Applied Intelligence (2019) 49:1200–1211
https://doi.org/10.1007/s10489-018-1326-8

Context-aware pedestrian detection especially for small-sized instances with Deconvolution Integrated Faster RCNN (DIF R-CNN)

Han Xie[1] · Yunfan Chen[1] · Hyunchul Shin[1]

Published online: 29 October 2018

Abstract

Pedestrian detection is a canonical problem in computer vision. Motivated by the observation that the major bottleneck of pedestrian detection lies on the different scales of pedestrian instances in images, our effort is focused on improving the detection rate, especially for small-sized pedestrians who are relatively far from the camera. In this paper, we introduce a novel context-aware pedestrian detection method by developing the Deconvolution Integrated Faster R-CNN (DIF R-CNN), in which we integrate a deconvolutional module to bring additional context information which is helpful to improve the detection accuracy for small-sized pedestrian instances. Furthermore, the state-of-the-art CNN-based model (Inception-ResNet) is exploited to provide a rich and discriminative hierarchy of feature representations. With these enhancements, a new synthetic feature map can be generated with a higher resolution and more semantic information. Additionally, atrous convolution is adopted to enlarge the receptive field of the synthetic feature map. Extensive evaluations on two challenging pedestrian detection datasets demonstrate the effectiveness of the proposed DIF R-CNN. Our new approach performs 12.29% better for detecting small-sized pedestrians (those below 50 pixels in bounding-box height) and 6.87% better for detecting all case pedestrians of the Caltech benchmark than the state-of-the-art method. For aerial-view small-sized pedestrian detection, our method achieve 8.9% better performance when compared to the baseline method on the Okutama human-action dataset.

Keywords Computer vision · Pedestrian detection · Deep learning · Neural network · Deconvolution · Feature map

1 Introduction

Pedestrian detection [1–6] has wide application in video surveillance, intelligent vehicles, robotics, and smart drones monitoring systems. Although steady improvement over the last decade has been made, accurate detection of the presence of pedestrians who are relatively far from the camera remains a challenge. State-of-the-art detectors typically work reasonably well with large-sized pedestrians, but they usually fail to detect small-sized (i.e., far-scaled) ones. Recognizing objects at vastly different scales is a fundamental challenge in computer vision. For a pedestrian of interest, captured features are effective only at a certain scale of the corresponding receptive field, especially in complex scenes that contain pedestrians of different scales. A fixed receptive field cannot cover the multiple scales at which objects appear in natural scenes. Additionally, it has been observed that far-scale instances often result in pedestrians with obscure appearances and blurred boundaries, as shown in Fig. 1a. This distortion makes it difficult to distinguish them from background clutters and other overlapped instances. Large pedestrians can provide rich information for pedestrian detection, while smaller instances of pedestrians cannot be easily recognized. Figure 1a shows an example image from the Caltech benchmark [7]. One pedestrian in the red box has a large scale and four other pedestrians in green boxes are at a small scale. Figure 1b shows five corresponding feature maps. The scale distribution of pedestrian heights in the Caltech training set is illuminated in Fig. 1c. One can observe that small-sized instances indeed dominate the distribution. This suggests that effective detection of small instances of pedestrians is essential to improve the overall detection accuracy.

✉ Han Xie
xiehan@hanyang.ac.kr

Yunfan Chen
chenyunfan@hanyang.ac.kr

Hyunchul Shin
shin@hanyang.ac.kr

[1] Division of Electronical Engineering, Hanyang University, 55 Hanyangdeahak-ro, Sangnok-gu, Ansan, Republic of Korea

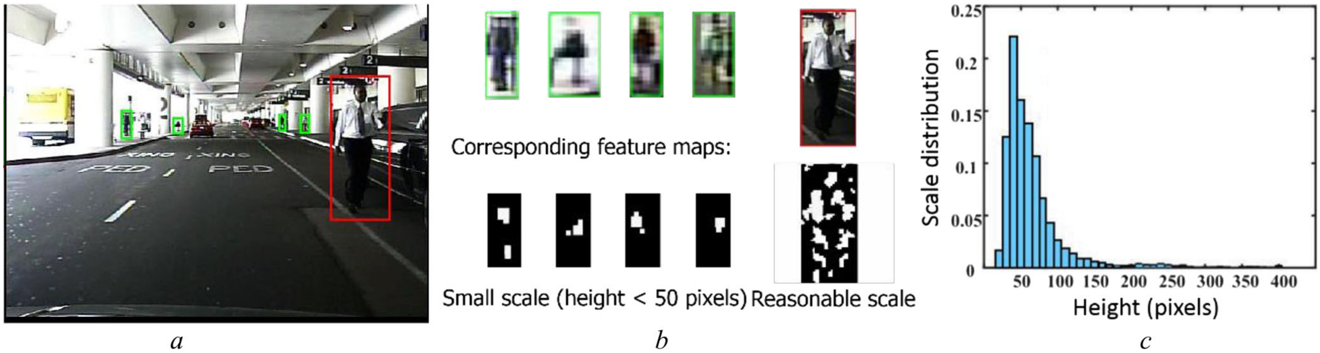


Fig. 1 **a** An example image from the Caltech benchmark [7]. A typical input image usually contains multiple pedestrian instances of different scales. **b** The corresponding feature maps of different scale instances. **c** Scale distribution of pedestrian heights in the Caltech training set. One can observe that small-sized (i.e. short) instances dominate the distribution

Multiple-scale detection problem are often addressed by combining feature maps as the representations of multiple layers in a neural network. SA-Fast RCNN [6] used a divide-and-conquer strategy based on Fast R-CNN, in which multiple built-in subnetworks are designed to adaptively detect pedestrians of different scales. Similarly, MS-CNN [4] worked with multiple layers to match objects of different scales. Another strategy, adopted by [8, 9], generated region proposals based on a single feature map using different anchors with different scales and aspect ratios that correspond to different receptive fields. This strategy avoids the repeated computation of feature maps and tends to be more efficient. Recent works such as Faster-RCNN [9] address the issue with a multi-scale region proposal network (RPN), which achieves excellent object detection performance.

A number of recent approaches have improved the feature extraction of small objects by using additional context information and increasing the spatial resolution of feature maps. DSSD [10] used deconvolution layers in combination with existing multiple layers to reflect the large-scale context. MS-CNN [4] applied deconvolution on shallow layers to increase the feature map resolution before using the layers to extract region proposals and pool features. Recently, Long et al. [11] introduced the Fully Convolution Network (FCN), which demonstrated impressive performance in semantic segmentation [11, 12], and object detection [13]. In [11], the authors combined coarse high-layer information with fine low-layer information for semantic segmentation. Additionally, the atrous convolution represent a powerful and convenient tool to effectively enlarge the field of view of filters and incorporate larger context without increasing the number of parameters or the amount of computations. It amounts to inserting holes ("trous" in French) between nonzero filter taps. This technique has a long history in signal processing. It was originally developed for efficient computation of the undecimated wavelet transform in a scheme referred to as "algorithm a trous" [14]. It has been used as an upsampling filter in DeepLab [13] for image segmentation tasks and shows good performance. However, these kind of technology such as deconvolution and atrous convolution are less explored in pedestrian detection area. In our paper, we will explain how to integrate these techniques to a deep neural network architecture carrying more context information to solve the small-sized pedestrian detection problem.

Motivated by the above ideas, a novel effective pedestrian detection framework based on the Faster R-CNN [9] pipeline is introduced; which named as Deconvolution Integrated Faster R-CNN (DIF R-CNN). This framework can achieve state-of-the-art performance. Our work possesses the following five major contributions:

First, a novel pedestrian detection framework is proposed by adding the deconvolutional module to the traditional Faster R-CNN network. The deconvolutional module can bring in more semantic context information to enhance the feature map, thereby improving the detection performance.

Second, we propose using a reduced network model, in which we adopt the prior layer as the initial feature map with a relatively large spatial resolution, instead of using the last layer as the output feature map. In addition, proper adjustment of the network has been made to avoid downsampling. Both tricks help retain more detailed information for small-sized pedestrians.

Third, a synthetic feature map that combines the initial feature map and the deconvolution layer with semantic information is proposed, instead of using multi-layer feature map based method, such as MS-CNN [4], in which low-level layers have less semantic information regarding small instances.

Fourth, we propose applying atrous convolution on the synthetic feature map. The synthetic feature map captures a rather smaller receptive field. To compensate for this, the atrous convolution can enlarge the receptive field and inject detailed context information. Larger receptive fields help the detection of large-sized instances and detailed context information helps the detection of small-sized instances of pedestrians. Therefore, application of atrous convolution can improve the detection accuracy of multi-scale object detection.

Finally, our approach is demonstrated empirically to achieve superior performance on well-known benchmarks. For example, it noticeably improves 12.29% for detecting far-scale pedestrian (those below 50 pixels in bounding-box height) and 6.87% for detecting all case pedestrian of the Caltech benchmark when compared to the state-of-the-art method.

The remainder of this paper is organized as follows. Section 2 introduces works related to deep learning and multi-scale pedestrian detection. Section 3 describes the proposed context-aware deep neural network (DIF R-CNN) in detail. Experimental results and analysis are presented in Section 4. Finally, conclusions and future works are summarized in Section 5.

2 Related works

Convolutional neural networks (CNNs) have recently been successfully applied in generic object recognition [9, 10, 12]. CNN-based models such as AlexNet [15], VGG [16], GoogleNet [17], InceptionNet [18], and ResNet [19] have also been developed. With the rise of deep learning methods, some recent works [2, 8, 20] have shown significant progress in pedestrian detection. Pierre et al. [8] proposed an unsupervised method based on convolutional sparse coding to pre-train the filters at each stage. F-DNN + SS [2] used a derivation of the Faster R-CNN, which adopted multiple parallel classifiers with soft-rejection-based network fusion. Zhang et al. proposed a model, referred to as RPN + BF [20], in which Faster R-CNN was applied successfully to pedestrian detection for two reasons: to improve the resolution of the feature maps and mine hard-negative examples. To improve the detection performance, most of these methods adopted pre-trained models from ImageNet, which have been successfully used for classification. However, the pre-trained models are always very deep in the sense that they have multiple pooling layers. These models are good for coarse-grained classification tasks but have some limitations in fine-grained ones, such as small object detection and semantic segmentation.

Several multi-scale detection schemes have been considered in order to achieve good performance in both big- and small-scale detection. In [4], default boxes of different scales were set to multiple layers within the convolutional neural network to predict objects at a certain scale on each layer. Since the nodes of different layers correspond to different receptive fields, it is natural to predict large objects from layers with large receptive fields and to use layers with small receptive fields to predict small objects. However, in order to perform small object detection well, these methods need to use some information from dense feature maps as well as from shallow layers with small receptive fields. Since shallow layers have less semantic information about objects, this may result in low performance for detecting small objects. Another multi-scale object detection method is to use different anchors in the feature map. For example, the Region Proposal Network (RPN) stage of Faster R-CNN [9] can generate a series of different scales of anchors, which can cover different scales of objects. The original Faster R-CNN did not perform well in small object detection, as demonstrated in RPN + BF [20]. During network layer down sampling, deep layers lose information related to small-sized objects. Therefore, it is natural to try to improve the small-scale object detection performance by retaining information related to small objects, even in deep layers.

Therefore, we propose a new context-aware deep neural network by using the Faster R-CNN pipeline, in which the deconvolutional module is added to bring in more semantic context information. A new synthetic feature map is generated by combining the deconvolution layer and high-level feature map to enhance the feature representation and to solve the problem of the shrinking resolution of feature maps in convolution neural networks. Furthermore, the receptive field of our DIF R-CNN is extended by atrous convolution, which also helps to inject more context information. Compared with multi-layer feature extraction methods like MS-CNN [4], our method reuses the higher-resolution maps in the feature hierarchy, instead of adding several feature maps for shallow layers; this results in a more effective procedure for detecting small objects.

3 Proposed context-aware deep neural network

The overview diagram of proposed pedestrian detection framework is illustrated in Fig. 2. As shown, our system consists of the base network for initial feature map generation, the deconvolution part for synthetic feature map generation, and region proposals generated with atrous convolution and classification. Starting from the left, the entire image is forwarded through the convolution layers to generate the initial feature map (FM1). Based on the initial feature map, we apply deconvolution with the encoder-decoder structure, combining the deconvolution layer with the initial feature map to generate the synthetic feature map that collects additional context information. Finally, atrous convolution is applied to the synthetic feature map to generate region proposals. These proposals are then classified and adjusted with the detection module. In following section, our approach is explained in detail.

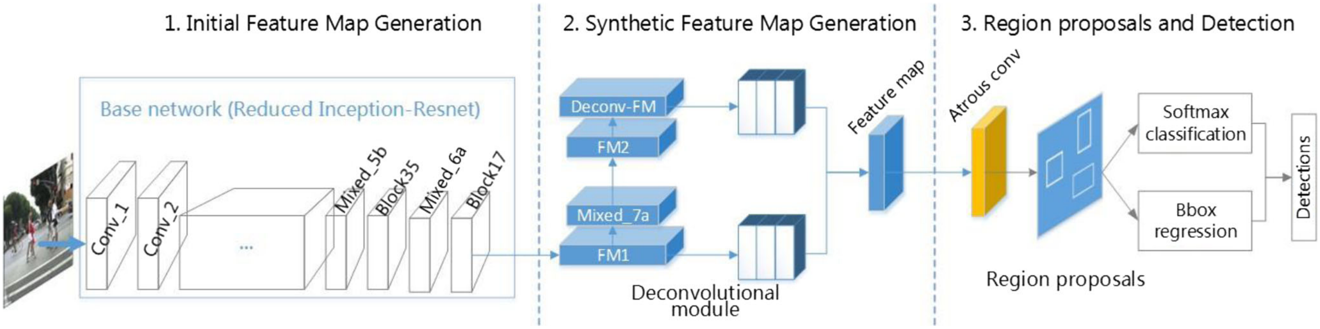


Fig. 2 The proposed DIF R-CNN framework for pedestrian detection

3.1 Initial feature map generation

In the base network, we use Inception ResNet to generate the initial feature map (FM1). Inception ResNet, proposed by Szegedy et al. [18], combines the optimization benefits conferred by residual connections with the computation efficiency of inception units. The inception module approximates a sparse CNN with a normal dense construction. It uses convolutions of different sizes (5×5, 3×3, 1×1) to capture details at various scales. As shown in Fig. 3, the Inception ResNet replaces the filter concatenation stage of the Inception module with residual connections. Compared with VGG16, Inception ResNet can be made more efficient by making the architecture deeper and wider. It also shows better performance than VGG-16 in the ILSVRC 2014 classification and detection challenges. Therefore, we used the more advanced model (i.e., Inception ResNet) in our method.

By the repeated combination of max-pooling and downsampling ('striding'), performed at consecutive layers of deep convolutional neural networks, the feature maps will be significantly reduced in terms of their spatial resolution. This lower spatial resolution facilitates the detection of small pedestrians.

Followed the Inception-Resnet-v2 [18], a reduced Inception-ResNet is adopted in our method. Each layer's kernel size, filter number, stride, and output size are shown in Fig. 4. Instead of using the full network and getting the feature map at the Mixed_7a layer, the output of the Inception ResNet block17 is taken as an initial feature map. If the Mixed_7a layer is used, the resolution of the feature map is too small, losing almost all of the small object details. Therefore, the feature map after block17 is selected. Furthermore, by setting the stride=1, downsampling can be avoid, in order to retain as much information about the small objects as possible in the deep layers of the network Mixed_5b layer to the block17 layer. The output shape of the initial feature map is [33 × 33 × 1088]. Thus, we adopt a feature map that is approximately 4 × 4 times larger when compared with the ordinary feature map (with an 8 × 8 resolution) obtained from Mixed_7a in the original full network. Finally, a deconvolution layer is used in the base network to add the semantic information; this will be explained in Section 3.2.

3.2 Synthetic feature map generation with deconvolutional module

In order to help integrate information from the initial feature map and the deconvolution layer, we used a deconvolutional module that was shown to be helpful for small object detection in DSSD [10]. The original deconvolutional module is inspired by Pinheiro et al. [21] who suggested that a factored variant of the deconvolutional module for a refinement network can lead to evenly matched accuracies as a more sophisticated network, and the deconvolutional module can make the network more efficient. Therefore, to strengthen features, adding extra deconvolution layers is proposed. The deconvolutional module in our experiments is built at the end of the base network.

The deconvolutional module we adopted is shown in Fig. 5, where 3 × 3 convolution and rectified linear activation are used. For the deconvolution branch, the encoder-decoder structure with 2×2 deconvolution is used followed by a 3×3 convolution. A batch normalization layer (BN) is added after each convolution layer. FM2 is extracted after Mixed_7a as an intermediate feature map. Then, the deconvolution layer is added to enlarge the feature map size in order to match the size of the initial feature map (FM1). Finally, element-wise product is performed as a combination method, which is followed by rectified linear activation to generate the synthetic feature map.

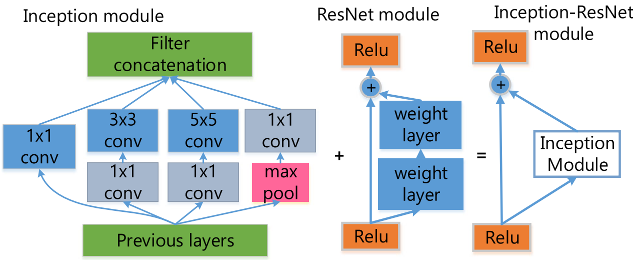
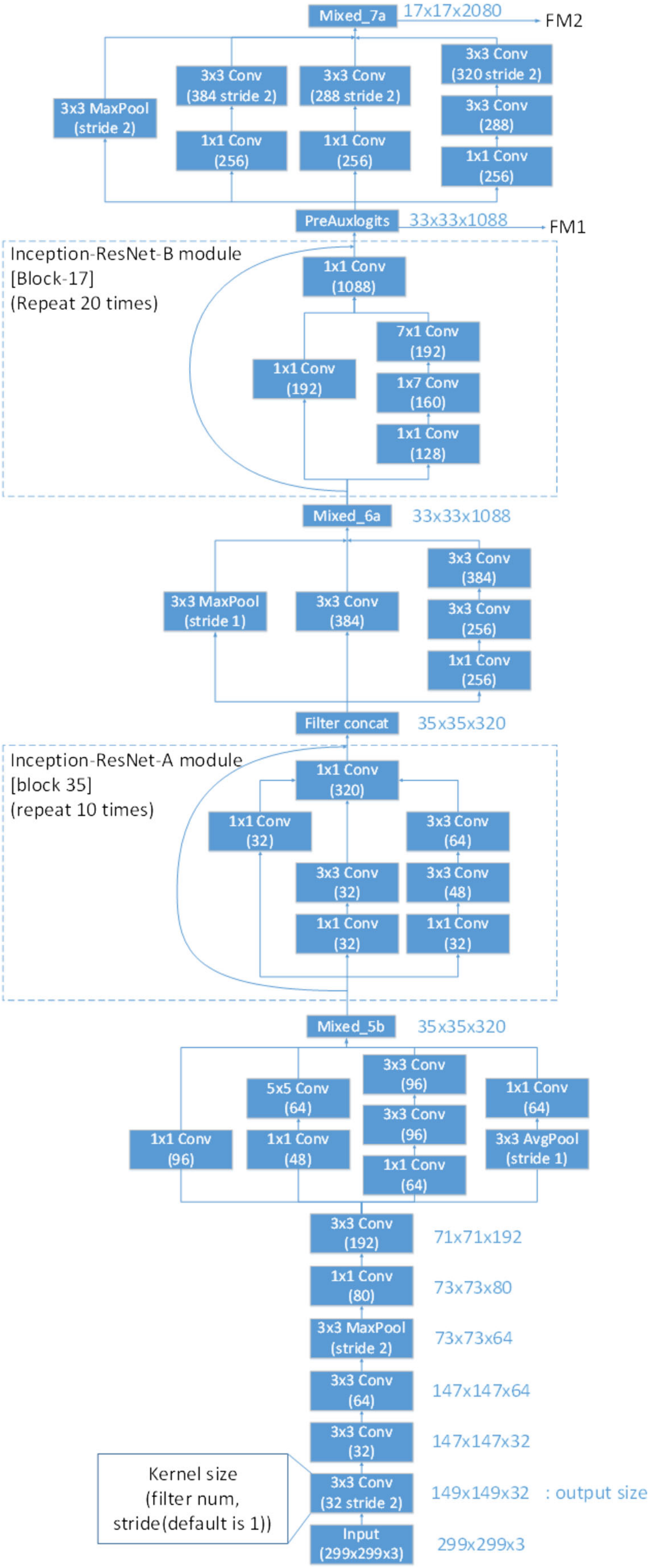


Fig. 3 Inception ResNet module structure

Fig. 4 Reduced Inception-ResNet module structure



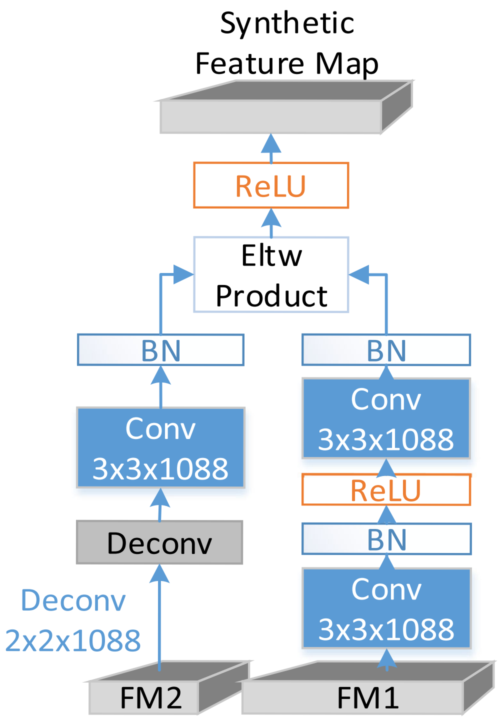


Fig. 5 Deconvolutional module

3.3 Region proposals with atrous convolution

Atrous convolution, which is a powerful tool in dense prediction tasks, allows us to effectively enlarge the field of view of filters to incorporate larger context without increasing the number of parameters or the amount of computation. Another advantage is that atrous convolution can be conveniently and seamlessly integrated to compute the responses of any layer.

The synthetic feature map is derived from the mid-level of the network that do not have big enough receptive field. To compensate for this, we designed our network to apply atrous convolution onto the synthetic feature map to enlarge the receptive field and inject context information. Figure 6 illustrates an example of feature extraction with atrous convolution in 2-D. Feature map *a* is produced from feature map *b* by an atrous convolution with rate $r = 2$. Feature map *a* corresponds to a receptive field of 9×9.

Atrous convolution with a rate r introduces $r - 1$ zeros between consecutive filter values, effectively enlarging the kernel size of a $k \times k$ filter to $k^{'} \times k^{'}$ by using Eq. (1) without increasing the number of parameters or the amount of computation.

$$k^{'} = k + (k-1)(r-1) \tag{1}$$

In our experiment, the atrous convolution is used with a 3×3 kernel size and rate $r = 2$. Therefore, $k^{'} = 5$. After block17, the output shape is $[33 \times 33 \times 1088]$. The corresponding receptive field of each element is 47×47. After atrous convolution, the output shape is still $[33 \times 33 \times 1088]$, but the receptive field of each element is 79×79; this means that we can obtain more context information.

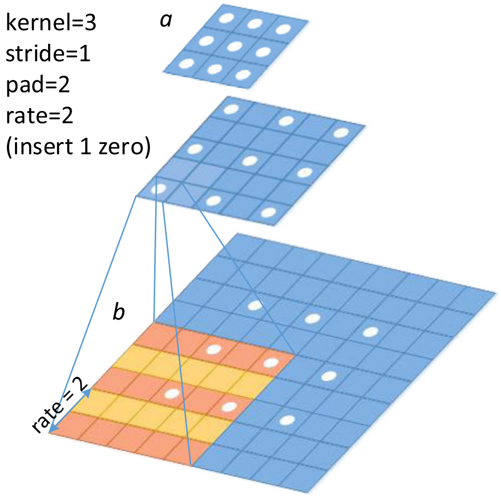


Fig. 6 Feature extraction with atrous convolution in 2-D

In Fig. 2, the right part illustrates the generation of proposals with atrous convolution based on the synthetic feature map and classification. To solve the multiple-scale detection problem, different anchors are used with four scales [0.25, 0.5, 1.0, 2.0] and three aspect ratios [0.5, 1.0, 2.0]. For training, a binary label is assigned to each box according to two different classes: pedestrian and non-pedestrian (i.e., background). Our loss function is defined as

$$L\left(p,\hat{p},b,\hat{b}\right) = L_{cls}(p,\hat{p}) + \lambda L_{reg}\left(b,\hat{b}\right), \tag{2}$$

where the classification loss L_{cls} is a softmax logistic loss over the two classes, $\hat{p}$ and p are the ture and predicted labels separately. The second task loss L_{reg} is the bounding box regression for positive boxes, $L_{reg}\left(b,\hat{b}\right) = R\left(b-\hat{b}\right)$, where R is the robust smooth-L1 loss as defined in Faster R-CNN [9], and $b = (b_x, b_y, b_w, b_h)$ represents the ground truth bounding box associated with a positive anchor. Then $\hat{b} = \left(\hat{b}_x, \hat{b}_y, \hat{b}_w, \hat{b}_h\right)$ denotes the predicted bounding box and λ is a trade-off parameter (empirically set to 10), for which a larger value places a stronger emphasis on good bounding box locations.

4 Experimental results

4.1 Implementation details

Our experiments are based on the open source framework of TensorFlow Object Detection [22], and the model we used is based on Inception-ResNet-v2 [18], which is pre-trained on the ILSVRC-CLS image classification dataset [23] . We change the Mixed_6a stage's convolution stride from 2 to 1 to increase the feature map resolution. However, the reduced stride also shrinks the receptive field. In order to offset, atrous convolution is used to enlarge the receptive field by increasing the dilation rate to 2. Table 1 illustrates the resolution of blocks of DIF R-CNN based on the pre-trained model. Then, the resulting model is fine-tuned using SGD with an initial learning rate of 0.0001, momentum of 0.9, and batch size of 1; the learning rate is reduced by a factor of 10 after 500,000 iterations and again after 700,000 iterations. Learning stops after 800,000 iterations. With the fine-tuned network of region proposals, non-maximum suppression (NMS) is adopted to eliminate highly overlapped bounding boxes with lower scores. The value of the intersection over union (IoU) is defined as

$$\mathrm{IoU}(b_1, b_2) = \frac{area(b_1 \cap b_2)}{area(b_1 \cup b_2)} \tag{3}$$

where b_1 and b_2 are the two proposal bounding boxes. In this paper, we set the threshold to 0.7 because it is experimentally demonstrated that this threshold can improve the detection efficiency without affecting the performance. The generated bounding boxes are ranked by their scores. If IoU > 0.7, it means that b_1 and b_2 highly overlap. Then the detection bounding boxes with the lower score will be eliminated. After using non-maximum suppression (NMS), a total of 100 proposals are generated for the second stage detection part. The full training and testing codes are built on Tensorflow v1.4.0. The entire network is trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB of memory.

4.2 Caltech pedestrian dataset

The Caltech dataset [7] is one of the most popular datasets for pedestrian detection. It contains 250 k frames captured from 10 h of urban traffic videos. The training data (set00-set05) consists of six training sets, each with 6–13 one-minute long sequence files, along with all annotation information. The testing data (set06-set10) consists of five sets, again along with all annotation information. The training and testing dataset have different video sequences with respect to the difficulty of pedestrian height, visibility, and aspect ratio. In our experiments, the training images are extracted with one out of every frame. There are 128,419 images for training and 4024 images for testing.

Table 1 The resolution of blocks of DIF R-CNN based on the pre-trained Inception-ResNet-v2 model

Layer	Resolution
Mixed_5b - block35	35 × 35
Mixed_6a - block17	33 × 33
Initial Feature Map (FM1)	33 × 33
Mixed_7a	17 × 17

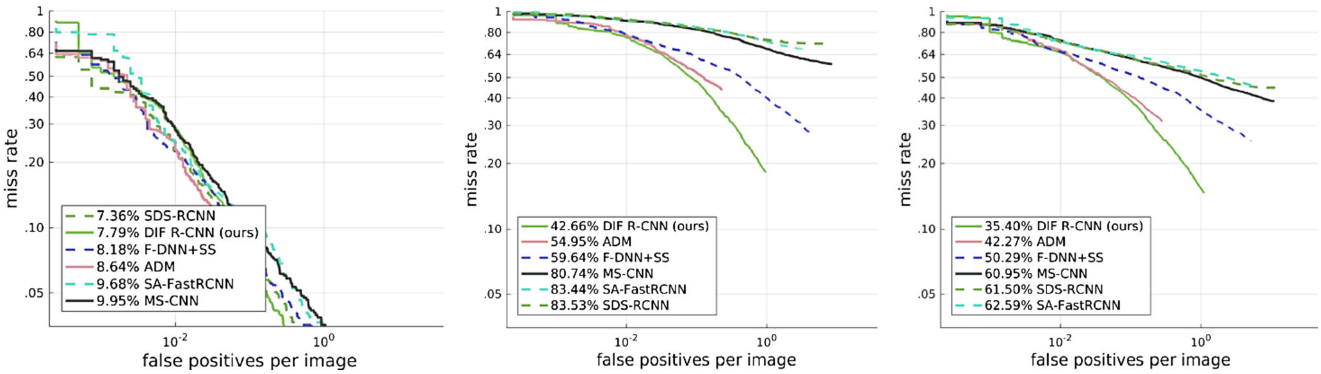
4.2.1 Detection evaluation on the Caltech pedestrian dataset

Our detection framework has been compared with five of the latest fully deep learning methods: ADM [1], F-DNN + SS [2], SDS-RCNN [3], MS-CNN [4] and SA-Fast RCNN [6]. Note that the competing methods [3, 4, 6] used the same training set. For ADM [1], a joint training set consisting of both Caltech and INRIA [24] training images is used. For F-DNN + SS [2], Caltech training set, ETH [25] and TudBrussels dataset [26] are used. We evaluate the performance of various detectors using the log-average miss rate (MR) which is computed by averaging the miss rate at false positive rates spaced evenly between the 10^{-2} to 1 false-positive-per-image (FPPI) range. The comparison results are evaluated for pedestrian instances of three scenarios: (a) reasonable case, i.e. no less than 50 pixels in height and at least 65% unoccluded, (b) far-scale case, i.e. shorter than 50 pixels, and (c) overall case, which is a combination of all scales and occlusions.

Figure 7a displays the quantitative results of the reasonable case. Our approach achieves a very low log-average miss rate of 7.79%, which is competitive with the state-of-the-art SDS-RCNN method [3]. However, SDS-RCNN [3] performs poorly in small-sized pedestrian detection. As exhibited in Fig. 7b, for the far-scale case, our method demonstrates a noticeable improvement (over12%) compared to the state-of-the-art method. Our approach achieves the lowest miss rate of 42.66%, where the next best method (ADM [1]) has a miss rate of 54.95%. Figure 7c presents the overall performance. Our method again significantly outperforms the others following a performance trend that is similar to that of the far-scale case; this makes sense because the number of far-scale instances dominates the overall pedestrian population of the Caltech benchmark. Our approach outperforms all comparison methods and achieves the lowest log-average miss rate of 35.40%, which clearly exceeds the two next best results: 42.27% for ADM [1] and 50.29% for F-DNN + SS [2].

Figure 8 presents four different example images in four rows from the Caltech testing set. The detection results of the related works [1–3] are shown in the figure and compared with ours. The small-sized instances, which are labelled as ignored instances with ground truth bounding boxes less than 20 pixels in height, are also shown. One can observe that most pedestrians, including far-scale instances, can now be detected correctly by our approach. In contrast, the state-of-the-art methods, such as ADM [1], F-DNN + SS [2] and SDS-RCNN [3], generate more false positives as well as more false negatives.

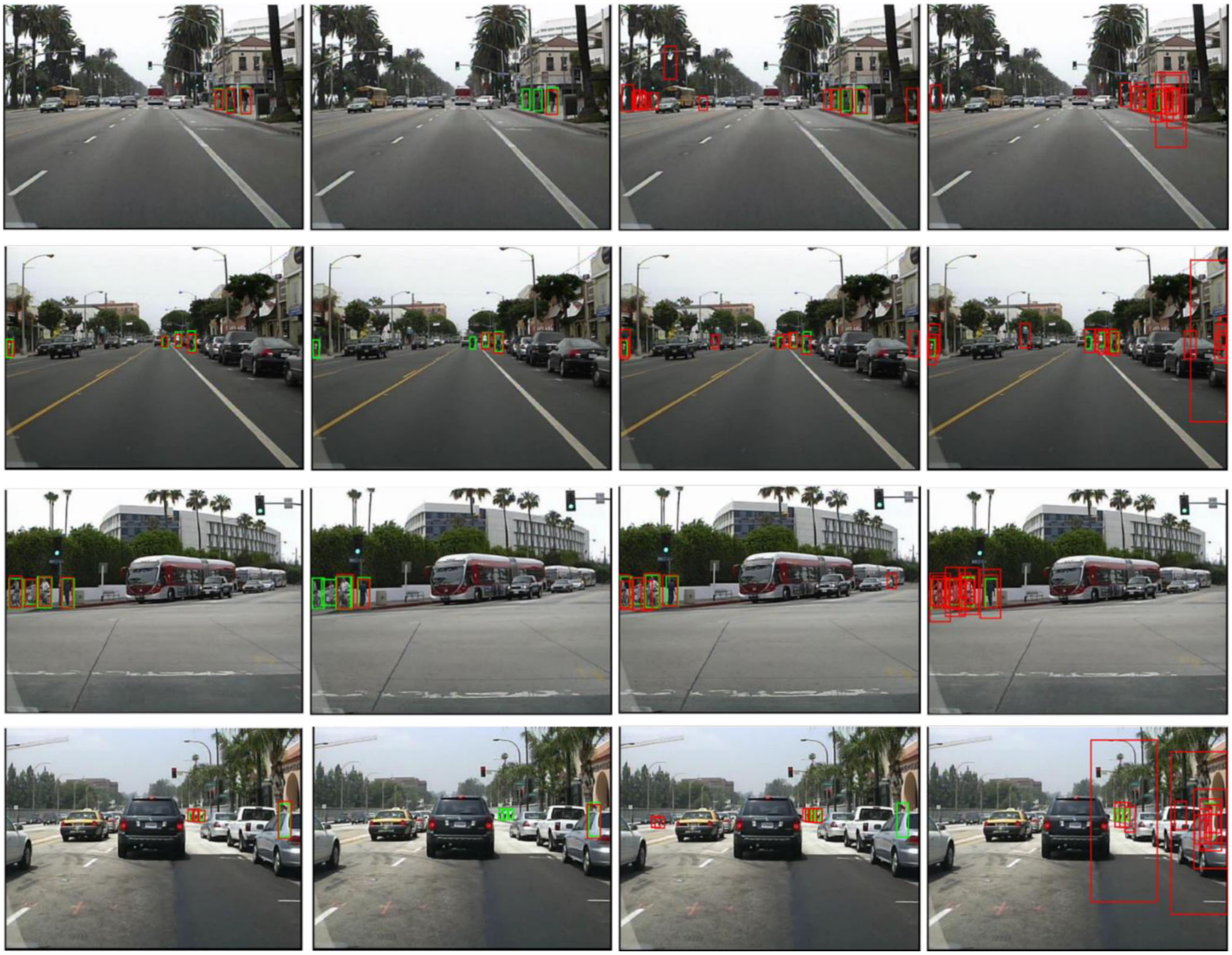
Table 2 shows a comparison report between our method and recent popular pedestrian detection methods in terms of



(a) Reasonable (person height > 50 pixels) **(b) far-scale (person height < 50 pixels)** **(c) overall**

Fig. 7 Comparisons of detection results (miss rate versus false positive per image) on the Caltech pedestrian dataset. **a** Reasonable (person height > 50 pixels), **b** far-scale (person height < 50 pixels), **c** overall

their computation efficiency. Images with 720×960 pixels are used which is the same size with other compared methods [2–4, 6], while the ADM [1] use images with 480×640 pixels for testing runtime. A single NVIDIA TITIAN X GPU is used for computation. The efficiency of DIF R-CNN surpasses the current state-of-the-art methods for pedestrian detection. This shows that our method outperforms both in accuracy and in runtime.



(a) DIF R-CNN (ours) **(b) ADM [1]** **(c) F-DNN+SS [2]** **(d) SDS-RCNN [3]**

Fig. 8 Visual comparisons of our detection results vs. those of two state-of-the-art methods on the Caltech benchmark. The red bounding boxes show the detection results, and the green bounding boxes denote the ground truth. **a** DIF R-CNN (ours), **b** ADM [1], **c** F-DNN+SS [2], **d** SDS-RCNN [3]

Table 2 Comparison of DIF R-CNN with other state-of-the-art methods based on the Caltech miss rate and runtime performance

Method	Miss rate (overall)	Runtime
SDS-RCNN [3]	61.50	0.21 s
MS-CNN [4]	60.95	0.4 s
SA-Fast RCNN [6]	62.59	0.59 s
F-DNN + SS [2]	50.29	2.48 s
ADM [1]	42.27	0.58 s
DIF R-CNN (ours)	35.40	0.18 s

4.2.2 Ablations experiments

This subsection is devoted to investigating the effectiveness of different components of DIF R-CNN. The simulation experiments are performed on the Caltech pedestrian dataset. The initial simulation is the original Inception-ResNet with Faster R-CNN framework. Then the atrous convolution has been added to make an improved version, which is the IncepResNet+Atrous. Finally, to form our DIF R-CNN, two changes are made. First, a deconvolutional module is added. Second, the base network is changed to reduced Inception-ResNet. Because the deconvolutional module includes once feature map upsampling, to match the deconvolution layer, the reduced network module is used with adopting prior layer instead of the last layer as the output feature map. Figure 9 shows the comparisons of the performance while training. During training, the mean Average Precision at 0.5 IoU (Intersection over union) threshold (mAP@0.5) is used as an evaluation metric. A quarter of training data are used for validation. The highest precision based on original Inception-ResNet is 62.59% after 843.8 k iterations, while the IncepResNet+Atrous achieves 66.13% after 804.4 k iterations. The DIF R-CNN brings a noticeable improvement that achieves the precision of 80.78% after 800.0 k iterations. The green curve (present DIF R-CNN) becomes almost flat from 720.0 k, but it decreases with further iterations. The precision drops to 77.06% after 819.6 k iterations. Therefore, we catch the model at 800.0 k and stop the training. Figure 10 presents the detection miss rates with these three simulations on Caltech testing set. From the original Inception ResNet to the proposed DIF R-CNN, comparing three simulations, the miss rate gets smaller and smaller as the neural network architecture is optimized.

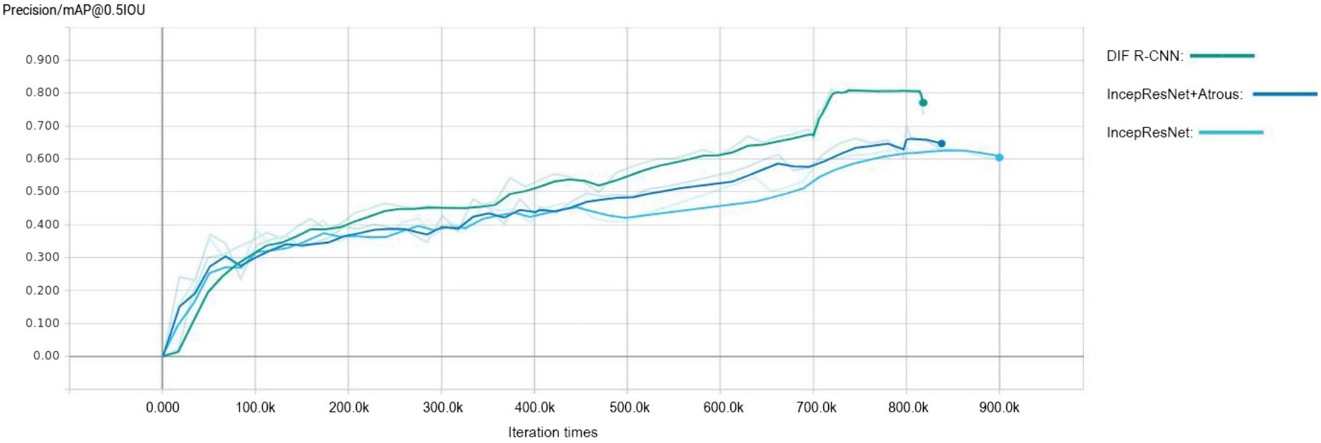
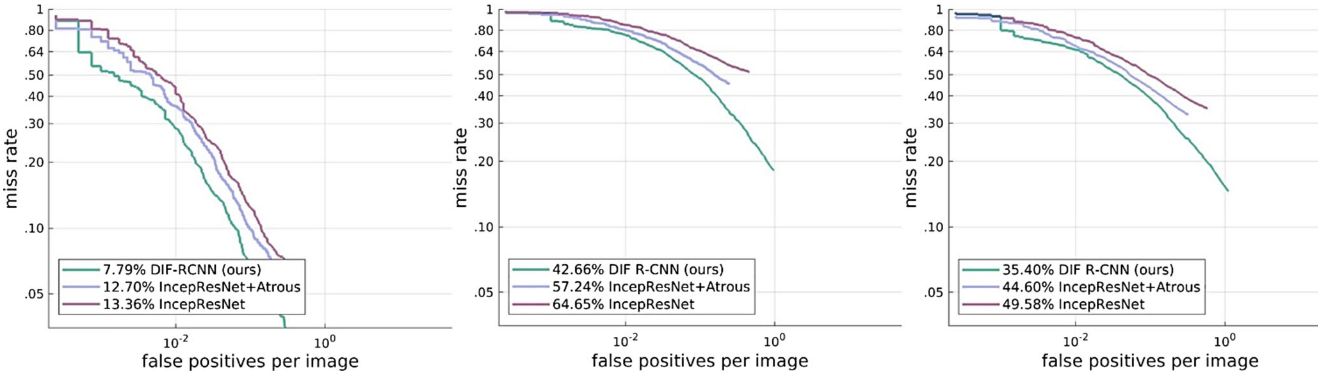


Fig. 9 The validation precision of original Inception-ResNet with Faster RCNN, IncepResNet+Atrous, and DIF R-CNN during training



(a) Reasonable (person height > 50 pixels) **(b) Far-scale (person height < 50 pixels)** **(c) Overall**

Fig. 10 Comparisons of original Inception-ResNet with Faster RCNN, IncepResNet+Atrous, and DIF R-CNN detection results on the Caltech pedestrian testing dataset. **a** Reasonable (person height > 50 pixels), **b** Far-scale (person height < 50 pixels), **c** Overall

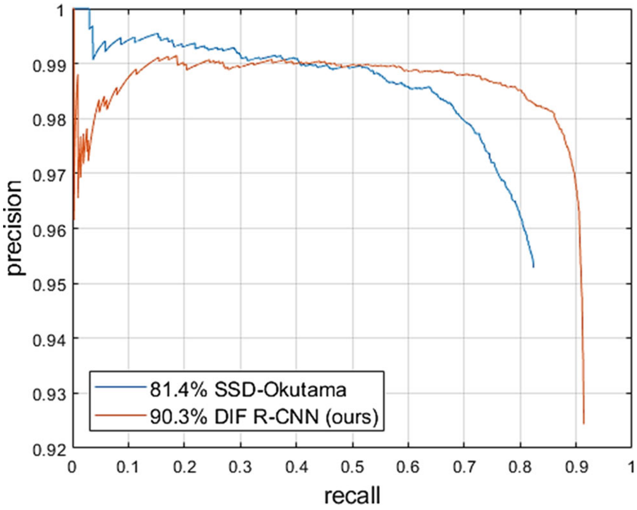
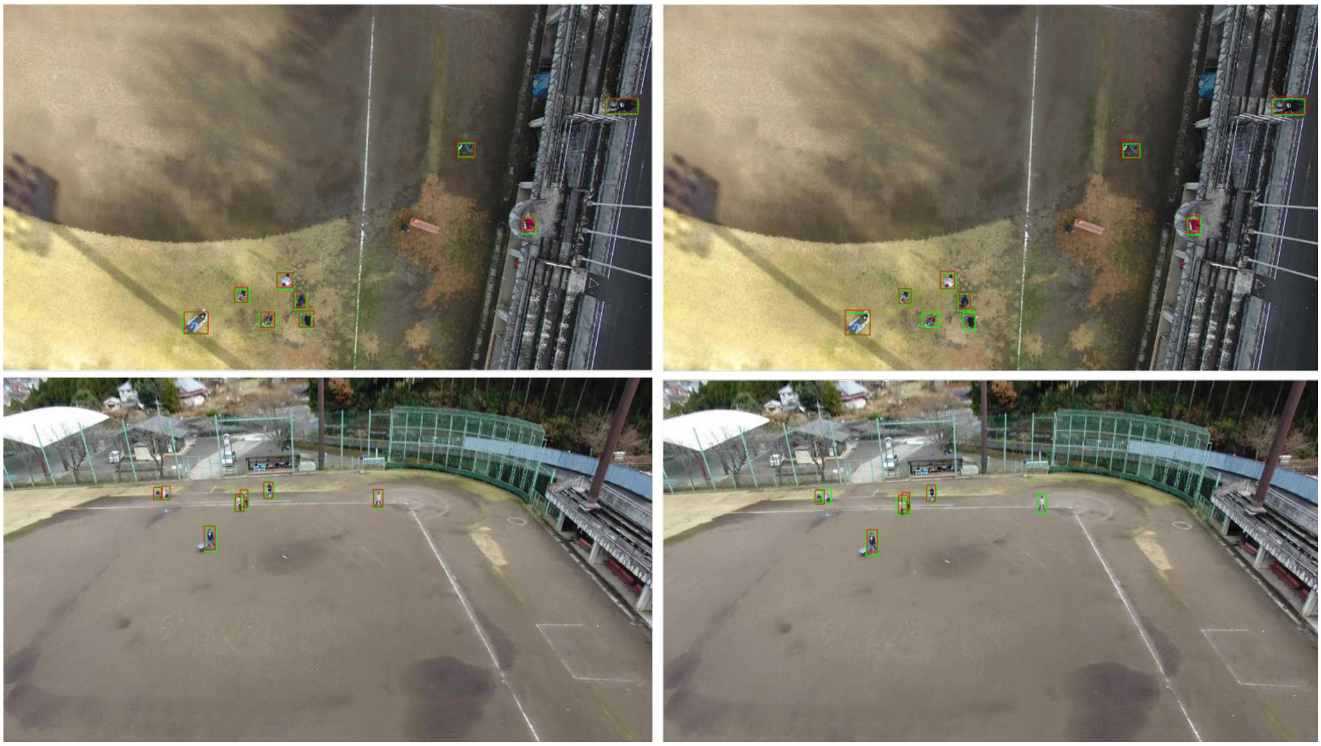


Fig. 11 The pedestrian detection comparison of DIF R-CNN and the baseline method on the Okutama human-action dataset

4.3 Okutama human-action datasets

In this section, the results obtained by our algorithm are presented based on another new aerial view dataset, Okutama human-action dataset [5], which is used for human detection and human action understanding from a real-world aerial view. In our case, only the human detection task is considered. The Okutama dataset contains a total of 43 video sequences (33 training video sequences and 10 testing video sequences) at 30 FPS and 77,365 frames in 4 K resolution. These sequences were recorded using 2 UAVs flying at altitudes varying between 10 and 45 m and with camera angles of 45 or 90 degrees. This dataset is fully-annotated by providing all bounding boxes of the label *Pedestrian*. The training images are extracted from every 10 frames. In total, we get 5904 images with 3480×2160 pixels. Since the annotation of the testing set are not available, the data is split with the common trade-offs: 70% of the data into training subset and 30% of the data into validation subset. This dataset is more challenging due to the aerial view angle and rather small person's height in large-sized images.

Regarding training and fine-tuning, the model is trained with an initial learning rate of 0.0003, momentum of 0.9, and batch size of 1. The learning rate is reduced by a factor of 10 after 170,000 iterations and learning stops after 200,000 iterations. Since the ratio of pedestrian size over image size of this dataset is small, the anchors are set from 0.125 instead of 0.25. The mean Average Precision at 0.5 IoU threshold (mAP@0.5) is used as an evaluation metric like many other object detection methods [22]. The baseline method SSD which is given with Okutama dataset [5] achieved 81.4% of mAP for the validation set. Since the labels for testing set are not available, the authors recommended to use the validation set for comparisons. Our approach achieves 90.3% in terms of mAP, resulting significant improvement of 8.9%. Figure 11


(a) Detection results of DIF R-CNN (ours)

(b) Detection results of baseline method, SSD-Okutama [6].

Fig. 12 Visual comparisons of our detection results vs. the baseline methods on the Okutama human-action dataset. The red bounding boxes show the detection results, and the green bounding boxes denote the ground truth. **a** Detection results of DIF R-CNN (ours), **b** Detection results of baseline method, SSD-Okutama [6].

Table 3 Comparison of DIF R-CNN and the baseline method based on the Okutama human-action dataset in terms of mAP@0.5 and runtime performance

Method	mAP (%)	Runtime
SSD-Okutama [5]	81.4	0.028 s
DIF R-CNN (ours)	90.3	0.22 s

shows the precision-recall curve of our proposed DIF R-CNN and the baseline method. Figure 12 is the visual comparisons of the detection results, by using two different example images in two rows from the Okutama validation set. The SSD method produces more false positives than those of our approach.

Table 3 shows a comprehensive comparison between our method and the baseline method in terms of their performance and computation efficiency. We tested on a single NVIDIA TITIAN X GPU. Although the SSD method achieved better speed of 0.028 s per image (our network takes 0.22 s per image), the precision gap between SSD and ours is significant, almost 9%. In order to realize the real-time detection, we plan to improve the speed in the future works.

5 Conclusion and future works

In this paper, a novel context-aware Deconvolution Integrated Faster R-CNN (DIF R-CNN) is proposed for pedestrian detection, especially for small-sized instances. It is based on the Faster R-CNN pipeline with a deconvolutional module and atrous convolution adopted to capture more context information. A synthetic feature map is generated to provide both visual details and semantic context representation. Furthermore, the state-of-the-art CNN model Inception-ResNet is integrated into our approach. Extensive experiments demonstrated that the proposed DIF R-CNN is superior in detecting small-sized pedestrian instances and achieves comparable or better performance relative to other state-of-the-art methods on several challenging datasets. For future works, we plan to improve the detection speed by further simplifying the network structure and making the model size smaller with network compression technology and to resolve the challenges of detecting heavily occluded pedestrian instances.

Acknowledgements This work was supported by Basic Research Project in Science and Engineering through the Ministry of Education of the Republic of Korea and National Research Foundation of Korea (National Research Foundation of Korea 2017-R1D1A1B04-031040).

References

1. Zhang X, Cheng L, Li B, Hu H-M (2018) Too far to see? Not really!—pedestrian detection with scale-aware localization policy. IEEE Trans Image Process 27(8):3703–3715
2. Du X, El-Khamy M, Lee J, Davis L (2017) Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 953-961. https://doi.org/10.1109/WACV.2017.111
3. Brazil G, Yin X, Liu X (2017) Illuminating pedestrians via simultaneous detection & segmentation. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp 4950–4959. https://doi.org/10.1109/ICCV.2017.530
4. Cai Z, Fan Q, Feris RS, Vasconcelos N (2016) A unified multi-scale deep convolutional neural network for fast object detection. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham. https://doi.org/10.1007/978-3-319-46493-0_22
5. Barekatain M, Marti M, Shih H-F, Murray S, Nakayama K, Matsuo Y, Prendinger H (2017) Okutama-action: an aerial view video dataset for concurrent human action detection. In: 30th IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 2153–2160. https://doi.org/10.1109/CVPRW.2017.267
6. Li J, Liang X, Shen S, Xu T, Feng J, Yan S (2018) Scale-aware fast R-CNN for pedestrian detection. IEEE Trans Multimedia 20(4): 985–996
7. Dollár P, Wojek C, Schiele B, Perona P (2009) Pedestrian detection: a benchmark. In: 2009 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 304–311. https://doi.org/10.1109/CVPR.2009.5206631
8. Sermanet P, Kavukcuoglu K, Chintala S, LeCun Y (2013) Pedestrian detection with unsupervised multi-stage feature learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 3626–3633. https://doi.org/10.1109/CVPR.2013.465
9. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
10. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC (2017) DSSD: deconvolutional single shot detector. arXiv:1701.06659 [cs.CV]. http://arxiv.org/abs/1701.06659. Accessed 23 Jan 2017
11. Long J, Shelhamer E, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(4):640-651. https://doi.org/10.1109/TPAMI.2016.2572683
12. Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 447–456. https://doi.org/10.1109/CVPR.2015.7298642
13. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834-848. https://doi.org/10.1109/TPAMI.2017.2699184
14. Holschneider M., Kronland-Martinet R., Morlet J., Tchamitchian P. (1990) A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform. In: Combes JM., Grossmann A., Tchamitchian P. (eds) Wavelets. inverse problems and theoretical imaging. Springer, Berlin, Heidelberg, pp 286–297. https://doi.org/10.1007/978-3-642-75988-8_28
15. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (NIPS). Commun. ACM, pp 1097–1105. https://doi.org/10.1145/3065386

16. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV]. http://arxiv.org/abs/1409.1556. Accessed 4 Sep 2014
17. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 1–9. https://doi.org/10.1109/CVPR.2015.7298594
18. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First Conference on Artificial Intelligence. AAAI Press, pp 4278–4284. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806. Accessed 12 Feb 2017
19. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
20. Zhang L, Lin L, Liang X, He K (2016) Is faster R-CNN doing well for pedestrian detection?. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9906. Springer, Cham, pp 443–457. https://doi.org/10.1007/978-3-319-46475-6_28
21. Pinheiro PO, Lin TY, Collobert R, Dollár P (2016) Learning to refine object segments. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham, pp 75–91. https://doi.org/10.1007/978-3-319-46448-0_5
22. Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S (2017) Speed/accuracy trade-offs for modern convolutional object detectors. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 3296–3297. https://doi.org/10.1109/CVPR.2017.351
23. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
24. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 886–893. https://doi.org/10.1109/CVPR.2005.177
25. Ess A, Leibe B, Schindler K, Van Gool L (2008) A mobile vision system for robust multi-person tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 1–8. https://doi.org/10.1109/CVPR.2008.4587581
26. Wojek C, Walk S, Schiele B (2009) Multi-cue onboard pedestrian detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 794–801. https://doi.org/10.1109/CVPR.2009.5206638

Han Xie received the B.S. degree in Electronics Engineering from the Wuhan University of Science and Technology of China, in 2015. Currently, she is working toward the master to Ph.D. degree with the Department of Electronic and Communication, Hanyang University of Korea. She focuses on computer vision and pattern recognition.

Yunfan Chen received the B.S. degree in Electronics Engineering from the Wuhan University of Science and Technology of China, in 2015. Currently, she is working toward the master to Ph.D. degree with the Department of Electronic and Communication, Hanyang University of Korea. She focuses on computer vision and pattern recognition.

Hyunchul Shin received the B.S. degree in electrics engineering from the Seoul National University, Seoul, Korean, in 1978, the M.S. degree in Electrical Engineering from KAIST, Korea, in 1980, and the Ph.D. degree in Electrical Engineering and Computer Science, U.C. Berkeley, CA, USA, in 1987. Currently he is a Professor with Hanyang University, Korea. He has more than 180 publications in journals and international conferences. His current research interests include computer vision and artificial intelligence.