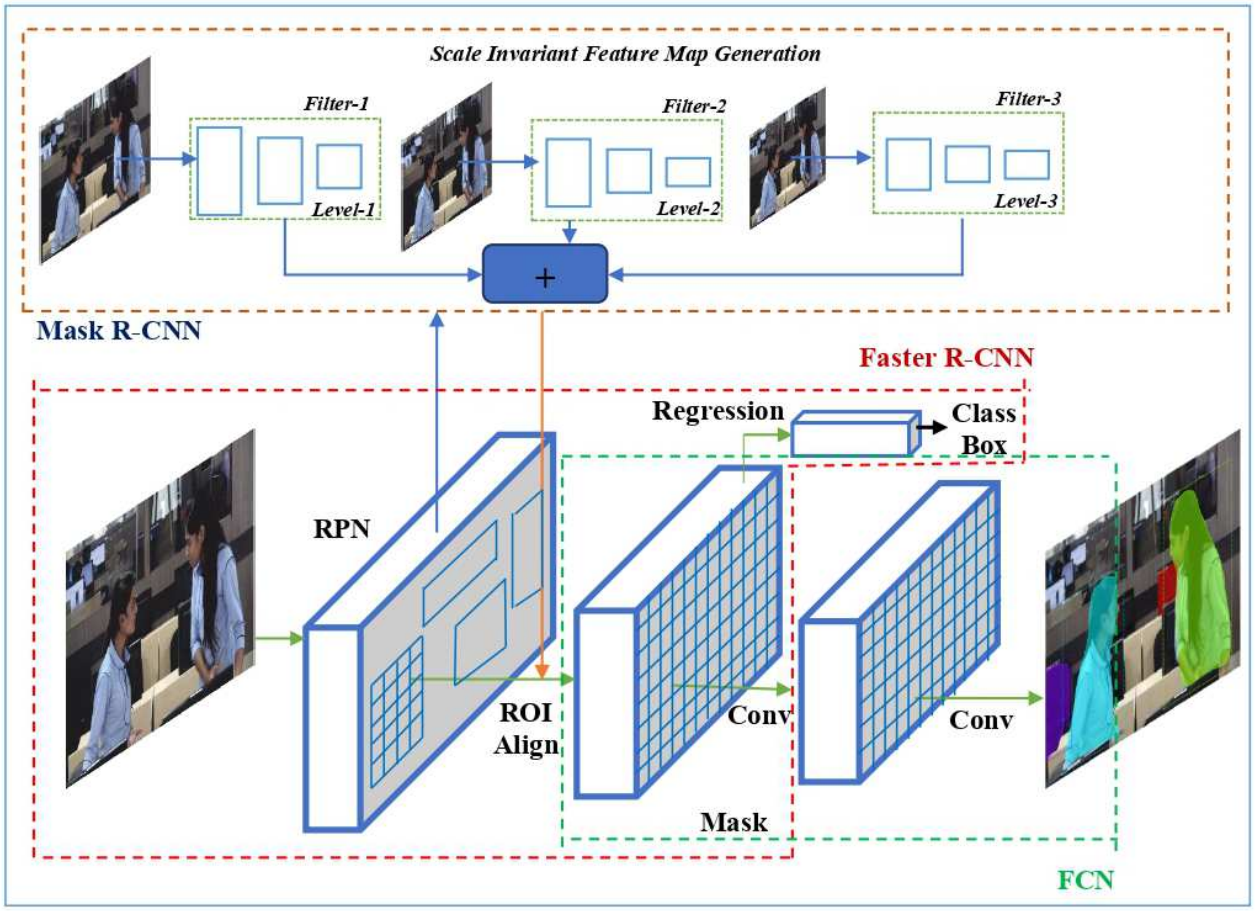

Figure 5: Improved Mask R-CNN Architecture for instance segmentation and pedestrian detection.

3.2 Pedestrian Region Proposals Extraction

3.2.1 Region Proposal Network (RPN)

The first stage of architecture is RPN. It is a lightweight deep neural network. The RPN proposes an RoI by Feature Pyramid Networks (FPN). In this, the processed input image is convolved in the bottom-up and top-down approach for extracting the features. The process of RPN is divided into different steps as shown in Figure 6. It consists of viz. 1) input image acquisition, 2) feature extractors, 3) bounding box regression and class prediction, 4) the RoI Alignment, 5) binary-Mask Prediction. Each of this step is described in brief as follows:

3.2.1.1 Input image acquisition The input to the network is the image or frame of the video. The video is read from the database and converted into frames. Each frame is processed using an RPN to generate the

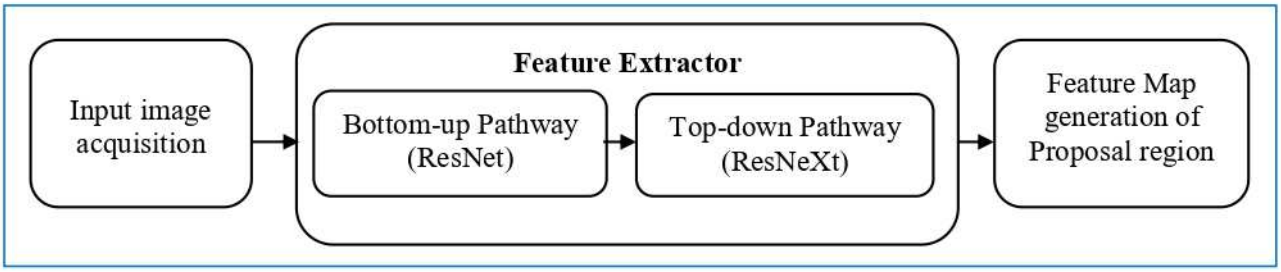

Figure 6: The process of Region Proposal Network.

bounding box and class prediction for the specific object in the image.

3.2.1.2 Feature Extractor From the input image, high-level features are extracted using a Residual neural network (ResNet), called a bottom-up pathway, and ResNeXt, called a top-down pathway as shown in Figure 7. ResNet architecture is the same as the other neural network. In a traditional neural network, each layer feeds into the next layer. But in ResNet instead of the subsequent layer, it can be feed into layers which are two-threes hops away from the current layer. Consider the input x and learning the true distribution of x is represented as $T(x)$. The residual between this is represented as $Res(x)$ and computed using the Eq. 1.

$$Res(x) = T(x) - x \tag{1}$$

The true distribution $T(x)$ for input x is computed by solving the Eq. 1 and after solving it represents as in Eq. 2.

$$T(x) = Res(x) + x \tag{2}$$

The traditional neural network learns using $T(x)$. Opposite to that ResNet learns using the $Res(x)$. ResNet and ResNeXt generate the FPN for the complete image at multiple scales. FPN is used for feature detection. FPN consists of a pyramid structure because each object with different sizes can be considered in any one of the levels in the pyramid. The bottom-up pathway uses CNN architecture for feature extraction. As the scanning progresses from bottom to top the resolution decreases. The output of the convolution module at the end is used as input in the top-down pathway.

In this, 1 x 1 convolution filters are implemented to reduce channel depth to $256 - d$ and generate the feature map layer i.e. $M5$ that is used for object prediction. As the scanning progress, convolved layered output is upsampled by 2 and 1 x 1 convolution filter of the bottom-up pathway is applied to respective feature maps. These extracted feature map layers are used as input in region proposals network for class prediction and bounding box regression purposes.

3.2.1.3 Region Proposals In this step, the previous layer convolved with a 3 x 3 sliding window is used to generate the multiple feature map. Next, again, 1 x 1 convolution filters are applied for class prediction and bounding box regression called RPN head. 3 x 3 mask sliding window has been used because the mask applied from the center will lead to efficient information. K anchor boxes with different sizes are used at the time of convolution so that different shape objects can be detected efficiently as shown in Figure 8(a). The output is generated as a $2k$ score for class layers and a $4k$ score of bounding box regression.

The generated $2k$ score indicates that the extracted information consists of background or not. $4k$ score consists of the parameter of the bounding box represented by (x, y, w, h). Where w represents the width and h represents the height of the bounding box. x and y indicate the spatial plane coordinate. k represents the count of anchor boxes of the different size bounding box. The next stage includes steps such as scale-invariant

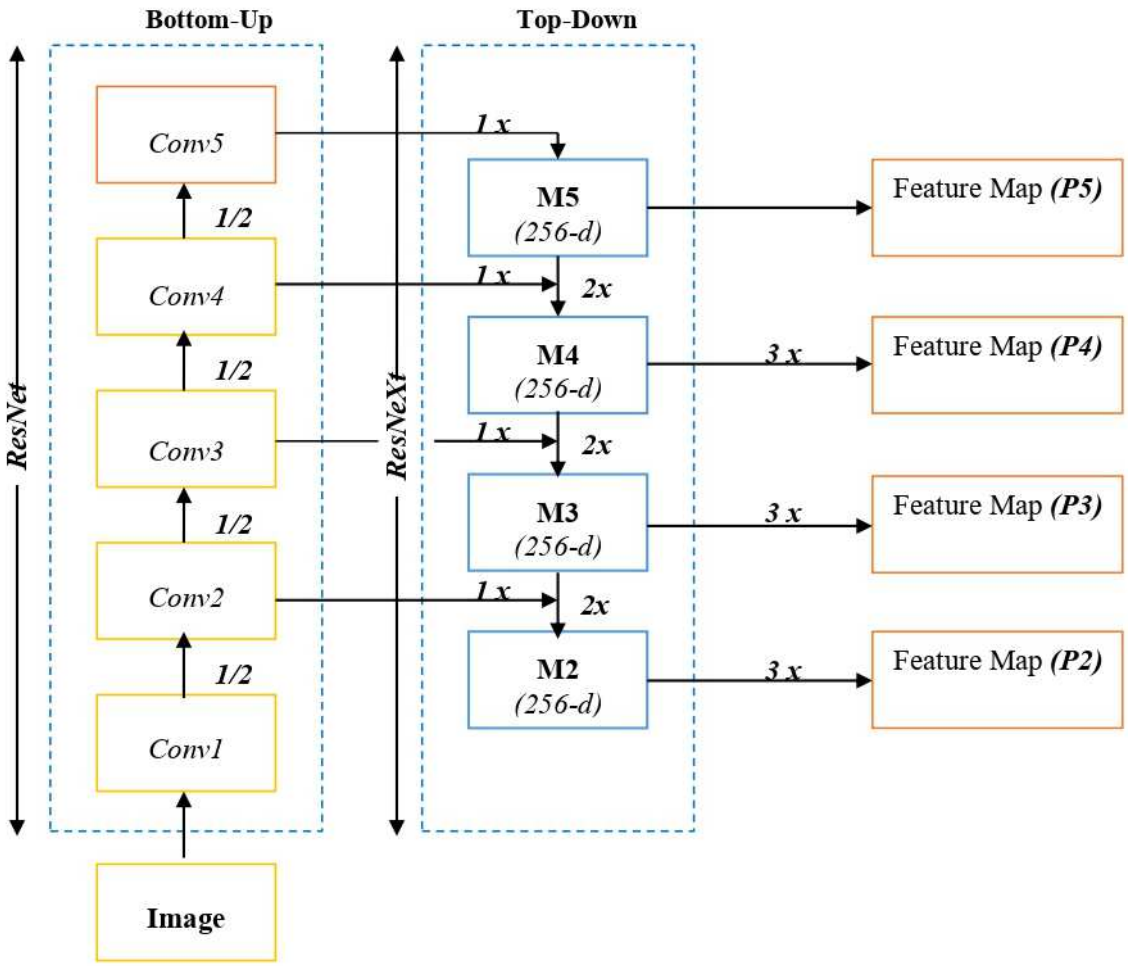

Figure 7: Feature extraction using bottom-up and top-down pathway.

feature map generation, RoI alignment, bounding box regression, class prediction, and binary mask prediction, describe in brief as follows.

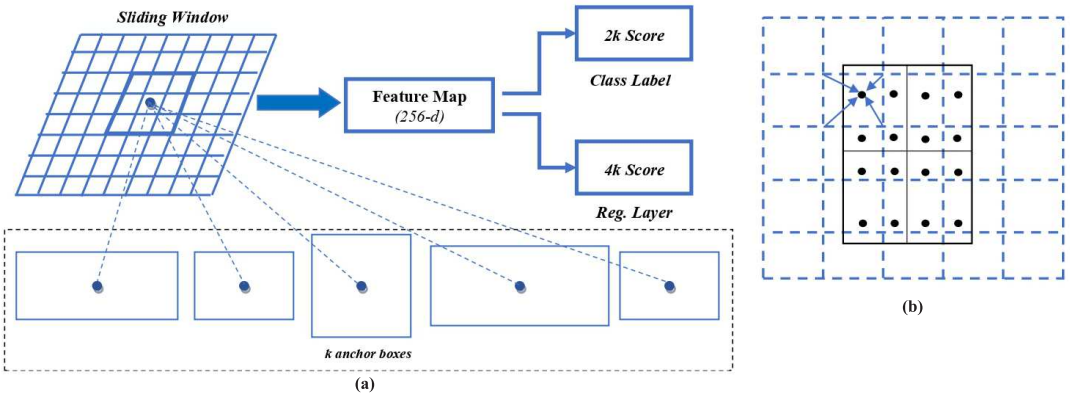

Figure 8: The process of sliding window and RoI align (a) sliding window for feature map generation (b) the RoI align process for input region proposal

3.3 Scale-invariant feature map generation

The human instances which are different in scale are not classified efficiently in the Mask R-CNN. Hence it is necessary to address this issue by generating the scale-invariant feature map for the different scale of human instances. The different scale images are convolved with various sizes of anchor box filters to generate the confidence score without ignoring the low-resolution layer, which was ignored in the exiting Mask R-CNN RPN network [6]. Later this score is integrated with the feature map generated in an earlier stage. Here the region proposals accuracy increases due to the scale-invariant feature map. The steps of scale-invariant map generation is described as follows:

Algorithm 1: Scale-invariant feature map generation

Input : Mf_j, the different Scale filtered Mask and $sf(x, y)$, scale image of size $w \times l$
Output : f_m, the scale-invariant feature map
parameter: $Confiscore$, Initialize to zero

Read the total number of scale image up to n;
for $sf_i \leftarrow 1$ **to** n **do**
  Convolved image with different scale map m;
  for $j \leftarrow 1$ **to** m **do**
    Mask convolved over-scaled image;
    $Confiscore \leftarrow$ Conv(sf,Mf) ;
    if $Confiscore$ = $null$ **then**
      $break$;
    else
      $Cummscore \leftarrow Cummscore$ + $Confiscore$;
    end
  end
  $f_m \leftarrow Cummscore$;
end

Scale-invariant feature maps are fused with other feature maps to locate the different scale objects efficiently. Next, each region's proposals are aligned using the RoI align process. Other steps of Mask R-CNN are applied

in further processing. We have discussed the Mask R-CNN architecture used for pedestrian detection. In the end, all the detected objects with the bounding box and segmentation mask are represented on the original image. Then human detected class IDs are separated from all detected objects.

3.4 RoIAlign

The RoI alignment is necessary for compact feature map representation and for maintaining uniformity at the time of convolving, predicting class, segmentation mask, and bounding box. The location of the object matters for the proper placement of the bounding box and to avoid the quantization. RoI alignment is a bilinear interpolation of points in a sliding window [6] as shown in Figure 8 (b). Bilinear interpolation generates a smooth interpolation as compare to the nearest neighborhood method. The pixel location in the original image is represented by $f(x)$ and is calculated using Eq. (3) and (4).

$$f(x) = \rho(x) \times \frac{S_w}{\rho_w} \tag{3}$$

$$f(y) = \rho(\gamma) \times \frac{S_h}{\rho_h} \tag{4}$$

$\rho(x)$ and $\rho(\gamma)$ are the current proposal coordinate location. S_w and S_h are the source width and height of the original image. ρ_w and ρ_h are the width and height of the proposal. $f(x)$ and $f(y)$ are the target pixel location of the current image.

Consider the input image which is represented as $f(x)$. The adjacent neighborhood pixel coordinates for $f(1, 2, 3.4)$ is represented by $f(1, 3)$, $f(2, 3)$, $f(1, 4)$, and $f(2, 4)$. The pixel location $f(i + \mu , j + \alpha)$, where μ = 0.2, α = 0.4, i = 1, and j = 3. The computation is performed using Eq. (5).

$$\begin{aligned} f(i+\mu, j+\alpha) = (1-\mu)(1-\alpha)f(i,j) + \alpha f(i,j+1) \\ +\mu(1-\alpha)f(i+1,j) + \mu\alpha f(i+1,j+1) \end{aligned} \tag{5}$$

After sampling, four points are generated and max-pooling is performed on each sampled point to represent the aligned coordinate on the image.

3.5 Bounding Box Regression, class, and mask prediction

The generated four points in RoI alignment are used as bounding box parameters at the top left x, top left y, width, and height of the bounding box. The detected objects are represented using the bounding box parameter. The class label is assigned for each bounding box. It is the index number of detected objects in an image. Next, for each bounding box parameter segmentation mask is generated. The binary mask has been generated at the end. The size of a binary mask is represented by Km^2. Where K indicates the class number and m represents spatial plane resolution. The size of a mask is resized to obtain an accurate mask for each object in an image.

4 Experiments

We evaluated the effectiveness of the proposed SIM R-CNN framework on several popular pedestrian detection datasets including Caltech [1], INRIA [2], MS COCO [3], ETH [4], and KITTI [5] and our pedestrian database. All experiments were implemented in a machine with a single GPU and a CPU Intel Core $i5$ 3.4GHz processor having 16GB RAM and NVIDIA graphics processor.

4.1 Proposed Pedestrian database in the academic environment

We proposed a pedestrian database, that contains a video sequence of student behavior such as student studying in a classroom, practical lab, examination hall scenarios, a student doing the cheating in the exam hall, a student taking answer book outside the exam hall, student stealing the mobile phone, student stealing the laboratory material, student dispute in the college premises, student disturbing another student, student threatening another student, etc. Student's behavior in college premises is recorded using a digital camera from a different viewing angle. The video is recorded at 30 frames/second. The database comprises of 100 sample videos of approx. 20-30 minutes duration for each sample video. The sample frames of our database shown in Figure 9. The Table 1 shows the proposed pedestrian database details. The different student behavior of students is recorded in the daylight condition. Each behavior category consists of 10 videos.

Table 1: Details of proposed pedestrian database in the academic environment

Pedestrian Behaviour	Images or Video Clip	Environment or location	Annotation
Student doing the cheating	3,60,000 frames/sec.	Examination Hall	215K annotated obj
Student taking answer book	3,60,000 frames/sec.	Examination Hall	210K annotated obj
Student stealing the mobile phone	3,60,000 frames/sec.	Labs	225K annotated obj
Student stealing material	3,60,000 frames/sec.	Labs	220K annotated obj
Student dispute in the college premises	3,60,000 frames/sec.	Class room and Labs	210K annotated obj
Student disturbing another student	3,60,000 frames/sec.	Class room and Labs	195K annotated obj
Student threatening another student	3,60,000 frames/sec.	Class room and Labs	185K annotated obj

Figure 9: Sample images of our database. The first row shows a scenario where a two girls dispute in the laboratory. The second row shows the scenario of students stealing the mobile phone of another student. The third row shows a scenario of a student threatening another students side view. The fourth row shows the same threatening scenario with the front view. The fifth row shows the scenario of students stealing the laboratory material. The sixth row shows the scenario of the student doing the cheating in the exam hall.

4.2 Comparison with State-of-the-Art Pedestrian Detection Methods

4.2.1 Caltech

The proposed model is also trained using Caltech training and testing dataset. The results are illustrated in Figure $10(a)$. The proposed approach is compared with the existing techniques such as TA-CNN [44], Checkerboards [45], CompACT-Deep [46], and SAF R-CNN [31]. It can be observed that SIM R-CNN outperforms other methods by a large margin and obtains the lowest miss rate of 8.32%, which achieves state-of-the-art performance for pedestrian detection using Mask R-CNN.

4.2.2 INRIA and ETH

The SIM R-CNN is also trained and tested with the INRIA and ETH datasets. The comparison results are illustrated in Figure $10(b)$ and Figure $10(c)$. First, for the INRIA dataset, the proposed approach gives the miss rate of 7.32%, which outperforms the existing method [44]. Second, for the ETH dataset, the miss rate of the proposed model is 32.64% compared with 34.98% of [41] and 37.37% of [47]. In general, the proposed method achieves a higher detection rate on both the dataset.

4.2.3 KITTI

The proposed framework is also tested on challenging the KITTI dataset. The pedestrian detection results and performance comparisons of the SIM R-CNN with other existing approaches [44] [45] [46] [47] are shown in Figure $10(d)$. The proposed approach gives promising results on the KITTI dataset, i.e., 76%, 64%, and 60%.

4.2.4 MS COCO and proposed pedestrian dataset

The proposed model is also trained using the MS COCO and proposed pedestrian dataset. The results are shown in Figure $10(e)$ and Figure $10(f)$. The proposed approach compared with the existing techniques in [44][45][46][47]. It can be observe that SIM R-CNN outperform other methods and gives a miss rate of 8.57% on MS COCO and miss rate of 8.69% on the proposed pedestrian dataset. We have represented the result using the receiving operating characteristics curve. The proposed SIM R-CNN based pedestrian detector is compared with state-of-the-art pedestrian detection methods. Figure $10(g)$ shows the results of the recall and precision of the proposed model compare with TA-CNN [44], Checkerboards [45], CompACT-Deep [46], and SAF R-CNN [31]. It can observes that the SIM R-CNN. Mask-RCNN outperformed existing approaches in terms of accuracy, speed, and time needed for pedestrian detection. The precision, recall indicates how accurately the pedestrian is detected in an image.

The SIM R-CNN applied over the various dataset and pedestrian detection results are represented in the Figure 11. Again, the SIM R-CNN detect the small-size pedestrian instances efficiently. It is observed from the experimental results that SIM R-CNN can be accurately and efficiently detect the variant size pedestrian instances that were not detected in the TA-CNN and CompACT-Deep framework. In heavy occlusion also SIM R-CNN detects the pedestrians accurately. Table 2 shows the comparison of the proposed pedestrian detector with the existing state of the art pedestrian detector. The average log miss rate and testing time metric are used for the comparison. The overall accuracy of the SIM R-CNN is 76% at the miss rate of 8.32% and 0.51% that is better than the existing approaches. The object rotation, illumination changes affects the performance of proposed SIM R-CNN because it is scale-invariant not rotation and illumination invariant. Variation in object pose and illumination condition can be handled in the future by adding the rotation and illumination invariant features to the existing detector.

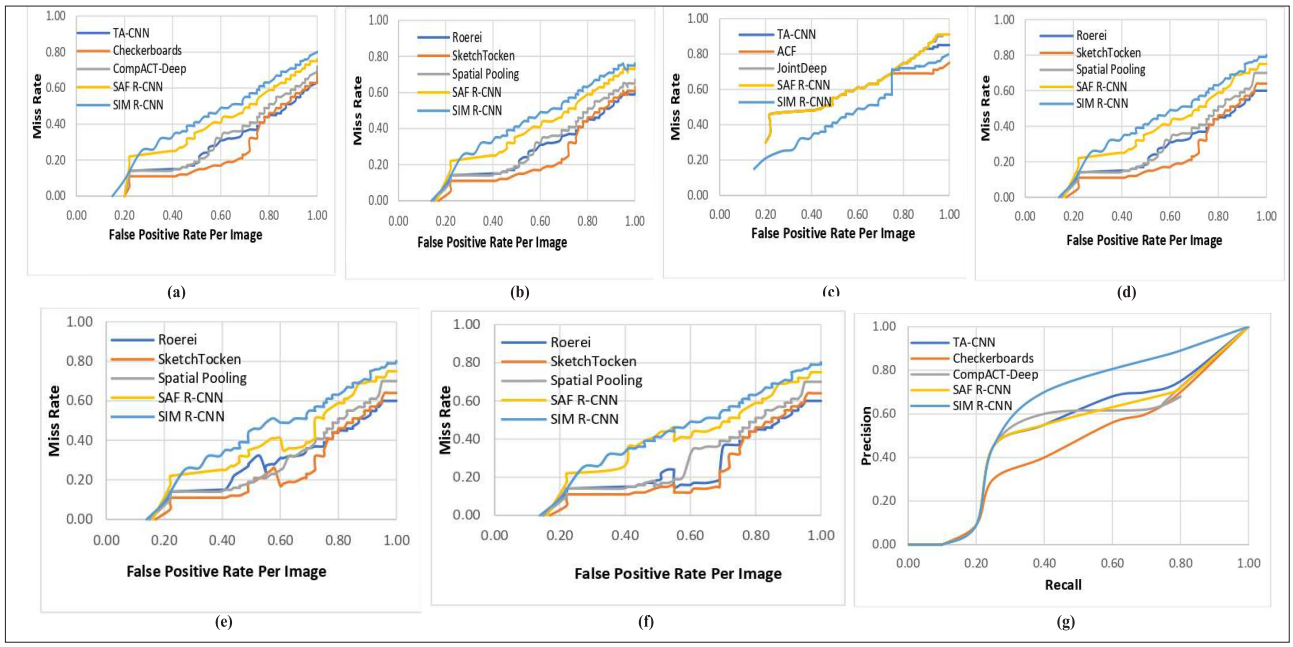

Figure 10: The comparison of SIM R-CNN pedestrian detector with recent state-of-the-art methods and datasets. (a) Caltech dataset: The SIM R-CNN outperforms other methods with the lowest log-average miss rate of 8.32%. (b) INRIA dataset: The SIM R-CNN outperforms other methods with the lowest log-average miss rate of 7.32%. (c) ETH dataset: The SIM R-CNN outperforms other methods with the miss rate of the proposed model is 32.64%. (d) KITTI dataset: The SIM R-CNN gives promising results 76%, 64%, and 60%. (e) MS COCO dataset: SIM R-CNN outperforms other methods and gives a miss rate of 8.57%. (f) Proposed pedestrian dataset: SIM R-CNN outperforms other methods and gives a miss rate of 8.69%. (g) Precision vs Recall: The comparison of SIM R-CNN pedestrian detector with recent state-of-the-art methods.

Table 2: Comparison of the proposed SIM R-CNN pedestrian detector with the state-of-the art pedestrian detector based on Miss rate and testing time

Methods	Faster R-CNN	Fast R-CNN Single Scale	Fast R-CNN Multi Scale	R-CNN	SAF R-CNN	Proposed SIM R-CNN
Miss rate (%)	17.60	13.70	11.67	12.77	9.32	8.32
Testing Time (%)	0.22	0.34	3.04	5.31	0.59	0.51

Figure 11: Results of the SIM R-CNN model on various datasets. The first row shows the original sample images in the proposed pedestrian database. The second row shows the pedestrian detected images using the proposed approach. The third row shows an original sample of pedestrian images of the MS COCO dataset [3]. The fourth row shows the pedestrian detected images using the proposed approach on the Caltech pedestrian dataset. The fifth row shows the original sample images of the Caltech pedestrian dataset [1]. The sixth row shows the pedestrian detected images using the proposed approach on the Caltech pedestrian dataset.

5 Conclusion

In this paper, we proposed a scale-invariant Mask-RCNN based pedestrian detector. Scale-invariant features extracted from the RPN are used jointly with convolutional features to detect the different scale pedestrian efficiently in the image. The proposed detector achieves competitive result on the various pedestrian benchmark datasets such as Caltech [1], INRIA [2], Microsoft Common Object in Context (COCO) [3], ETH [4], KITTI [5], and our proposed academic environment pedestrian database. The experimental results showed that the proposed SIM R-CNN pedestrian detector gives 1) the lowest miss rate of 8.32% on the Caltech dataset, 2) the lowest log-average miss rate of 7.32% INRIA, 3) miss rate of 32.64% on ETH dataset, pedestrian detection accuracy of 76% on the KITTI dataset and 4) miss rate of 8.69% on the proposed database. The proposed method is superior in detecting different size pedestrian compared with the existing state of the art techniques such as TA-CNN [44], Checkerboards [45], CompACT-Deep [46], and SAF R-CNN [31]. In the future, the proposed model can also be used for detecting the human pose of different size pedestrians.

References

[1] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 34(4):743-761, 2012. DOI: https://doi.org/10.1109/TPAMI.2011.155

[2] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886-893, San Diego, CA, USA, 20th-25th June 2005. DOI: https://doi.org/10.1109/CVPR.2005.177

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context, 13th European Conference

on Computer Vision (ECCV), Springer, Zurich, Switzerland, pp. 1-15, 6th-12th September 2014. DOI: https://doi.org/10.1007/978-3-319-10602-1_48

[4] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis, IEEE International Conference on Computer Vision (ICCV), pp. 1-8, Venice, Italy, 22nd-29th October 2017. DOI: https://doi.org/10.1109/iccv.2007.4409092

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite, International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354-3361, RI, United States, 16th-21st June 2012. DOI: https://doi.org/10.1109/CVPR.2012.6248074

[6] Kaiming He, Georgia Gkioxari, Piotr Dollar and Ross Girshick. Mask R-CNN, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 42(2):386-397, Feb. 2020. DOI: https://doi.org/10.1109/TPAMI.2018.2844175

[7] Kaiming He, Georgia Gkioxari, Piotr Dollar and Ross Girshick. Mask R-CNN, IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988, Venice, Italy, 22nd-29th October 2017. DOI: https://doi.org/10.1109/iccv.2017.322

[8] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Las Vegas, Nevada, USA, 26th June-1st July 2016. DOI: https://doi.org/10.1109/CVPR.2016.91

[9] Jifeng Dai, Yi Li, Kaiming He and Jian Sun. R-FCN: Object Detection via Region-based Fully Convolutional Networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-11, Las Vegas, Nevada, USA, 26th June-1st July 2016.

[10] Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu and Alexander C. Berg. SSD: Single Shot MultiBox Detector, 14th European Conference on Computer Vision (ECCV), pp. 1-17, Amsterdam, Netherlands, 11th-14th October 2016. DOI: https://doi.org/10.1007/978-3-319-46448-0_2

[11] Jonathan Long Anguelov, Evan Shelhamer and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation, Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Boston, Massachusetts, 8th-10th June 2015. DOI: https://doi.org/10.1109/CVPR.2015.7298965

[12] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 39(6):1137-1149, June 2017. DOI: https://doi.org/10.1109/TPAMI.2016.2577031

[13] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks, Neural Information Processing Systems (NIPS), Montreal, pp. 1-9, Quebec, Canada, 7th-12th December 2015.

[14] Ross Girshick. Fast R-CNN, International Conference on Computer Vision (ICCV), pp. 1441-1448, Santiago, Chile 7th-13th December 2015. DOI: https://doi.org/10.1109/ICCV.2015.169

[15] Alec Radford, Luke Metz and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Boston, Massachusetts, 8th-10th June 2015.

[16] K. He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition, Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Boston, Massachusetts, 8th-10th June 2015. DOI: https://doi.org/10.1109/CVPR.2016.90

[17] Christian Szegedy, Sergey Ioffe and Vincent Vanhoucke, Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Las Vegas, Nevada, USA, 26th June-1st July 2016.

[18] on Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Boston, Massachusetts, 8th-10th June 2015.

[19] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks, Computer Vision and Pattern Recognition (CVPR), pp. 1-11, Portland, Oregon, USA, 23rd-28th June 2013. DOI: https://doi.org/10.1007/978-3-319-10590-1_53

[20] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks, 25th International Conference on Neural Information Processing Systems (NIPS), pp. 1-9, Lake Tahoe, Nevada, United States, 3rd-6th December 2012. DOI: https://doi.org/10.1145/3065386

[21] Liu Kangming. Research on an improved pedestrian detection method based on Deep Belief Network (DBN) classification algorithm, Journal of Information Systems and Technologies (RISTI), 17(3):77-87, March 2016.

[22] Y.Lecun, L. Bottou, Y. Bengio and P. Haffner. Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86(11): 2278-2324, Nov. 1998. DOI: https://doi.org/10.1109/5.726791

[23] Seonghoon Kang, Hyeran Byun and Seong-Whan Lee. Real-Time Pedestrian Detection Using Support Vector Machines, First International Workshop on SVM: Pattern Recognition with Support Vector Machines, pp 268-277, Niagara Falls, Canada, 10th August 2002. DOI: https://doi.org/10.1142/S0218001403002435

[24] David Geronimo, Angel D. Sappa, Antonio Lopez and Daniel Ponsa. Pedestrian detection using AdaBoost learning of features and vehicle pitch estimation, International Conference on Visualization, Imaging, and Image Processing, pp. 1-8, Spain, 28th-30th August 2006.

[25] C.Wu, J. Yue, L.Wang and F. Lyu. Detection and Classification of Recessive Weakness in Superbuck Converter Based on WPD-PCA and Probabilistic Neural Network, MDPI Electronics, 8(290):1-17, March 2019. DOI: https://doi.org/10.3390/electronics8030290

[26] Asvadi, Alireza, Karami-Mollaie, Mohammad Reza, Baleghi, Yasser, Seyyedi Andi, and Hosein. Improved Object Tracking Using Radial Basis Function Neural Networks, 7th Iranian Conference on Machine Vision and Image Processing (MVIP), Tehran, Iran, pp. 1-5, Nov 16th-17th Nov 2011. DOI: https://doi.org/10.1109/IranianMVIP.2011.6121604

[27] Neagoe, Victor Emil, Tudoran, Cristian, Neghina, and Mihai. A neural network approach to pedestrian detection, 13th WSEAS International Conference on COMPUTERS (ICCOMP), pp. 374-379, Wisconsin, United States, 23rd July 2009.

[28] Juncheng Wang andGuiying Li. Accelerate proposal generation in R-CNN methods for fast pedestrian extraction, The Electronic Library, Emerald, 37 (3): 1-19, May 2019. DOI: https://doi.org/10.1108/EL-09-2018-0191

[29] Bruno Artacho, Andreas Savakis. Waterfall Atrous Spatial Pooling Architecture for Efficient Semantic Segmentation, Sensors, MDPI, 19, (24): 1-17, 2019. DOI: https://doi.org/10.3390/s19245361

[30] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multiscale orderless pooling of deep convolutional activation features, 13th European Conference on Computer Vision (ECCV), pp. 392-407, Zurich, Switzerland, 6th-12th September 2014. DOI: https://doi.org/10.1007/978-3-319-10584-0_26

[31] Jianan Li, Xiaodan Liang, Shengmei Shen, Tingfa Xu, Jiashi Feng, Shuicheng Yan. Scale-aware Fast R-CNN for Pedestrian Detection, IEEE Transaction on Multimedia, 20(4):985-996, April 2018. DOI: https://doi.org/10.1109/TMM.2017.2759508

[32] Kelong Wang and Wei Zhou. Pedestrian and cyclist detection based on deep neural network fast R-CNN, International Journal of Advanced Robotic Systems, SAGE, 16(2):1-10, April 2019. DOI: https://doi.org/10.1177/1729881419829651

[33] MiranPobar and Marina Ivasic-Kosm. Mask R-CNN and Optical flow-based method for detection and marking of handball actions, 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI 2018), pp. 1-6, Beijing, China, 13th-15th October 2018. DOI: https://doi.org/10.1109/CISP-BMEI.2018.8633201

[34] AsatiMinkesh, Kraisittipong Worranitta and Miyachi Taizo. Human extraction and scene transition utilizing Mask R-CNN, International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-6, California, United States, 16th-20th June 2019.

[35] Gawande, Ujwalla and Hajari, Kamal and Golhar, Yogesh. Pedestrian Detection and Tracking in Video Surveillance System: Issues, Comprehensive Review, and Challenges, Recent Trends in Computational Intelligence, IntechOpen,1-24, April 2020. DOI: https://doi.org/10.5772/intechopen.90810

[36] Ujwalla Gawande, Kamal Hajari and Yogesh Golhar. Deep Learning Approach to Key Frame Detection in Human Action Videos, Recent Trends in Computational Intelligence, IntechOpen,1-17, Feb 2020. DOI: https://doi.org/10.5772/intechopen.91188

[37] Xiaoyu Wang, Tony X. Han and Shuicheng Yan. An HOG-LBP human detector with partial occlusion handling, IEEE 12th International Conference on Computer Vision (ICCV), pp. 32-39, Kyoto, Japan, 29th September-2nd October 2009. DOI: https://doi.org/10.1109/ICCV.2009.5459207

[38] Piotr Dollar, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features, British Machine Vision Conference (BMVC), London, UK, pp. 1-11, 7th-10th September 2009. DOI: https://doi.org/10.5244/C.23.91

[39] Ujwalla Gawande, Yogesh Golhar. Biometric security system: a rigorous review of unimodal and multimodal biometrics techniques, International Journal of Biometrics (IJBM), InderScience, 10(2):142-175, April 2018. DOI: https://doi.org/10.1504/IJBM.2018.10012749

[40] U. Gawande, M. Zaveri and A. Kapur. Bimodal biometric system: feature level fusion of iris and fingerprint, 14th European Conference on Computer Vision (ECCV), ScienceDirect, Elsevier, 2013(2): 7-8, Feb. 2013. DOI: https://doi.org/10.1016/S0969-4765(13)70035-3

[41] Nathan Silberman, David Sontag, Rob Fergus. Instance Segmentation of Indoor Scenes Using a Coverage Loss, 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6th-12th September 2014. DOI: https://doi.org/10.1007/978-3-319-10590-1_40

[42] Ziyu Zhang, Sanja Fidler, Raquel Urtasun. Instance-Level Segmentation for Autonomous Driving with Deep Densely Connected MRFs, Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Boston, Massachusetts, 8th-10th June 2015. DOI: https://doi.org/10.1109/CVPR.2016.79

[43] Xiaogang Wang, Meng Wang, and Wei Li. Scene-specific pedestrian detection for static video surveillance, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 36(2):361-374, 2014. DOI: https://doi.org/10.1109/TPAMI.2013.124

[44] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks, Computer Vision and Pattern Recognition (CVPR), pp. 1-10, Boston, Massachusetts, 8th-10th June 2015. DOI: https://doi.org/10.1109/CVPR.2015.7299143

[45] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Filtered channel features for pedestrian detection, Computer Vision and Pattern Recognition (CVPR), pp. 1751-1760, Boston, Massachusetts, 8th-10th June 2015. DOI: https://doi.org/10.1109/CVPR.2015.7298784

[46] Mohammad Saberian Zhaowei Cai and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection, International Conference on Computer Vision (ICCV), pp. 1-10, Santiago, Chile 7th-13th December 2015. DOI: https://doi.org/10.1109/ICCV.2015.384

[47] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features, 13th European Conference on Computer Vision (ECCV), Springer, Zurich, Switzerland, pp. 546-561, 6th-12th September 2014. DOI: https://doi.org/10.1007/978-3-319-10593-2_36