Pattern Recognition 115 (2021) 107846

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Pedestrian detection with super-resolution reconstruction for low-quality image

Yi Jin[a,1], Yue Zhang[a,b,1], Yigang Cen[a,b,*], Yidong Li[a], Vladimir Mladenovic[c], Viacheslav Voronin[d]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China
[c] Faculty of Technical Sciences, University of Kragujevac, Cacak, Serbia
[d] Center for Cognitive Technology and Machine Vision, Moscow State University of Technology "STANKIN", Moscow, Russian Federation

ARTICLE INFO

Article history:
Received 15 July 2020
Revised 6 December 2020
Accepted 2 January 2021
Available online 14 February 2021

Keywords:
Pedestrian detection
Low-quality
SRGAN
Faster R-CNN

ABSTRACT

Pedestrian detection has emerged as a fundamental technology for autonomous cars, robotics, pedestrian search, and other applications. Although many excellent object detection algorithms can be used for pedestrian detection, it is still a challenging problem due to the complicated real-world scenarios, e.g., the detection of pedestrians in low-quality surveillance videos. In this paper, we aim to study the challenging topic of pedestrian detection in low-quality images. Low-quality images are interpreted as those taken with a low-resolution camera, heavy weather or a blurred scene, making it difficult to distinguish pedestrians from the background. To solve this problem, we first introduce a dataset called playground (PG) for low-quality image detection. Images from PG are shot using two different camera views, and pedestrian images are taken at different periods, including day and night. The dataset contains a total of 5,752 images with 31,041 annotations. The average size of the pedestrian is 87×41 and the image size is 480×640, indicating that these images are taken from very long distances. Then, we propose a super-resolution detection (SRD) network to enhance the resolution of low-quality images that can help distinguish pedestrians from the blurred background. Finally, based on these enhanced images, we adopt and improve the Faster R-CNN network to help relocate occluded pedestrians. Experimental results on this new dataset proved the efficiency and effectiveness of our algorithm on low-quality images.

1. Introduction

Pedestrian detection has always been a hot issue in computer vision research. Given an arbitrary image, the goal of pedestrian detection is to determine whether or not there are any pedestrians in the image, and if a pedestrian is in fact present, return the image location of the pedestrian. Pedestrian detection is one of the most fundamental problems with a wide range of real-world-applications. It is usually combined with pedestrian tracking [1,2], pedestrian recognition [3,4], and other technologies and is used in applications such as automotive driverless systems (ADAS) [5,6], intelligent robots [7,8], intelligent video surveillance [9], human behavior analysis [10], passenger flow statistics system [11], and intelligent transportation [12]. Pedestrian detection algorithms mainly include handcrafted feature-based methods (e.g. histogram of oriented gradients (HOG) with boosting classifier [13,14]), hybrid methods (e.g. handcrafted features for proposed generation and deep CNN used to classify these proposals [15–17]), and CNN-based methods [18–20]. With the development of deep learning, more available pedestrian public datasets have been built, and pedestrian detection has achieved remarkable accuracy.

However, pedestrian detection is still a challenging task due to the wide variation in the appearance of the person, background, and visibility conditions. A particular challenging example is pedestrian detection from low-quality surveillance video. First, it is difficult to distinguish pedestrians from the background from surveillance camera footage. In many surveillance scenarios, pedestrians are very small and dense with serious occlusion. Some objects are so similar in appearance, shape, color, and texture to the human body that existing models trained on high-quality images cannot accurately distinguish pedestrians. Second, few algorithms have been developed for pedestrian detection in low-quality images. For this task, we propose a super-resolution

* Corresponding author at: School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China.
E-mail address: ygcen@bjtu.edu.cn (Y. Cen).
[1] The first two authors (Yi Jin and Yue Zhang) contribute equally.

https://doi.org/10.1016/j.patcog.2021.107846
0031-3203/

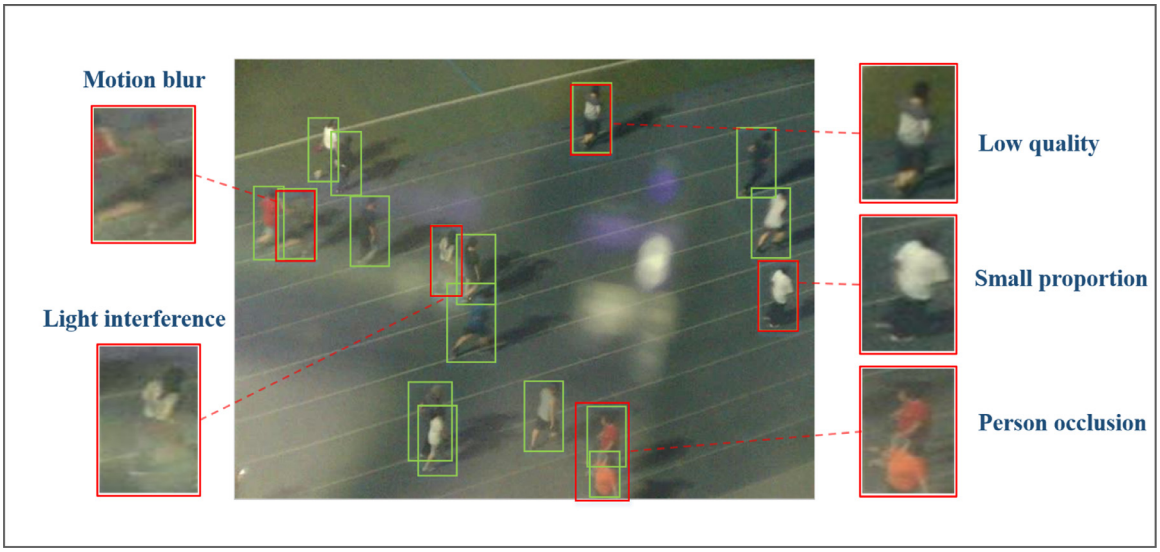


Fig. 1. Main challenges of the newly introduced PG dataset.

pedestrian detection (SRD) algorithm. Specifically, we first reconstruct the image with super-resolution to help distinguish the pedestrian from the background. Then, based on Faster R-CNN, we employ the repulsion loss [21] to reduce the distance between the regression box and the ground truth and push away the distance between the regression box and surrounding boxes so as to improve the positioning accuracy.

With the improvement in algorithm performance, it is necessary to propose more challenging datasets that will motivate the continuous development of new algorithms. Most of the existing datasets, such as Cityperson [20], and KITTI [22], are collected by vehicle-mounted cameras. These dataset images have larger pedestrian image sizes, but there are few pedestrians, and some images have no pedestrians. Furthermore, current pedestrian datasets are derived from relatively clear camera shots and rarely include images from surveillance cameras. In addition, most of the images in these datasets are only taken during the day, and very few are taken at night. Overall, the currently used datasets contain fewer pedestrians, the images are clearer, and persons are found in a wide range of scenarios, which is not beneficial for pedestrian detection from surveillance camera images. Thus, we develop a low-quality and diverse pedestrian detection dataset called PG[2] that is shot in a playground. This dataset was created using two cameras that shoot simultaneously both during the day and at night. The dataset contains many challenges, such as low-quality pedestrians, motion blur, light interference, and pedestrian with occlusion, as shown in Fig. 1. The dataset consists of 5,752 images and 31,041 bounding boxes. The training set includes 2,754 images and 24,334 bounding boxes. Three test sets are collected at different times. The easy subset includes 320 images and 1,280 bounding boxes that were shot at the same time as the training set (in the summer). The medium-difficulty (medium) subset was shot in the winter and includes 475 images and 2,180 bounding boxes. The hard subset consists of 498 images and 3,247 bounding boxes that were shot at night in the summer. We show an example of the proposed SRD method, including super-resolution reconstruction, the designed anchors, and bounding box regression. The detection results obtained from the images in different time periods prove the effectiveness and efficiency of the detectors trained with the new dataset and demonstrate promising performance.

[2] https://github.com/IvyYZ/PG.

The contributions of this paper are as follows:

- This paper proposes a new end-to-end pedestrian detection method called the super-resolution detection (SRD) network that aims to solve the low-quality and occlusion problems in intelligent video surveillance. Specifically, we perform super-resolution reconstruction of the image to help distinguish the pedestrians from the background and adopt an improved Faster R-CNN network to relocate occluded pedestrians.
- To verify the effectiveness of the proposed SRD algorithm, a new low-quality playground (PG) dataset for pedestrian detection is collected that provides dense and occluded pedestrians with light interference and motion blur in the surveillance images. The dataset was shot from two different cameras with different perspectives in two seasons and includes day and night surveillance video pedestrian images. There are a total of 5,752 images with 31,041 annotations, i.e., 5.39 annotations per image.
- Compared with the state-of-the-art methods, our proposed SRD method achieves higher accuracy of pedestrian detection based on the PG dataset. In particular, we demonstrate improved results for more difficult detection cases (light interference and occluded), and overall higher localization precision.

The rest of this paper is organized as follows. In Section 2, related works on the super-resolution method and detection method are reviewed. Description of the proposed dataset is presented in Section 3. The concept, framework, and details of the proposed SRD are presented in Section 4. In Section 5, the datasets and the experimental results are presented. Section 6 concludes the paper.

2. Related work

2.1. Pedestrian detection

Prior to the success of deep convolutional neural networks in computer vision tasks [23–25], a large number of manual features such as HOG features [13], Haar features [26] and ICF features [27] were applied to pedestrian detection. With the development of deep convolutional networks (e.g. VGG [28], ResNet [29]), many visual tasks have adopted deep learning to extract powerful features, including image classification [30,31], image feature representation [32,33], and instance segmentation [34,35].

The successes of deep learning have also lead to its incorporation in object detection. In terms of image target detection, R-CNN [36] was the first to use CNN for target detection. The candidate region generation algorithm is used instead of a sliding window to extract the CNN feature and enhance robustness, and SVM was employed to implement classification. Fast R-CNN [18] uses the multitasking loss function to unify feature extraction, classification tasks, and box regression into a framework. SSD [37] adopts multi-scale feature images, and samples candidate regions on feature maps of different scales, improving the detection recall rate and the detection effect of small targets. Faster R-CNN [19] adopts the RPN layer on the basis of Fast R-CNN to obtain proposals, effectively improving the detection speed. YOLO [38] directly performs feature extraction, candidate frame regression, and classification in the same convolution network, simplifying the network structure. YOLO v2 [39] employs k-means to cluster the training set box to generate a suitable a priori box. YOLO v3 [40] uses FPN-like multi-scale prediction based on YOLO v2 and adopts a better basic classification network in small target classifications. YOLO v4 [41] combines some existing CNN technologies to achieve improved results. In addition, CSP [42] uses box-free settings and does not use the complex postprocessing strategies used in recent keypoint-pairing based detectors.

In terms of image pedestrian detection, the Faster R-CNN network has been widely applied. Li et al. [43] merged a large-scale sub-network and a small-scale sub-network into a unified framework, and the candidate region height was used to estimate the scale-aware weights of the two sub-networks. Tian et al. [44] proposed deep parts to address the issues related to occlusion; this approach divides the human body into multiple parts and detects them separately. Wang et al. [21] designed a repulsion function that effectively solves the problem of intra-class occlusion. Mao et al. [45] proposed HyperLearner's pedestrian detection algorithm by introducing apparent-to-semantic channels, temporal channels, and depth channels. Zhang et al. [46] proposed an attention network with self or external guidances based on the Faster R-CNN architecture for occluded pedestrian detection. Han et al. [4] proposed a re-ID driven localization refinement framework to refine detection boxes for pedestrian detection. Zhang et al. [47] proposed an active pedestrian detector that explicitly operates over multiple-layer neuronal representations to deliver accurate prediction of pedestrian locations. Chen et al. [48] presented a novel hierarchical knowledge distillation framework to learn a lightweight pedestrian detector that significantly reduces the computational cost and still maintains high accuracy at the same time. Liu et al. [42] proposed CSP (center and scale prediction) that simplified object detection as a straightforward center and scale prediction task through convolutions with a box-free setting. Pang et al. [49] introduced a misguided attention network that fits naturally into popular pedestrian detection pipelines. Considering that there are many pedestrians in the surveillance video and the pedestrians are small, we apply repulsion loss in the Faster R-CNN framework [21] to accurately locate the rectangular box of pedestrians.

2.2. Image super-resolution

Super-resolution focuses on recovering high-resolution images from a low-quality observation to achieve the effect of enhancing image details. It includes multiple low-quality images to synthesize a high-resolution image and a single low-quality image to obtain high-resolution images. In this work, we adopt single-image super-resolution reconstruction technology (SISR). SRCNN [50] was the first to use CNN [28] for super-resolution reconstruction and adopts the three-end depth full convolution network end-to-end to achieve the most advanced high-resolution performance. Chao et al. [51] proposed FSRCNN to let the network directly learn the deconvolution filter, further improving the speed and accuracy. VDSR [52] was the first method to introduce global residuals into super-resolution, improving the training speed significantly, and has greatly improved the PSNR and SSIM evaluation indicators. SRGAN (superresolution generative adversarial network) [53] introduced GAN into high-resolution reconstruction and relied on a loss function closer to the perceptual similarity to recover visually more convincing HR images. EDSR [54] applied residual skip connections and eliminated range flexibility from the networks by normalizing the features. SRDenseNet [55] applied dense block in high-resolution reconstruction. Zhang et al. [56] combined local/global residual and dense skip connections in their RDN. Li et al. [57] proposed SRFBN to use hidden states in an RNN with constraints to achieve such feedback manner so as to refine low-level representations with high-level information. Zhang et al. [58] proposed the very deep residual channel attention networks (RCAN) method that allows abundant low-frequency information to be bypassed through multiple skip connections so that the main network focuses on learning high-frequency information. Dai et al. [59] proposed a second-order attention network (SAN) for more powerful feature expression and feature correlation learning. Although many different methods have been developed for super-resolution networks, most network structures adopt skip connections that fuse high-level semantic features and low-level texture features. Since SRGAN can improve the realism of the recovered images by using perceptual loss and adversarial loss, we adopt SRGAN to perform super-resolution processing on our PG dataset.

2.3. Overview of previous datasets

The MIT dataset [60] is the earliest pedestrian dataset with 924 persons; it contains both front and back views and has no negative samples. The average size of the pedestrian is approximately 128×64. In the INRIA dataset [61] static pedestrian database, the training set contains 288 positive samples and 1,218 images of negative samples. Most pedestrians are standing with a height of more than 100 pixels. The USC database [67] contains three datasets (USC-A, USC-B, and USC-C). The USC-A includes 205 images with 313 standing pedestrians with no occlusion. The USC-B dataset is mainly from the CAVIAR video library and includes 271 pedestrians with various perspectives, and mutual occlusion between the pedestrians. The USC-C includes 100 images and 232 pedestrians (multi-angle) with no mutual occlusion. Daimler pedestrian detection benchmark [62] is obtained by the vehicle-mounted camera and includes 21,790 training images with the minimum height of pedestrians of 72 pixels. TUD database [64] provides image pairs for calculating the optical flow information for the evaluation of the role of motion information in pedestrian detection. The training set includes 1,092 images with 1,776 pedestrians and 192 pairs of non-pedestrian images. NICTA pedestrian database [65] is currently a large-scale static pedestrian database, with 25,551 images containing a single person. ETH pedestrian database [66] constructs a pedestrian database based on the binocular vision for multi-person pedestrian detection and tracking research. It adopts a pair of vehicle-mounted cameras to shoot with a frame rate of 13-14 FPS. CVC-01 dataset includes 1,000 pedestrian samples and 6,175 non-pedestrian samples. Currently, the most popular datasets are the Caltech-USA [63], Cityperson [20], and KITTI [22] datasets that are relatively large datasets proposed in recent years. They are shot by the vehicle-mounted cameras and have been used as benchmark datasets for testing many algorithms. However, the pedestrian distribution in these datasets is scattered. In some images, there is no pedestrian. WiderPerson [68] is collected from crowd scenarios with highly occluded pedestrians. A sufficient amount of training data enables the development of deep models and the demonstration of their representation power

Table 1
Comparison of pedestrian images in different datasets.

Dataset	train num	BBoxes	Cameras	persons	resolutions	annotations	time	source
MIT [60]	924	924	1	-	64 × 128	full	day	-
INRIA [61]	1,832	2,416	1	1,126	640 × 480	full	day	static person
Daimler [62]	21,790	56,492	1	4,800	640 × 480	full	day	vehicle-mounted camera
Caltech [63]	250,000	350,000	1	-	640 × 480	full,visible	day	vehicle-mounted camera
TUD [64]	2,568	1,776	-	1,776	640 × 480	full	day	vehicle-mounted camera
NICTA [65]	25,551	-	-	-	-	-	day	static person
ETH [66]	924	1	-	-	640 × 480	full	day	vehicle-mounted camera
CVC-01	7,175	-	-	1,000	640 × 512	full	day,night	-
USC-A [67]	205	313	-	313	-	-	day	Internet
USC-B [67]	54	271	-	271	-	-	day	vehicle-mounted camera
USC-C [67]	100	232	-	232	-	-	day	Internet
KITTI [22]	3,712	2,322	2	2,322	1240 × 376	full	day	vehicle-mounted camera
CityPersons [20]	2,975	19,654	-	19,654	2048 × 1024	full,visible	day	vehicle-mounted camera
WiderPerson [68]	8,000	236,073	-	236,073	-	full	day	Internet
PG	2,754	24,334	2	24,334	640 × 480	full,visible	day,night	Surveillance cameras

in person detection. Although current algorithms have achieved high accuracy on these datasets, pedestrian detection is far from being solved and widely applied in real scenarios. Therefore, it is necessary to analyze the limitations of existing datasets. Furthermore, existing pedestrian detection algorithms are trained on high-quality images. Meanwhile, for the low-quality images, these algorithms fail to distinguish the pedestrians from the background.

3. PG dataset

A comparison of pedestrian images from different datasets such as Caltech, USC, KITTI, CityPersons, WiderPerson (Table 1) shows that current datasets are limited in four ways: 1) The images are high quality, with a large proportion of individual pedestrians in the image, and pedestrians that can be separated from the background. 2) Pedestrian images are clear, the pedestrian appearance in the dataset can be seen clearly, and there is no motion blur or light interference. 3) There is an absence of images from the monitoring perspective. These datasets include images obtained by vehicle-mounted cameras and online and lack a top-down perspective. 4) Most of the existing datasets are collected in daylight, while pedestrian datasets at night are rarely collected. To address the abovementioned limitations, we collect a new multi-time person detection dataset called Playground (PG) by simulating the surveillance scenarios as much as possible.

3.1. Description of the PG dataset

3.1.1. Capacity

The position of the cameras that we deployed is greater than 3 meters above the ground. We selected 2 days with different weather conditions for image collection. For each day, 1 h of the video was taken in the afternoon and 1 h at night. Our final raw video set contains 4 h of video. Finally, 5,752 images were selected, and 31,041 bounding boxes were annotated. The training set was collected in the summer and includes 4,459 images and 24,334 bounding boxes. The test set includes 1,293 images and 6,707 bounding boxes. The test set is divided into three parts according to the time of collection, as shown in Fig. 3. The easy subset includes 320 images with 1,280 bounding boxes that are shot during the day in the summer. The medium-difficulty subset is shot during the day in the winter. It includes 475 images and 2,180 bounding boxes. The hard subset consists of 498 images and 3,247 bounding boxes that were shot at night in the summer.

3.1.2. Scale

We generate statistics on the pedestrian height of the proposed dataset, as shown in Fig. 4. It is observed from Fig. 4(a) that the pedestrian heights are distributed in the range of 20-140 pixels, and pedestrians with a height in the range of 75–80 pixels make the largest contribution to the overall distribution. For the convenience of observation, we draw a pie chart for the ratio of quantitative height, as shown in Fig. 4(b). Pedestrians with heights between 70 and 80 pixels account for 23% of the total dataset, those with heights of $<$70 pixels account for 24%, and pedestrians with the heights of $>$120 pixels only account for 2%. Furthermore, for the same pedestrian size (such as 80-90 pixels) in different datasets, the pedestrian contours in the proposed dataset are more blurred than those in others.

3.1.3. Position

The persons on the Caltech-USA [63] and CityPersons [20] datasets are distributed in a narrow band across the center of the image. This is because these datasets are captured by vehicle-mounted cameras, and pedestrians are mainly concentrated on both sides of the road. The persons in the WiderPerson dataset [68] are distributed in the lower half of the image. By contrast, we calculate the center coordinates of all of the pedestrian boxes (the 31,041 bounding boxes) and plot the resulting heat map, as shown in Fig. 5(a). It is observed that the pedestrians in our dataset can be located in any position of the image. Unlike other datasets from horizontal shooting, the images that we collect are taken from a top-down camera without the sky background so that pedestrians are evenly distributed across the entire image. Specifically, we draw a scatter for the pedestrians in different test sets. The pedestrians in the images of the easy and hard datasets are distributed in all corners of the images, while the pedestrians in the images of the medium-difficulty dataset are mainly concentrated in the upper half of the image, indicating that our data are diverse.

Compared with existing datasets, we summarize the new features in PG as follows: 1) larger number of bounding boxes, lower-quality images with light interference. It is clear that the PG dataset presents a more challenging and realistic pedestrian detection task. 2) Multiple time periods result in severe lighting changes. The dataset is also constructed using both day videos and night videos, which has not been considered in previous datasets. This provides a better simulation of the real scenarios than the previous datasets but includes severe lighting variation. 3) More reliable bounding box detector. First, the Faster R-CNN detection algorithm is used to perform rough detection of pedestrians, and then, the position of the rectangular frame is manually adjusted.

3.2. Evaluation protocol

We report performance using a standard average-log miss rate (MR) and AP in experiments. According to the given thresh-

old of detection confidence score, the detected bounding boxes (N_p) can be divided into four categories, true positive (TP), false-positive(FP), true negative (TN), and false-negative (FN). TP denotes that the sample is a positive sample in reality and is predicted to be a positive sample. FP denotes that the sample is a negative sample in reality but is predicted to be a positive sample. TN denotes the sample that is a negative sample in reality and is predicted to be a negative sample. FN denotes that the sample is a positive sample in reality but is predicted to be a negative sample. Miss rate (MR) can be computed using the number of true positives (N_{tp}) and the number of ground-truth images (N_{gt}) as:

$$MR = 1 - N_{tp}/N_{gt}, \tag{1}$$

and the false positive per image (FPPI) can be calculated by dividing the number of false positives (N_{fp}) by the number of images (N_{img}), as follows:

$$FPPI = N_{fp}/N_{img}, \tag{2}$$

over the false positive per image (FPPI) range of $[10^{-2}, 10^{1}]$. For the miss rate plotted versus the number of false positive image in a log-log plot, a lower miss rate reflects better detection performance.

Average precision (AP) is the same as the COCO detection Task [69], and the confidence scores of all of the samples are ranked from high to low. Suppose that R recall values are obtained from topN $(1/R, 2/R..., R/R)$. For each recall value v, the maximum precision corresponds to $top - N$ at the recall threshold, and then average R precision is used to obtain the final AP. In particular, recall (R) is calculated using the number of true positives (N_{tp}) and the number of false-negative (N_{fn}) as:

$$R = N_{tp}/\left(N_{tp} + N_{fn}\right), \tag{3}$$

and the precision (P) is calculated using the number of true positives(N_{tp}) and the number of all detected bounding boxes (N_p).

$$P = N_{tp}/N_p, \tag{4}$$

AP is the area of the P-R curve, a higher average precision reflects better detection performance.

4. Super resolution and detection network

For low-quality surveillance images, the pedestrian contour is blurred, and thus, pedestrians cannot be separated from the background. To achieve more accurate pedestrian detection for low-quality images, we propose the super-resolution detection (SRD) network, Specifically, we perform super-resolution reconstruction of the image to help distinguish the pedestrians from the background, and adopt an improved Faster R-CNN network to locate the bounding box of the pedestrians. Fig. 2 illustrates the overall framework of the SRD network. As shown in Fig. 2, our framework is mainly composed of two network structures; specifically, the dashed box on the left represents the super-resolution reconstruction network, and the dashed box in the middle represents the pedestrian detection network. The module presents between the super-resolution network and the detection network is the DU module, and it means that we first perform the down sampling (D) operation, and then the upsampling (U) operation. The right dashed box represents the goal of the improvement method.

4.1. Network architecture

4.1.1. Super-resolution reconstruction

To better leverage the quality of the image, the collected image is reduced to a quarter of the original image as input, and then, an upsampling method is adopted (i.e., bilinear interpolation, super-resolution network) to recover the image, with the original image as the ground-truth. The super-resolution methods include the SRGAN, RAN, and SAN networks. As shown in Fig. 7, the edge of the pedestrian contains more detailed information when the SRGAN method is used; such information is convenient for distinguishing the pedestrian from the background. Therefore, we employ the SRGAN network [53] to obtain clear images.

Specifically, let $I^{LR} \in \mathbb{R}^{C\times H\times W}$ represents the low-quality image as input to the SRGAN network. SRGAN network is composed of a generator $G(\theta)$ and a discriminator $D(\theta)$. Its generator part

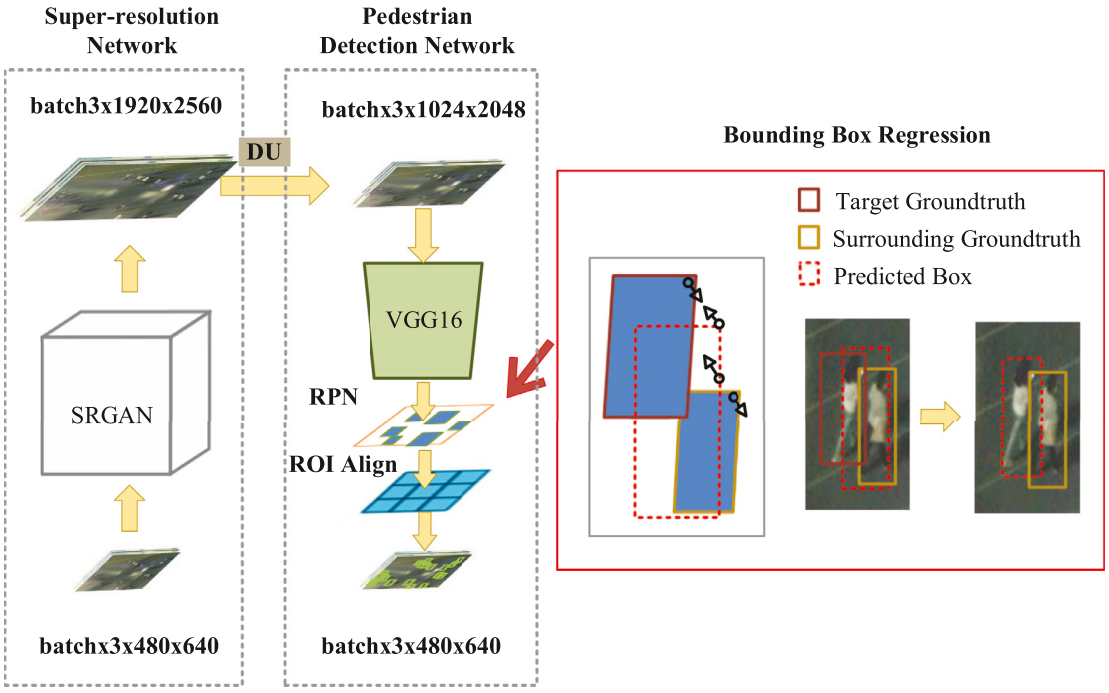


Fig. 2. A schematic overview of the SRD Network. It includes a super-resolution Network (the left part), DU resampled module (i.e. downsampling and upsampling by bilinear interpolation method), and a Detection network (the right part).

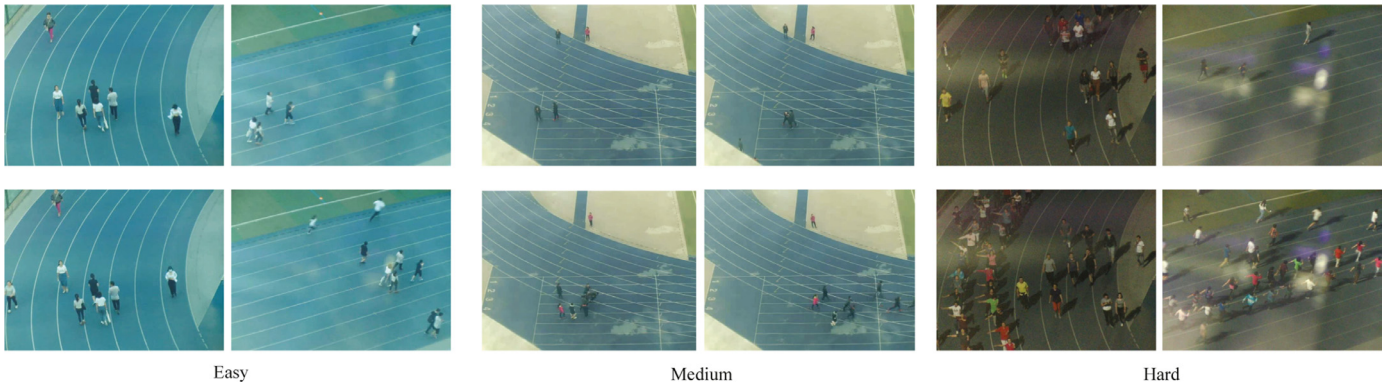
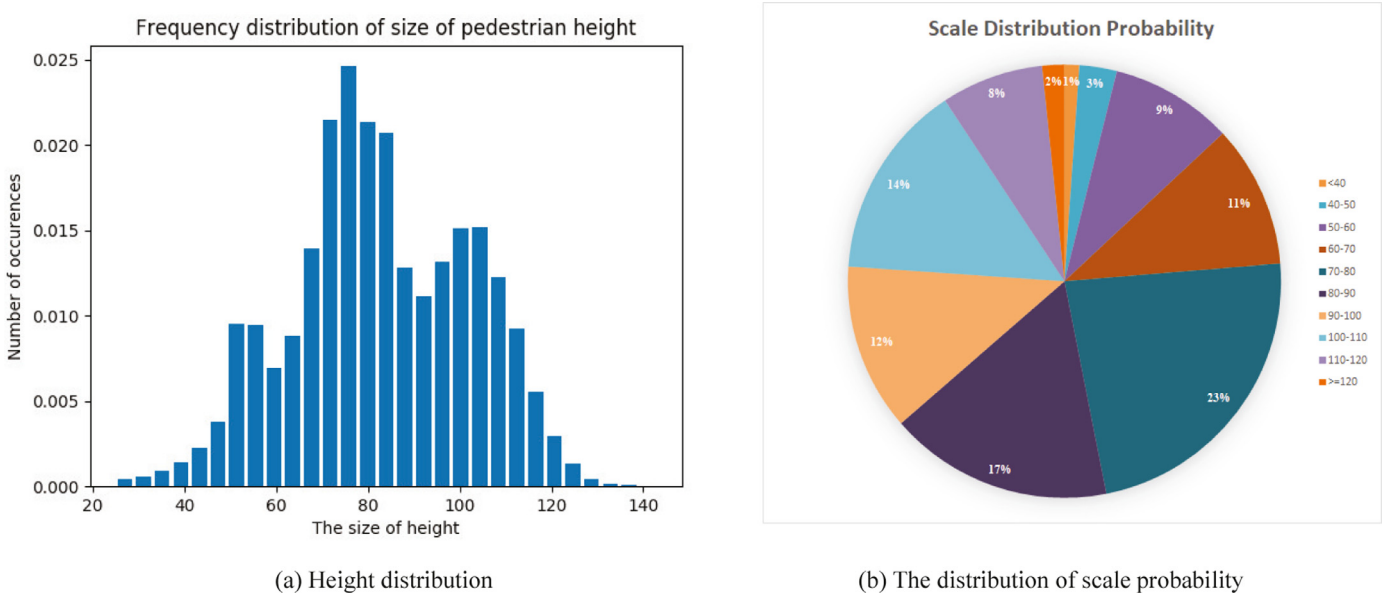
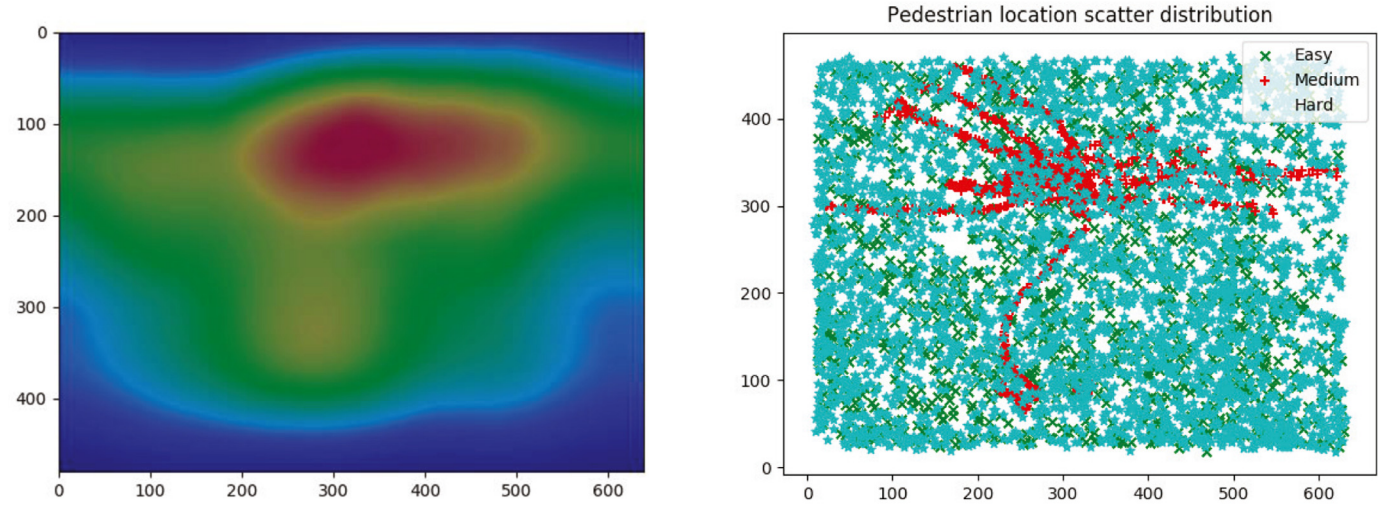


Fig. 3. PG test set. It includes an easy subset, a medium-difficulty (medium) subset, and a hard subset.



(a) Height distribution (b) The distribution of scale probability

Fig. 4. (a) Height distribution, we show histograms of height rate over the number of proposal dataset. (b) Distribution of scale probability; we draw the ratio of quantitative scale range (height in pixels) to the entire dataset.



(a) The distribution of pedestrian position (b) Position distribution of different dataset.

Fig. 5. (a) Distribution of pedestrian positions using a heat map drawn according to the statistics of the coordinates of all of the pedestrian center points. (b) Location distributions of different datasets, including easy, medium-difficulty (medium) and hard test sets.

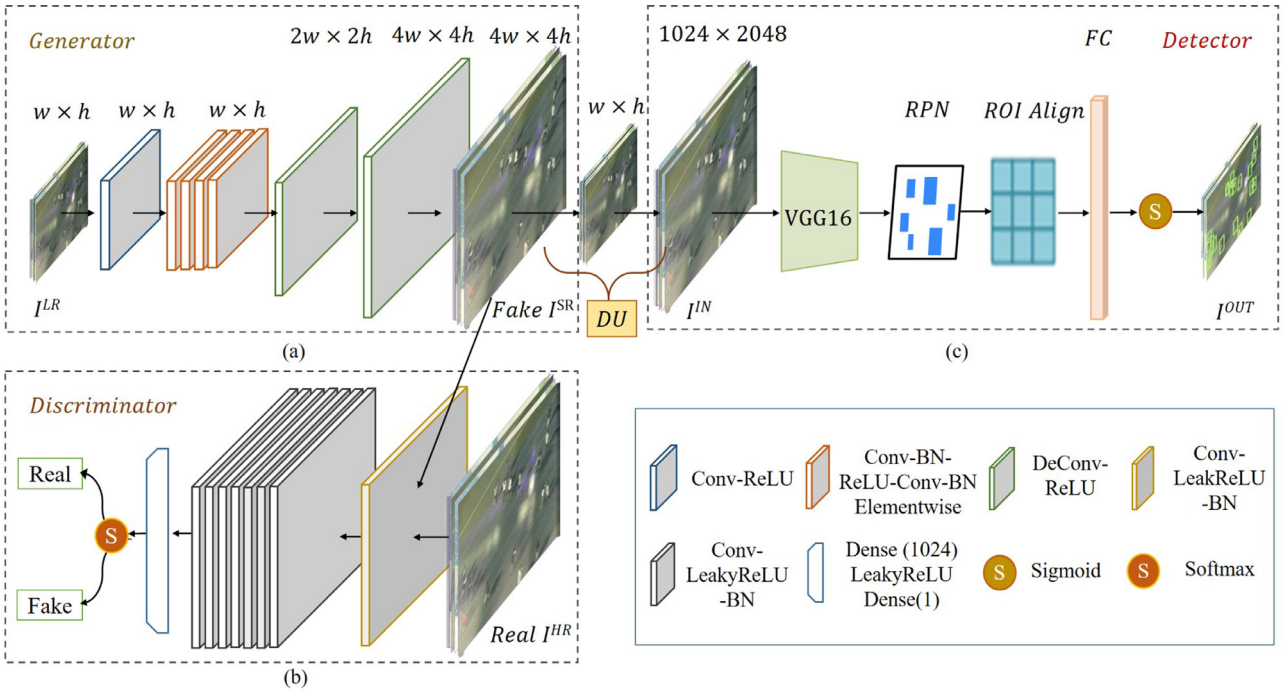


Fig. 6. Detailed framework description of the SRD network where DU module represents resampling by the bilinear interpolation method. (a) SRGAN generator network, (b) SRGAN discriminator network, (c) pedestrian detection network.

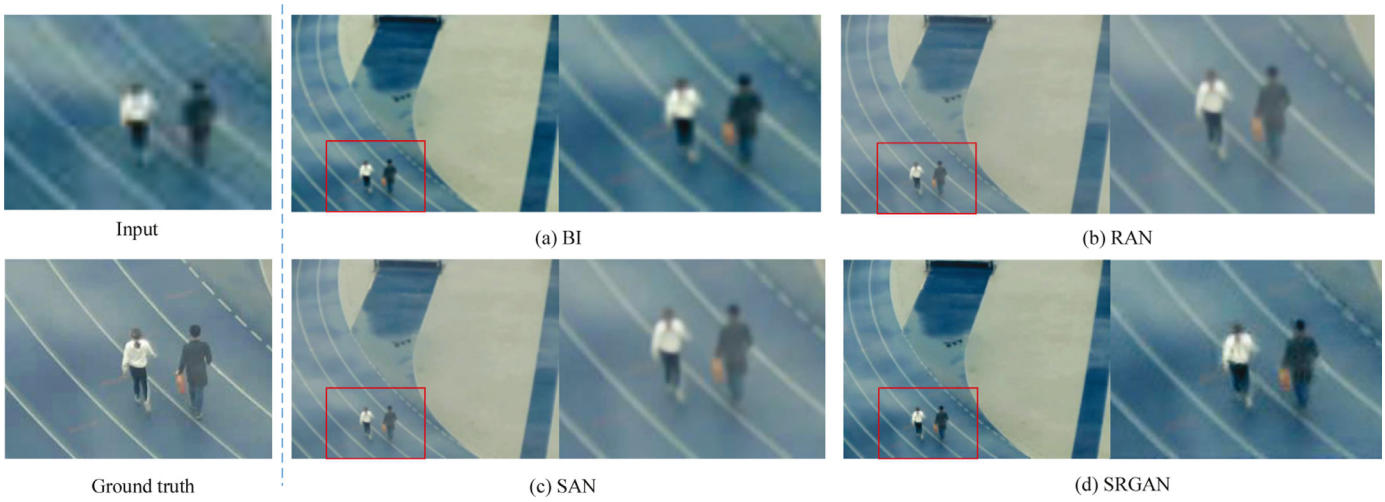


Fig. 7. High-resolution image reconstruction. (a), (b), (c), (d) are the images reconstructed by BI, RGAN, SAN, and SRGAN, respectively.

contains multiple residual blocks and two losses. As shown in Fig. 6(a), these residual blocks consist of a 1-block Conv-ReLU sub-network, 6-block Conv-BN-ReLU-Conv-BN-sum sub-network, and 2-block DeConv-ReLU sub-network. Each residual block contains two 3×3 convolution layers. The convolution layer is followed by the batch normalization layer and PReLU. The loss functions include perceptual content loss and adversarial loss. Super-resolution images can be generated through the generation network. The label map and the generated super-resolution image are input to the discriminator together, as shown in Fig. 6(b). The discriminator consists of 1-block Conv-Leaky ReLU sub-network, 7-block Conv-BN-Leaky ReLU sub-networks, and 1-block Dense-Leaky ReLU-Dense sub-network, followed by a sigmoid function for two classifications. The discriminant loss function is used to determine whether the generated image is a true high-resolution image.

4.1.2. Detection network

In this section, we will describe the pedestrian detection process shown in Figs. 2 and 6(c) in detail. The super-resolution images are input into the Faster R-CNN detection network $De(\theta)$, and the network outputs the pedestrian position regression box. This process consists of three steps: the corresponding feature maps are obtained after the images are fed into the backbone; RPN structure is used to generate candidate bounding boxes that are mapped to the feature map to obtain the corresponding feature matrix; ROI Pooling is employed to scale the feature map to a size of 7×7, and the feature map is flattened to obtain the prediction results through a series of fully connected layers. The Faster R-CNN algorithm has achieved outstanding performance in the general target detection task, but the performance of the Faster R-CNN algorithm for the low-quality image of the pedestrian with occlusion will significantly decrease. In this work, we improved the Faster R-CNN

based on VGG16 to boost the performance of pedestrian detection. Our main improvements of the Faster R-CNN are as follows:

Fixed Image Size. Since the input of a Faster R-CNN network can be determined according to the input image, that is, the maximum image size of the input images can be set as the input size, the smaller size images need to be expanded to the input size by padding with zero. In our dataset, all of our input images are given with an unified size of $3 \times 480 \times 640$. Due to poor image quality, it is difficult to distinguish pedestrians from the background. Scaling of the image to a larger size is beneficial for distinguishing the pedestrians from the background. The size of the image obtained through the super-resolution network will be four times of the original image, that is, $3 \times 1920 \times 2560$. Because the anchor ratio of Faster R-CNN is $\{1:1, 1:2, 2:1\}$, we consider changing the input obtained super-resolution image to a suitable size, i.e., $3 \times 1024 \times 2048$, that will be a better match to the anchor ratio. When the image details are generated by GAN, much noise will be introduced through the GAN method. To avoid the effect of noise, we use the following method to re-sample the obtained super-resolution images (with the size of $3 \times 1920 \times 2560$). First, the bilinear interpolation method is used for the downsampling of the super-resolution image to obtain the small image ($3 \times 480 \times 640$) by fusing the adjacent pixel information. Then, the neighboring information of the small image is resampled to enhance the details to obtain the large image ($3 \times 1024 \times 2048$). Thus, the influence of noise introduced by the GAN will be reduced. Additionally, the pedestrian detection performance can be improved. Then, the images are sent to the VGG16 network to generate features and the features are input into the RPN (region proposal network) layer. RPN consists of a 3×3 Conv-ReLU and can be used to generate regional proposals. Then, instead of ROI pooling, ROI align is employed to map the ROI to the corresponding position of the feature map according to the input image. It uses the proposals to extract the proposal features from the feature maps and send them to the subsequent fully connected layers. Finally, the prediction probability of foreground and background can be obtained by softmax classification.

Finer Bounding Boxes. The core objective of the target detection network is to make the predicted proposal closer to the real target box (ground truth). For the images obtained by the surveillance camera, there exist overlapping images of pedestrians or dense crowding of pedestrians, as shown in Fig. 1, and the predictive bounding boxes will be interfered and away from the ground truth due to the influence of other nearby targets. To prevent the predicted box from moving away from the ground truth, in addition to classification loss and regression loss, we introduce repulsion loss [21] in our detector. By narrowing the predicted box and the corresponding ground truth (GT) distance, repulsion loss can increase the predicted box and other target boxes (i.e. ground truth and predict boxes) to improve the detection performance of pedestrian occluded scenes, as shown in the rightmost part of Fig. 2. More specifically, the loss function consists of three parts, the attraction term, the repulsion ground truth term (RepGT), and the repulsion boxes term (RepBox). Attraction term loss can be generated by computing the distance between the predicted box and the corresponding GT, while RepGT aims to push away the predicted box and other adjacent GTs. RepBox aims to push away the predicted target bounding box and other predicted bounding boxes. Thus, repulsion loss can guide the bounding box to be more accurate in bounding box regression.

4.2. Diversity loss functions

4.2.1. Super-resolution loss function

Super-resolution loss functions consist of a content loss and an adversarial loss. The adversarial loss (generative loss) is defined based on the probabilities of the discriminator overall training samples as:

$$L_G = \sum_{n=1}^{N} -logD\left(\hat{y}_i\right) \tag{5}$$

Here, $D\left(\hat{y}_i\right)$ is the estimated probability that the reconstructed image $\hat{y}_i$ is an HR image.

Content loss is defined as the Euclidean distance between the feature representations of a reconstructed image $\hat{y}_i$ and the reference image y_i, $\phi_{i,j}(\cdot)$ represents the jth convolution operation before the ith max pooling layer:

$$L_{SR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{w=1}^{W_{i,j}} \sum_{h=1}^{H_{i,j}} \left(\phi_{i,j}(y)_{w,h} - \phi_{i,j}\left(\hat{y}\right)_{w,h}\right)^2 \tag{6}$$

In addition to the generator, the SRGAN contains a discriminator, and the discriminator loss is: Let $\hat{y}_i$ represent the generated image and y_i represent the target image. Y represents the real image distribution and $\hat{Y}$ represents the generated image distribution.

$$L_D = \max_D\left(E_{\hat{y}_i \in \hat{Y}}\left[\log D\left(\hat{y}_i\right)\right] + E_{y_i \in Y}[\log\left(1 - D(y_i)\right)]\right). \tag{7}$$

4.2.2. Detection loss

The detection network includes classification loss L_{cls} and regression loss L_{reg}. $L_{cls}$is a log loss over two classes (object vs. not object), and the binary cross entropy loss is shown in the following equation:

$$L_{cls} = \frac{1}{N} \sum_{i=1}^{N} -\left[p_i * log\hat{p}_i + (1 - p_i)log\left(1 - \hat{p}_i\right)\right] \tag{8}$$

Here, $\hat{p}_i$ is the probability that the prediction anchor is the target. Let $p_i = 1, 0$ be an indicator for matching the default box to the ground truth box of the positive label (negative label). For example, $\hat{p}_i$ is a positive sample, then p_i is equal to 1, the first term is retained to calculate the loss.

The regression loss is a smooth L1 loss between the predicted box $\hat{t}_i = \left\{\hat{t}_x, \hat{t}_y, \hat{t}_w, \hat{t}_h\right\}$ and the ground truth box $t_i = \{t_x, t_y, t_w, t_h\}$ parameters.

$$L_{reg} = \frac{1}{N} \sum_{i=1}^{N} \hat{p}_i \; smooth_{L1}\left(t_i - \hat{t}_i\right) \tag{9}$$

The smooth L1 loss can be written as:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if \;\; |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{10}$$

Compared with the L2 loss function, smooth L1 loss is insensitive to outliers, and has relatively small gradient change and good training convergence effect.

4.2.3. Repulsion loss

Repulsion loss can effectively improve the accuracy of the target location and is insensitive to the threshold of NMS. According to reference [21], repulsion loss can be written as :

$$L_{rep} = L_{Attr} + \alpha L_{RepGT} + \beta L_{RepBox}, \tag{11}$$

where L_{Attr} is the attraction term that requires a predicted box to approach its designated target, while L_{RepGT} and L_{RepBox} are the repulsion terms that require a predicted box to keep away from other surrounding ground-truth objects and other predicted boxes with different designated targets, respectively. Coefficients α and β act as the weights to balance auxiliary losses, we will discuss the influence of the values of α and β in Section 5.4.2.

5. Experiments

In this section, we will introduce our implementation details of data processing and training settings. All of the experiments are conducted based on the proposed PG dataset. The overall process is shown in Fig. 2: first, the SRGAN network is pre-trained. Then, the image that needs to be detected is input into the SRGAN network, and the output super-resolution image has the size that is four times that of the input image. The obtained super-resolution image is resampled by the DU module (i.e., downsampling and upsampling) to obtain the $3 \times 1024 \times 2048$ image that is then input into the improved Faster R-CNN network to complete the detection.

5.1. Training implementation details

We perform the experiment with a 1060 GPU and TensorFlow 1.4, tensorlayer 1.8.0, and PyTorch 0.4.0. We first trained the super-resolution network with 8 images in each mini-batch. In the pre-training stage of the super-resolution network, VGG19 is initialized by ImageNet-pre-trained weights, and the network is optimized by the random gradient descent method. In the detection network, all of the input images are preprocessed by resizing them to $3 \times 1024 \times 2048$. In the pre-training Faster R-CNN network, VGG16 is initialized by ImageNet-pre-trained weights. The initial learning rate was set to 0.001 and was decreased by 0.1 every 3000 epochs. Anchors with different scales $[3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33]$ are used in RPN. The features of each proposal are extracted by ROIAlign. The pooling size is set to 7×7.

5.2. Data processing

To make the super-resolution reconstructed image more realistic, we pre-process the images to train the SRGAN network. Specifically, the collected images are reduced to a quarter of the original image as the input of SRGAN, and the original image as ground-truth. They are input to the SRGAN network to pre-train the SRGAN network. The size of the obtained output image by the SRGAN network will be four times the input image. In the detection network, we perform a resampling operation to obtain $3 \times 1024 \times 2048$ images, and the horizontal image flipping is the only form of data augmentation.

5.3. Data analysis

We first introduce a dataset named playground (PG) dataset for low-quality image detection. The dataset is shot by two different cameras, including day and night surveillance video pedestrian images. There are a total of 5,752 images with 31,041 annotations, i.e., 5.39 annotations per image, and the average size of the pedestrian is 41×87. This means that this dataset contains small-size pedestrians with various kinds of occlusions. The experiments are performed on three subsets, as shown in Fig. 3, namely the easy subset, the medium-difficulty subset, and the hard subset. The images in the easy subset are taken at different times on the same day as the training set. In particular, the easy subset is shot on the playground during the day. The medium-difficulty dataset is also shot during the day but in a different season. This subset is a fighting video where the size of the person is smaller than the training set. The hard dataset is shot in the summer similar to the easy dataset, but not on the same day, and is also shot at night. It includes many pedestrians, and light interferences are present.

5.4. Model analysis

In this section, we perform some ablation experiments on the medium-difficulty dataset in the test set while analyzing our improved method. We used the same settings for all experiments, but changes were made to some components as specified below.

5.4.1. DU module

Pedestrian detection of low-quality images are mainly divided into two parts, namely, super-resolution processing and pedestrian detection. In terms of the super-resolution method, in addition to the popular GAN method (RCAN and SRGAN), some traditional methods such as the bilinear interpolation (BI) method are known. We conduct experiments and compare the results obtained by different methods. In addition, we find that performing super-resolution processing on the image alone does not lead to a clear improvement in detection accuracy. The obtained super-resolution image is resampled by a DU module (i.e., downsampling and upsampling) to obtain the $3 \times 1024 \times 2048$ image, and we find that it displays greatly improved detection performance, as shown in Table 2. This is because the addition of bilinear interpolation downsampling effectively enriches some image details.

Specifically, bilinear interpolation (BI), SRGAN, and SRGAN+DU are used to carry out upsampling images, respectively. In this paper, the input original image is $3 \times 480 \times 640$. The size of the image obtained by SRGAN is 4 times that of the original image, namely $3 \times 1920 \times 2560$. The size of the images obtained by BI and SRGAN+BI is $3 \times 1024 \times 2048$. We compare the detection AP of the different up-sampled images and the original images by the improved Faster R-CNN with the results presented in Table 2. The original image is input into the improved Faster R-CNN, and the detected AP obtained is 93.5%. After the BI-processed images were detected, the AP was improved by 0.8%. The AP detection results of super-resolution images obtained by SRGAN are higher than the original images and lower than the BI method. This is because the super-resolution images generated by the SRGAN network will produce unpredictable noise that will affect the detection performance. However, the image processed by the SRGAN+DU method can effectively remove some noise and obtains the highest AP values, which are 1.8%, 1.0% and 1.4% higher than the detection results of the original image, BI and SRGAN, respectively.

5.4.2. Parameter analysis

In repulsion loss, L_{Attr} is the attraction term that is important to the closer agreement between predicted box and the corresponding GTs. L_{RepGT} and L_{RepBox} are the repulsion terms, α and β act as the weights for balancing auxiliary losses. We investigated the sensitivity of the parameters α and β in Eq. (11) on our medium-difficulty test set. In particular, we set the values of α as 0.1, 0.3, 0.5, 0.7 and 0.9, respectively. Correspondingly, the values of β are set as 0.9, 0.7, 0.5, 0.3 and 0.1, respectively. An examination of the data in the Table 3 shows that for $\alpha = 0.1$ and $\beta = 0.9$, *MR* is the lowest, and AP reaches 95.3%; for $\alpha = 0.5$ and $\beta = 0.5$, *MR* is low and AP reaches the highest value of 95.8%. The *MR* of $\alpha = 0.1$ is 2.4% lower than the *MR* of $\alpha = 0.5$, and the AP of $\alpha = 0.5$ is 0.5% higher than the AP of $\alpha = 0.1$. In addition, we trained and tested the model for $\alpha = 0$, and $\beta = 1.0$, and obtained 70.8% *MR* and 92.0% AP. Based on the difference between the two evaluation criteria, we finally selected $\alpha = 0.1$ and $\beta = 0.9$ for the training of our dataset.

5.5. Super-resolution results

In this section, the collected image is reduced to a quarter of the original image as input, and then the super-resolution methods and bilinear interpolation upsampling method are used to recover the image, with the original image as ground-truth. The super-resolution methods include SRGAN, RAN, and SAN networks. As shown in Fig. 7(a), the image restored by the bilinear interpolation method is quite blurry. Fig. 7(b), (c) processed by RAN,

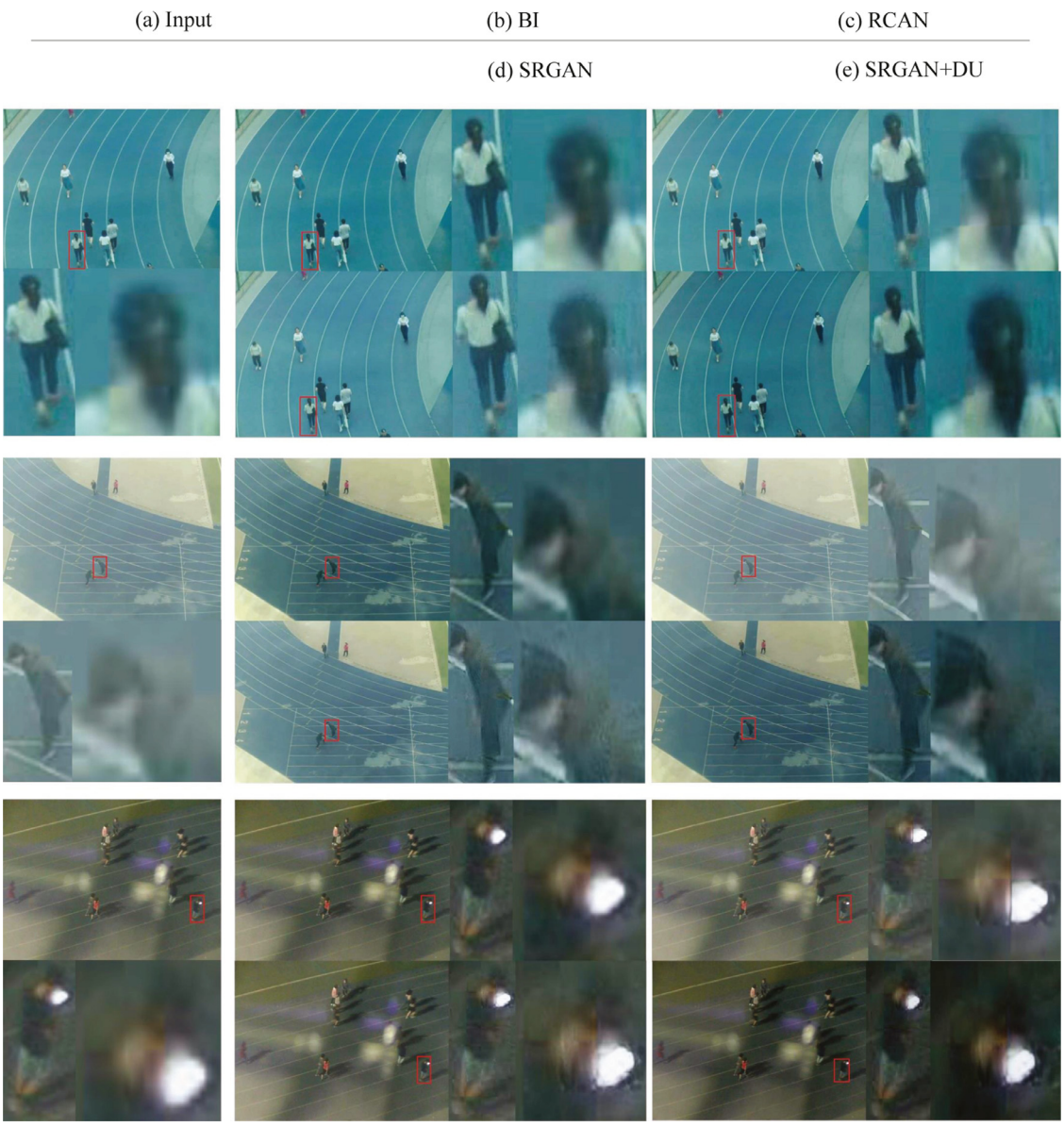


Fig. 8. Comparison the SR results of Bilinear interpolation (BI), RCAN, SRGAN and SRGAN+DU. The left part represents inputs.

SAN for super-resolution processing, respectively. Although they are clearer than Fig. 7(a), the pedestrians' contour edges are too smooth. Fig. 7(d) is the image processed by SRGAN for super-resolution processing. It is observed that the edge of the pedestrian contains more detailed information than other methods, which is convenient for distinguishing the pedestrian from the background. The above comparison experiment is based on images with ground truth. To illustrate the effectiveness of the SRGAN network, we performed super-resolution reconstruction on images without ground truth, with the results shown in Fig. 8. Comparison of the images reconstructed by BI, RCAN, SRGAN, and SRGAN+DU, it was found that it is easier to distinguish pedestrians from the background in the images obtained by the SRGAN+DU method.

5.6. Detection results

In this section, we first compare the detection results of the original image, the BI upsampled image, and the SRGAN super-resolution image based on the Faster R-CNN. The detection results of our three test subsets are shown in Fig. 9. Here, Fig. 9(a)–(c) are for the easy subset, medium-difficulty subset, and hard subset, respectively. The top line represents the detection results of the original images. The second line presents the detection results of images that are obtained by the BI method. The third row presents the detection result images obtained by the RCAN method. The bottom row presents the detection results of images that are obtained by the SRGAN method. It is observed from Fig. 9(a) that errors are present in both the first and second lines. Some pedestrians are detected repeatedly, indicating that some backgrounds are also considered as pedestrians. For Fig. 9(b), misdetection is found, both in the first, second, and third lines, and the bottom line pedestrians are completely detected by our method. For Fig. 9(c), for the subset with lower quality, the first row has

Table 2
Detection AP (%) on medium-difficulty subset based on improved Faster R-CNN networks.

Method	Original images	BI	SRGAN	SRGAN+DU
Improved Faster R-CNN	93.5	94.3	93.9	**95.3**

Table 3
The *MR* (%) and AP (%) of RepGT and RepBox Losses with different weights α and β on medium-difficulty test set.

α	0.1	0.3	0.5	0.7	0.9
β	0.9	0.7	0.5	0.3	0.1
MR	**66.6**	71.1	69.0	78.8	76.1
AP	95.3	94.0	**95.8**	92.1	91.0

redundant detection of pedestrians results in the original image, while the second and third rows can detect clearer pedestrian contours. In addition, the pedestrians in the bottom right corner of the image Fig. 9(c) are misdetected because the pedestrian is severely occluded. To solve the problem of pedestrian occlusion, after performing super-resolution processing, we propose an improved Faster R-CNN network by introducing repulsion loss and expanding the input size of the image. The pedestrian detection result of the improved Faster R-CNN is shown in Fig. 10. The red box represents the ground truth, and the green box represents the predicted bounding box. An examination of Fig. 10 shows that the number and positions of the regression boxes obtained by improved Faster R-CNN are the same as the ground truth, and some of our prediction box (green box) are more suitable for pedestrian size than the ground truth. In addition, we also find that for some pedestrians that are ignored by manual annotation (occluded pedestrian height $<$30 pixel, or occlusion area $>$60%), the regression box can also be obtained through super-resolution processing and the improved faster R-CNN algorithm. The detection results are shown in Fig. 11, where the red box represents ground truth, and the green box represents the predicted bounding box. As shown in Fig. 11(a), for the pedestrian, whose key parts with occlusion (such as the head), the rest part is not easy to identify as a pedestrian (should be ignored), the proposed algorithm draws an accurate bounding box. Fig. 11(b) shows light interference and it is difficult for humans to distinguish pedestrians, while the proposed algorithm also gives accurate location.

For an intuitive comparison of the effectiveness of our proposed method to those of the other methods, we perform the experiment for the log-average miss rate (*MR*) on SSD, YOLOv3, YOLOv4, Faster R-CNN, improved Faster R-CNN, and improved Faster R-CNN+SRGAN. From Dollar et al. [63], the detected bounding box and ground truth will form an overlap area, and decreasing the threshold of the overlap area below 0.5 has little effect on the reported performance. However, increasing it to over 0.6 results in rapidly increasing log-average miss rates because improved localization accuracy is necessary; therefore, we compute *MR* on the different test sets by setting the threshold of the overlap area to 0.5. As shown in Table 4, the obtained log average miss rate (*MR*) of each test set is quite different due to the different data of complex scenarios. As shown in Fig. 12, the range of false positives per image of our proposed dataset is $[10^{-3}, 10^{0}]$. This means that our dataset contains more challenges. Compared with the six methods, we can see that our proposed method achieves promising *MR* performances. For each dataset, the *MR* obtained by the SSD method is the highest, which is due to the under-fitting phenomenon and

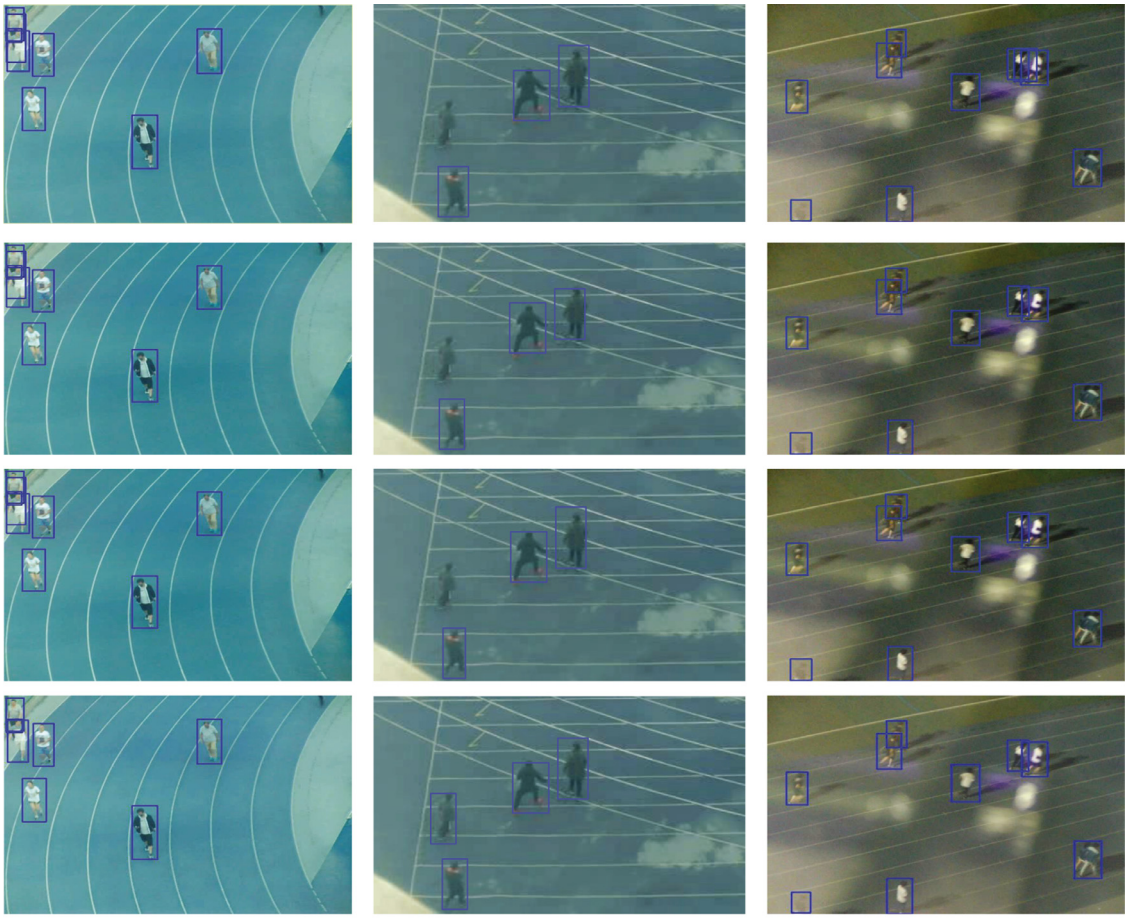


Fig. 9. Detection result images with the Faster R-CNN. The top row is the result of the original image. The second row presents the detection result images obtained after bilinear interpolation (BI). The third row presents the detection result images obtained by the RCAN method. The bottom row presents the result images obtained by SRGAN.

Fig. 10. Detection examples on proposed dataset using improved Faster R-CNN with VGG-16 (The red box represents ground truth, and green box represents the predicted bounding box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

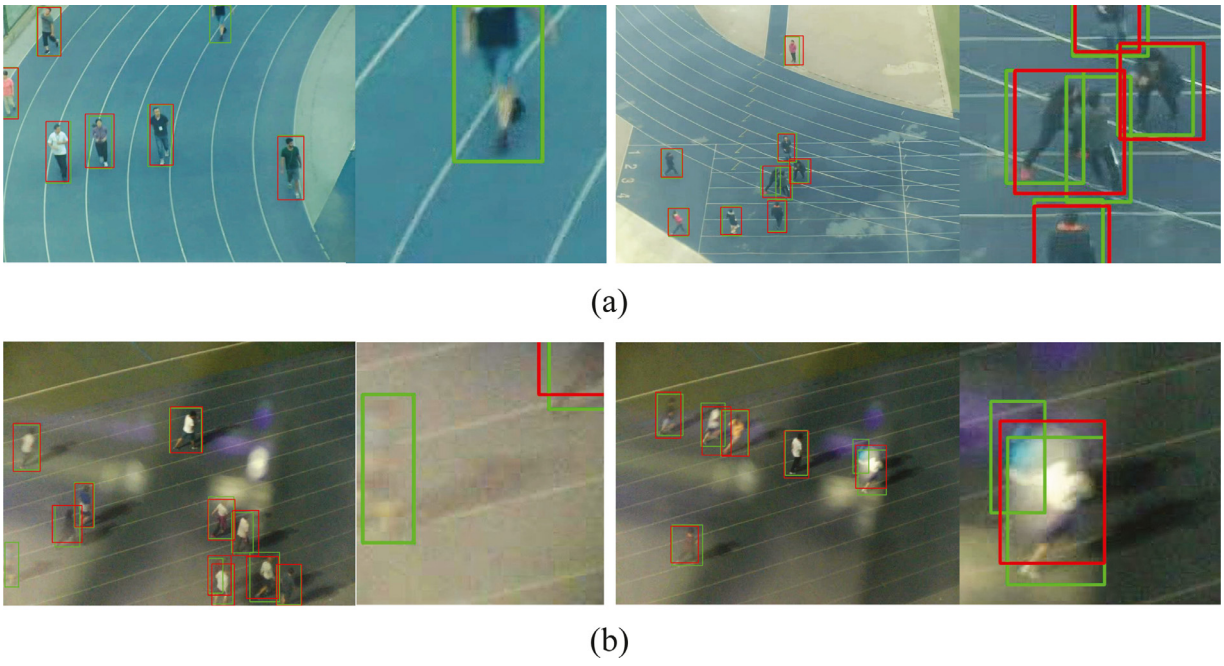


Fig. 11. Comparing the prediction box with ground truth (The red box represents ground truth, and green box represents the predicted bounding box): (a) the occlusion pedestrian and height pixels <30 (should be ignored), the proposed algorithm draw an accurate bounding box. (b) Light interference, it is difficult for humans to distinguish pedestrians, and the proposed algorithm also give accurate locations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Detection MR^{-2} (%) results on different HR images.

dataset	SSD	YOLOv3	YOLOv4	Faster R-CNN	Improved Faster R-CNN	Improved Faster R-CNN+SRGAN
Easy	95.16	75.33	40.02	43.22	38.25	**19.03**
Medium	97.90	88.21	72.56	75.55	71.13	**66.63**
Hard	96.57	92.55	74.73	76.20	69.48	**57.99**

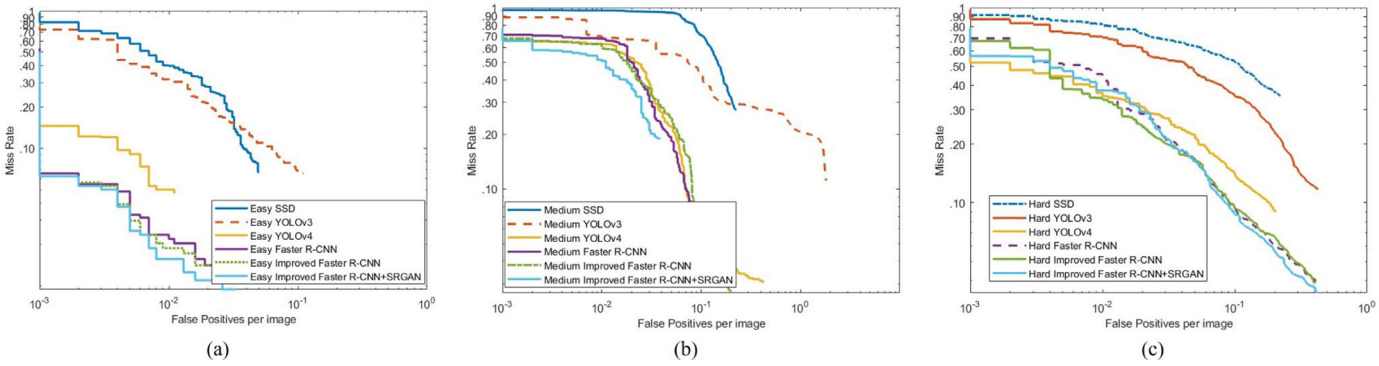


Fig. 12. Comparison the *MR* results of SSD, YOLO v3, YOLOv4, Faster R-CNN, improved Faster R-CNN, and improved Faster R-CNN+SRGAN on different test sets.

Table 5
Miss rate on the Caltech pedestrian validation set by different training methods. A→B represents pre-training on A and finetuning on B.

Methods		training	MR
DeepParts [44]	ICCV2015	Caltech	64.8
Hyper Learner [45]	CVPR2017	Caltech	61.5
FRCNN+ATT [46]	CVPR2018	Caltech	54.5
SAF RCNN [43]	TMM2018	Caltech	62.6
RepLoss [21]	CVPR2018	Caltech	59.0
AR-Ped [70]	CVPR2019	Caltech	58.9
CSP [42]	CVPR2019	Caltech	57.7
SRD		Caltech	**52.3**
SRD		PG→Caltech	**51.3**

there exist many missing detected boxes. Although the *MR* of the YOLOv3 method is lower than SSD, it is still high. This is because the model trained by the YOLOv3 method has an overfitting phenomenon, and there exist redundant detected boxes. Yolov4 adopts CIOU loss in the regression method. It considers the scale information of the aspect ratio of the boundary box in order to improve the detection accuracy of the occlusion targets compared with YOLOv3, SSD, and Faster R-CNN. The model trained by Faster R-CNN can effectively reduce the detected *MR*, but it is still higher than the detection result of YOLOv4. Compared with Faster R-CNN, the *MR* on the easy, medium-difficulty, and hard data sets obtained by improved Faster R-CNN dropped by 4.97%, 4.42%, and 6.72%, respectively, The image reconstructed by super-resolution is input into the improved Faster R-CNN, and the result is improved more significantly. The easy data set achieved 19.03 % *MR*.

5.7. Generalization capability

In this subsection, we compare our methods with some of the latest pedestrian detection methods in the Caltech pedestrian data set, as shown in Table 5. In this work, considering that the method we proposed can effectively alleviate the problems of low-quality pedestrian and pedestrian occlusion, instead of the test results of most of the algorithms on the R (reasonable) data, we tested the data of the whole validation set of Caltech. The valid data contains the pedestrians over 20 pixels in height with less than 80% occlusion. Our SRD model pre-trained on PG can reduce the MR from 52.3% to 51.3%, its MR is 13.5% lower than the MR of DeepParts. Although our data set is quite different from the existing public data sets scenario, our proposed data set can also serve as an effective pre-training data set for pedestrian detection tasks.

5.8. Discussion

In addition, we examine the test sets with all three levels of difficulty and calculated the AP of their test results, as shown in Tables 6. For the three subsets, the AP values of all of the methods are very high. Specifically, the AP of the easy subset obtained using our method (SRGAN+improved Fater R-CNN) reaches 99.10 %, which is 0.48 % greater than the baseline (only Faster R-CNN) detection result. For the medium-difficulty dataset that is a dataset of pedestrian images collected in the winter with different people's appearance and worse resolution, the input image is directly detected and AP of 91.03% is obtained. Our method obtains the AP of 95.25%. The hard subset is shot at the same time as the training set, but has different people. Compared with other test subsets, it has more pedestrians. Our method achieves AP of 96.98%.

Table 6
Improved Faster R-CNN detection AP (%) results on different HR images.

dataset	Baseline	Our method
Easy	98.62	**99.10**
Medium	91.03	**95.25**
Hard	95.80	**96.98**

The comparison of the results presented in Tables 4 and 6 shows that although we obtain an AP of 96.98% on the hard subset, the obtained MR is 57.99%. A higher AP value corresponds to the higher accuracy of the pedestrian detection. The MR value represents the ratio of the negative samples detected as pedestrians and ground-truth. In many application fields for pedestrian detection, a lower MR corresponds to a lower false detection rate. To clarify this point, we draw some failure cases and find that our algorithm not only obtains regression boxes that are close to the ground truth, but also obtains the regression box that misdetected the background as a pedestrian. As shown in Fig. 13, the image in the regression box is a negative sample but is detected as a positive sample by the detection network (FP). Therefore, there are three reasons for the high MR: (1) as shown in Fig. 13(a), misdetection of the background as a pedestrian; (2) as observed from Fig. 13(b), due to light interference, pedestrian features are disturbed, and three pedestrians are detected with four regression boxes. (3) As observed from Fig. 11, some pedestrians are ignored by manual annotation, but are detected by the proposed algorithm (this situation is relatively rare), this is also regarded as FP. Although the MR obtained by our data set is high, it also shows that it is difficult to distinguish between the background and pedestrian for the images in the proposed PG dataset that is more challenging than the other pedestrian detection datasets.

5.9. Computational efficiency

We run the tests on Inter(R) Core(TM) I7-5960XCPU, 8G memory strip, and NVIDIA GeForce GTX 1060 with 6GB memory. For the image with a width of 160 and a height of 120, we carry out four-fold pixel reconstruction to obtain an image of $3 \times 640 \times 480$, and

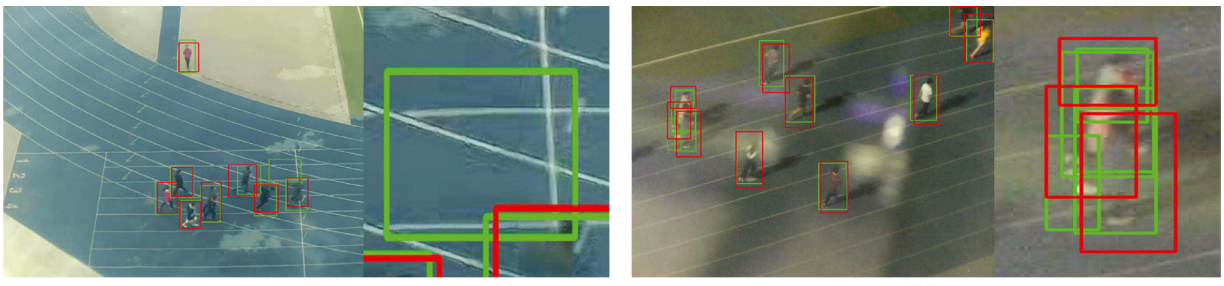


Fig. 13. Failure cases, the image in the regression box is a negative sample but is detected as a positive sample by detection network (FP). (a) Misdetecting the background as a pedestrian, (b) due to light interference, pedestrian features are disturbed, and three pedestrians are detected with four regression boxes.

the frame rate is 4.13 FPS. For the image with a width of 640 and height of 480, we carry out four-fold pixel reconstruction to obtain the image of $3 \times 1920 \times 2560$ with a frame rate of 0.23 FPS. The pedestrian detection speed of our improved Faster R-CNN is 7.08 FPS. The overall end-to-end framework that combines the super-resolution network and the detection network can achieve 2.32 FPS.

6. Conclusions

The existing pedestrian detection datasets were either obtained using vehicle-mounted cameras or were obtained online and contain clear images. To reduce the gap between the current pedestrian detection applications in the field of video surveillance, we develop a low-quality and diverse pedestrian detection dataset called PG. It includes many challenges, including low-quality pedestrians, motion blur, light interference, and pedestrian occlusion. The dataset consists of 5,752 images, and 31,041 bounding boxes. It enables new experiments both for training better models and as a new benchmark. In addition, we combined SRGAN and improved Faster R-CNN network to detect pedestrians in low-quality images. The experimental results prove the effectiveness of the proposed method. A new strategy is proposed to consider the problem of pedestrian detection under surveillance cameras. Our future work will continue this research direction, continuously improving and supplementing our dataset. Additionally, we will explore simpler and more effective methods to solve the challenges of low-quality pedestrian detection.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Key R&D Program of China 2019YFB2204200, in part by the National Natural Science Foundation of China under Grant 61872034, 61972030, 62011530042 and 62062021, in part by the Beijing Municipal Natural Science Foundation under Grant 4202055, in part by the Natural Science Foundation of Guizhou Province under Grant [2019]1064, in part by the Science and Technology Program of Guangzhou under grant 201804010271, in part by RFBR and NSFC according to the research project 20-57-53012 and by Project under Grant No. FSFS-2020-0031.

References

[1] M.C. Liem, D.M. Gavrila, Joint multi-person detection and tracking from overlapping cameras, Comput. Vis. Image Underst. 128 (2014) 36–50.
[2] X. Cao, S. Guo, J. Lin, W. Zhang, M. Liao, Online tracking of ants based on deep association metrics: method, dataset and evaluation, Pattern Recognit. (2020) 107233.
[3] Y. Zhang, Y. Jin, J. Chen, S. Kan, Y. Cen, Q. Cao, PGAN: part-based nondirect coupling embedded GAN for person reidentification, IEEE MultiMedia 27 (3) (2020) 23–33.
[4] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, N. Sang, Re-id driven localization refinement for person search, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9814–9823.
[5] C.M. Martinez, M. Heucke, F.-Y. Wang, B. Gao, D. Cao, Driving style recognition for intelligent vehicle control and advanced driver assistance: a survey, IEEE Trans. Intell. Transp.Syst. 19 (3) (2017) 666–676.
[6] L. Li, D. Wen, N. Zheng, L. Shen, Cognitive cars: a new frontier for ADAS research, IEEE Trans. Intell. Transp.Syst. 13 (1) (2012) 395–407.
[7] G. Matthews, P.A. Hancock, J. Lin, A.R. Panganiban, L.E. Reinerman-Jones, J.L. Szalma, R.W. Wohleber, Evolution and revolution: personality research for the coming world of robots, artificial intelligence, and autonomous systems, Pers. Individ. Differ. (2020) 109969.
[8] K. Goldberg, Robots and the return to collaborative intelligence, Nat. Mach. Intell. 1 (1) (2019) 2–4.
[9] A. Li, Z. Miao, Y. Cen, X.-P. Zhang, L. Zhang, S. Chen, Abnormal event detection in surveillance videos based on low-rank and compact coefficient dictionary learning, Pattern Recognit. 108 (2020) 107355.
[10] Z. Zheng, G. An, D. Wu, Q. Ruan, Spatial-temporal pyramid based convolutional neural network for action recognition, Neurocomputing 358 (2019) 446–455.
[11] W. Liu, X. Du, Q. Geng, J. Li, H. Li, L. Liu, Metro passenger flow statistics based on YOLOv3, in: IOP Conference Series: Materials Science and Engineering, vol. 688, IOP Publishing, 2019, p. 044025.
[12] A. Boukerche, Y. Tao, P. Sun, Artificial intelligence-based vehicular traffic flow prediction methods for supporting intelligent transportation systems, Comput. Netw. 182 (2020) 107484.
[13] J. Yan, X. Zhang, Z. Lei, S. Liao, S.Z. Li, Robust multi-resolution pedestrian detection in traffic scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3033–3040.
[14] W. Nam, P. Dollár, J.H. Han, Local decorrelation for improved pedestrian detection, in: Advances in Neural Information Processing Systems, 2014, pp. 424–432.
[15] T. Chen, S. Lu, J. Fan, S-CNN: subcategory-aware convolutional networks for object detection, IEEE Trans. Pattern Anal. Mach.Intell. 40 (10) (2017) 2522–2528.
[16] D.T. Nguyen, W. Li, P.O. Ogunbona, Human detection from images and videos: a survey, Pattern Recognit. 51 (2016) 148–175.
[17] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast R-CNN for pedestrian detection, IEEE Trans. Multimed. 20 (4) (2017) 985–996.
[18] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[19] S. Ren, R. Girshick, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach.Intell. 39 (6) (2017) 1137–1149.
[20] S. Zhang, R. Benenson, B. Schiele, CityPersons: a diverse dataset for pedestrian detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3213–3221.
[21] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, C. Shen, Repulsion loss: Detecting pedestrians in a crowd, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7774–7783.
[22] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
[23] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (5) (2014) 2019–2032.
[24] J. Yu, D. Tao, M. Wang, Y. Rui, Learning to rank using user clicks and visual features for image retrieval, IEEE Trans. Cybern. 45 (4) (2015) 767–779.
[25] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multi-view features for image re-ranking, IEEE Trans. Multimed. 16 (1) (2014) 159–168.
[26] S.-K. Pavani, D. Delgado, A.F. Frangi, Haar-like features with optimally weighted rectangles for rapid object detection, Pattern Recognit. 43 (1) (2010) 160–172.
[27] F. Bartoli, G. Lisanti, S. Karaman, A.D. Bimbo, Scene-dependent proposals for efficient person detection, Pattern Recognit. 87 (2019) 170–178.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014).
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[30] M. Martineau, D. Conte, R. Raveaux, I. Arnault, D. Munier, G. Venturini, A survey on image-based insect classification, Pattern Recognit. 65 (2017) 273–284.
[31] S. Khan, N. Islam, Z. Jan, I.U. Din, J.J.C. Rodrigues, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, Pattern Recognit. Lett. 125 (2019) 1–6.
[32] S. Kan, Y. Cen, Z. He, Z. Zhang, L. Zhang, Y. Wang, Supervised deep feature embedding with handcrafted feature, IEEE Trans. Image Process. 28 (12) (2019) 5809–5823.
[33] S. Kan, L. Zhang, Z. He, Y. Cen, S. Chen, J. Zhou, Metric learning-based kernel transformer with triplets and label constraints for feature fusion, Pattern Recognit. 99 (2020).
[34] X. Wang, X. Jiang, J. Ren, Blood vessel segmentation from fundus image by a cascade classification framework, Pattern Recognit. 88 (2019) 331–341.
[35] J. Peng, G. Estrada, M. Pedersoli, C. Desrosiers, Deep co-training for semi-supervised image segmentation, Pattern Recognit. (2020) 107269.
[36] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, pp. 580–587.
[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: European Conference on Computer Vision, 2016.
[38] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[39] M. Rastegari, V. Ordonez, J. Redmon, A. Farhadi, XNOR-Net: ImageNet classification using binary convolutional neural networks, in: European Conference on Computer Vision, 2016.
[40] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, (2018), arXiv:1804.02767.
[41] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, YOLOv4: optimal speed and accuracy of object detection, (2020), arXiv:2004.10934.
[42] W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: a new perspective for pedestrian detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[43] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast R-CNN for pedestrian detection, IEEE Trans. Multimed. 20 (4) (2018) 985–996.
[44] Y. Tian, L. Ping, X. Wang, X. Tang, Deep learning strong parts for pedestrian detection, in: IEEE International Conference on Computer Vision, 2016.
[45] J. Mao, T. Xiao, Y. Jiang, Z. Cao, What can help pedestrian detection? Computer Vision and Pattern Recognition, 2017.
[46] S. Zhang, J. Yang, B. Schiele, Occluded pedestrian detection through guided attention in CNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[47] X. Zhang, L. Cheng, B. Li, H. Hu, Too far to see? Not really! - pedestrian detection with scale-aware localization policy, IEEE Trans. Image Process. 27 (8) (2018) 3703–3715.
[48] R. Chen, H. Ai, C. Shang, L. Chen, Z. Zhuang, Learning lightweight pedestrian detector with hierarchical knowledge distillation, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1645–1649.
[49] Y. Pang, J. Xie, M.H. Khan, R.M. Anwer, F.S. Khan, L. Shao, Mask-guided attention network for occluded pedestrian detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4967–4975.
[50] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV, 2014, pp. 184–199.
[51] C. Dong, C.C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 391–407.
[52] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[53] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[54] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.
[55] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017.
[56] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[57] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, W. Wu, Feedback network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3867–3876.
[58] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, in: Lecture Notes in Computer Science, vol. 11211, Springer, 2018, pp. 294–310.
[59] T. Dai, J. Cai, Y. Zhang, S. Xia, L. Zhang, Second-order attention network for single image super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 11065–11074.
[60] C. Papageorgiou, T. Poggio, A trainable system for object detection, Int. J. Comput. Vis. 38 (1) (2000) 15–33.
[61] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, IEEE Computer Society, 2005, pp. 886–893.
[62] M. Enzweiler, D. Gavrila, Gavrila, D.M.: monocular pedestrian detection: Survey and experiments. IEEE Trans. on PAMI 31(12), 2179-2195, IEEE Trans. Pattern Anal. Mach.Intell. 31 (12) (2010) 2179–2195.
[63] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach.Intell. 34 (4) (2011) 743–761.
[64] C. Wojek, S. Walk, B. Schiele, Multi-cue onboard pedestrian detection, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, IEEE Computer Society, 2009, pp. 794–801.
[65] G. Overett, L. Petersson, N. Brewer, L. Andersson, N. Pettersson, A new pedestrian dataset for supervised learning, in: 2008 IEEE Intelligent Vehicles Symposium, IEEE, 2008, pp. 373–378.
[66] A. Ess, B. Leibe, L. Van Gool, Depth and appearance for mobile scene analysis, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[67] W. Bo, R. Nevatia, Cluster boosted tree classifier for multi-view, multi-pose object detection, in: IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, 2007.
[68] S. Zhang, Y. Xie, J. Wan, H. Xia, S.Z. Li, G. Guo, WiderPerson: a diverse dataset for dense pedestrian detection in the wild, IEEE Trans. Multimed. 22 (2) (2020) 380–393.
[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[70] G. Brazil, X. Liu, Pedestrian detection with autoregressive network phases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7231–7240.

Yi Jin is an associate professor with the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include image processing and signal processing. She received the Ph.D. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University, in 2010. She was a Visiting Scholar with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2013 to 2014. (SM'06-M'13), contact her at yjin@bjtu.edu.cn.

Yue Zhang is currently working toward the Ph.D. degree at the Institute of Information Science, Beijing Jiaotong University, and also with Beijing Key Laboratory of Advanced Information Science and Network Technology. Her main research interests include machine learning, person re-identification, metric learning, and deep learning. Contact her at 17112065@bjtu.edu.cn.

Yigang Cen is a professor in the Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology. His research interests include computer vision, image understanding/processing, IOT etc. He received the Ph.D. degree in control theory and control engineering from the Huazhong University of Science Technology, Wuhan, China, in 2006. In 2006, he joined the Signal Processing Centre, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, as a Research Fellow. From 2014 to 2015, he was a Visiting Scholar with the Department of Computer Science, University of Missouri, Columbia, MO, USA. (M'09), He is a corresponding author of this article. Contact him at ygcen@bjtu.edu.cn.

Yidong Li is the Vice-Dean and a professor in the School of Computer and Information Technology at Beijing Jiaotong University. Dr. Li received his B.Eng. degree in electrical and electronic engineering from Beijing Jiaotong University in 2003, and M.Sci. and Ph.D. degrees in computer science from the University of Adelaide, in 2006 and 2010, respectively. Dr. Li's research interests include big data analysis, data privacy and security, advanced computing and intelligent transportation. Dr. Li has published more than 100 research papers in various journals (such as IEEE Trans. on Information Forensics & Security, IEEE Trans. on Intelligent Transportation Systems, IEEE Trans. on Knowledge & Data Engineering, IEEE Trans. on Cybernetics), and refereed conferences (such as SIGKDD, CVPR, AAAI). He has also co-authored/co-edited 5 books (including proceedings) and contributed several book chapters. He has organized several international conferences and workshops and has also served as a program committee member for several major international conferences such as ICML/PKDD, PAKDD, NFOSCALE, WAC, SAC, PDCAT, DANTH, and PAAP. Contact him at ydli@bjtu.edu.cn.

Vladimir Mladenovic is an associate professor with Faculty of Technical Sciences University of Kragujevac. His research interests include big data analysis and security, image processing, computer vision. Contact him at vladimir.mladenovic@ftn.kg.ac.rs.

Viacheslav Voronin is the head of the Center for Cognitive Technology and Machine Vision at Moscow State University of Technology STANKIN, Moscow, Russian Federation. He received his BS (2006), MS (2008) in the communication system from the South-Russian State University of Economics and Service, and his Ph.D. in technics from Southern Federal University (2009). Voronin is a member of the Program Committee of the conference SPIE. His research interests include image processing, inpainting, and computer vision. Contact him at voronin_sl@mail.ru.