When Pedestrian Detection Meets Nighttime Surveillance: A New Benchmark

Xiao Wang[1], Jun Chen[†1], Zheng Wang[†2], Wu Liu[3], Shin'ichi Satoh[2], Chao Liang[1], Chia-Wen Lin[4]
[1]Wuhan University
[2]National Institute of Informatics
[3]AI Research of JD.com
[4]National Tsing Hua University
{hebeiwangxiao,chenj,cliang}@whu.edu.cn, {wangz,satoh}@nii.ac.jp, liuwu1@jd.com,
cwlin@ee.nthu.edu.tw

Abstract

Pedestrian detection at nighttime is a crucial and frontier problem in surveillance, but has not been well explored by the computer vision and artificial intelligence communities. Most of existing methods detect pedestrians under favorable lighting conditions (*e.g.* daytime) and achieve promising performances. In contrast, they often fail under unstable lighting conditions (*e.g.* nighttime). Night is a critical time for criminal suspects to act in the field of security. The existing nighttime pedestrian detection dataset is captured by a car camera, specially designed for autonomous driving scenarios. The dataset for nighttime surveillance scenario is still vacant. There are vast differences between autonomous driving and surveillance, including viewpoint and illumination. In this paper, we build a novel pedestrian detection dataset from the nighttime surveillance aspect: *NightSurveillance*. As a benchmark dataset for pedestrian detection at nighttime, we compare the performances of state-of-the-art pedestrian detectors and the results reveal that the methods cannot solve all the challenging problems of *NightSurveillance*. We believe that *NightSurveillance* can further advance the research of pedestrian detection, especially in the field of surveillance security at nighttime. https://github.com/xiaowang1516/NightSurveillance

1 Introduction

Pedestrian detection, which locates all pedestrians in an image, has aroused increasing attention in the computer vision and artificial intelligence communities. This technology is also a basis for many advanced applications, such as person re-identification [Liu *et al.*, 2018; Wanigasekara *et al.*, 2019; Zeng *et al.*, 2020], pedestrian representation [Sun *et al.*, 2019; Liu *et al.*, 2017; Gan *et al.*, 2016], pedestrian action prediction [Zhang *et al.*, 2015; Liu *et al.*, 2020]. Thanks to the contributions of various deep learning methods and plentiful pedestrian datasets, significant progress has been made in pedestrian detection. For example, the performance of pedestrian detection on the most popular dataset (Caltech [Dollar *et al.*, 2012]) is nearly saturated, with an average miss rate of 4.54% by the state-of-the-art detector [Liu *et al.*, 2019].

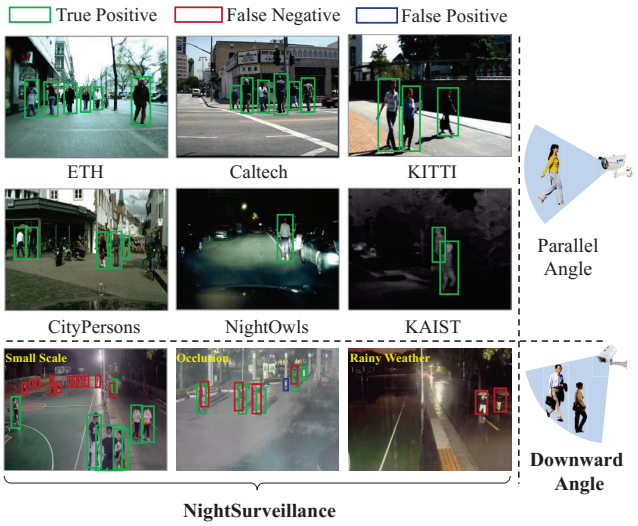

Figure 1: Some typical examples and results of the S3D detector to show the differences of our NightSurveillance dataset and existing datasets. Existing datasets include ETH, Caltech, KITTI, CityPersons, NightOwls, and KAIST. For our datasets, we show some real-world challenging conditions, such as small scale, occlusion, and rainy weather.

However, when pedestrian detection comes to the condition at nighttime, the performance drops a lot. Taking the SDS R-CNN detector [Brazil *et al.*, 2017] as an example, its average miss rate is 7.36% when applied to the daytime dataset (Caltech [Dollar *et al.*, 2012]), while the average miss rate increases drastically to 64% on a nighttime dataset (NightOwls [Neumann *et al.*, 2018]). Moreover, although the NightOwls dataset can serve the purpose of evaluating pedestrian detection performances at nighttime, it is built under an autonomous driving scenario, which ignores another important scenario in the field of security: surveillance. There are two main differences between autonomous driving and surveillance scenarios.

[†]Corresponding authors
Jun Chen is also with the National Engineering Research Center for Multimedia Software, Hubei Key Laboratory of Multimedia and Network Communication Engineering, Collaborative Innovation Center of Geospatial Technology.

Viewpoint. For autonomous driving, pedestrians are photographed by cameras at parallel angles. For surveillance, pedestrians are photographed by cameras at downward angles as illustrated in Figure 1. Difficult cases (small scale, occlusion, and rain) in such a situation are more challenging.

Imbalanced illumination. In autonomous driving at nighttime, the light distribution is more concentrated in the bottom and middle of a picture with a light source mainly coming from the car itself. However, the illumination in surveillance is relatively imbalanced at nighttime, involving weak street lights and strong vehicle lights (see Figure 2).

Pedestrian detection at nighttime is a challenging problem largely underrepresented in the literature, while it is crucial in surveillance applications. Current benchmarks are insufficient to bridge such a gap between autonomous driving and surveillance scenarios. In this paper, to advance this research field and benefit practical security applications, we construct a new pedestrian detection dataset from the nighttime surveillance aspect, namely, *NightSurveillance*. Our dataset is designed to comprehensively cover the following challenging factors: 1) *Scales.* The scales of pedestrians in surveillance are various. The small scale is particularly important. The number of pixels in a small-scale pedestrian is very limited, leading to less discriminant information for classification. Detecting pedestrian with small scales remains a challenging problem due to the lack of discriminating details. 2) *Lightness.* The light sources in nighttime surveillance are unstable, which brings imbalanced illumination. Imbalanced illumination induces large contrast variations in images. Detecting pedestrians in low-contrast regions results in loss of color or shape information, thereby making it difficult to separate foreground and background regions. 3) *Occlusion.* Pedestrians are randomly distributed in a surveillance area. Pedestrians are inevitably occluded by other objects. These pedestrians are notoriously hard to be detected due to the lack of key information and the substantially various appearances in the tricky occlusion situations. 4) *Rain.* Weather often change along time in long-term surveillance. Different weather conditions cause different visual variations that significantly impact the performance of detectors. Rain is another frequent weather except sunny day. Rain reduces the contrast dramatically and adds reflections to road surfaces. Moreover, the interference of an umbrella held by a pedestrian usually affects the detection of the pedestrian. 5) *Blur.* The temperature difference and aging equipment caused by long-term shooting lead to another common abnormal situation: blur. In the early morning, the fog always interferes with camera imaging.

For the challenges above, our *NightSurveillance* provides complete annotations which have covered all the cases discussed above. Images are collected from cover 20 cameras with an extended period from 17:00 to the next 5:00. We find that state-of-the-art pedestrian detection methods do not perform well on our *NightSurveillance* dataset, while the improvement is critical in security. We believe that *NightSurveillance* can further advance the research of pedestrian detection, especially in the field of surveillance security at nighttime.

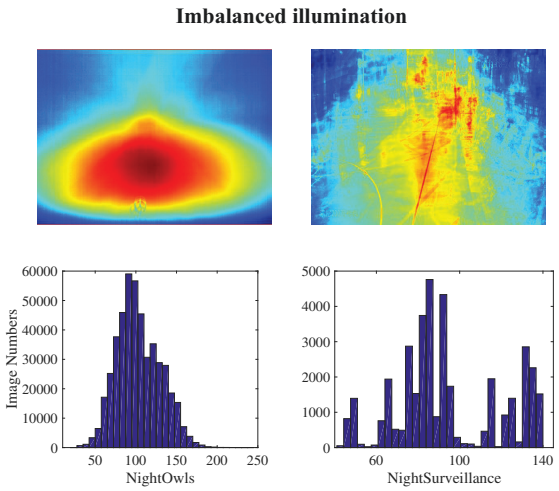

Figure 2: Distribution of illumination in the image and histogram of the image illumination in NightOwls (left) and *NightSurveillance* (right) datasets.

2 Related Datasets

In this section, we will make a brief survey of related datasets for pedestrian detection, including daytime dataset, nighttime dataset, and the differences between the surveillance and the autonomous driving scenarios at nighttime.

2.1 Daytime Datasets

Several datasets have been built for pedestrian detection at daytime, such as INRIA [Dalal and Triggs, 2005], ETH [Ess *et al.*, 2008], TUD-Brussels [Wojek *et al.*, 2009], Daimler [Enzweiler and Gavrila, 2009]. INRIA ($2k$ images), ETH (2k images), and TUB-Brussels (508 images) are not suitable for deep learning model training due to the limited quantity (as shown in Table 1). Daimler [Enzweiler and Gavrila, 2009] contains only grayscale images result in less discriminative information. Since then, the larger and richer datasets have been proposed, such as Caltech [Dollar *et al.*, 2012], KITTI [Geiger *et al.*, 2012] and CityPersons [Zhang *et al.*, 2017]. Caltech [Dollar *et al.*, 2012] has been widely used due to the large scale annotation, approximately $285k$ pedestrian bounding boxes and $250k$ images. Later researcher [Zhang *et al.*, 2016b] annotated the misaligned samples and improved the quality of the Caltech dataset. KITTI [Geiger *et al.*, 2012] focuses on the multi-sensor cameras for multiple tasks, and the portion for pedestrian detection is relatively small. CityPersons [Zhang *et al.*, 2017] annotated $31k$ pedestrians and $5k$ images. Pedestrian detection datasets at daytime have been comprehensive, leading to outstanding progress. The detection performance is not satisfactory when the SDS R-CNN detector (learned from Caltech [Dollar *et al.*, 2012]) is applied to the nighttime dataset (NightOwls [Neumann *et al.*, 2018]). For the security field, the night scene is more important than the daytime scene.

2.2 Nighttime Datasets

There are two kinds of existing datasets constructed at nighttime, *i.e.*, the multispectral dataset KAIST [Hwang *et al.*,

	Dataset	Train	Test	All	
		#images/#bboxes	#images/#bboxes	#images	#pedestrian/frame ↑
Daytime	KITTI [Geiger *et al.*, 2012]	$7k$/$4k$	–	$7k$	0.6
	Daimler [Enzweiler and Gavrila, 2009]	$22k$/$14k$	–	$22k$	0.65
	INRIA [Dalal and Triggs, 2005]	$2k$/$1k$	288/589	$2k$	0.86
	Caltech [Dollar *et al.*, 2012]	$128k$/$153k$	$121k$/$132k$	$250k$	1.14
	TUD [Wojek *et al.*, 2009]	508/$1k$	–	508	2.95
	CityPersons [Zhang *et al.*, 2017]	$3k$/$17k$	$1.5k$/$14k$	$5k$	7
	ETH [Ess *et al.*, 2008]	$2k$/$14k$	–	$2k$	7.85
Nighttime	NightOwls [Neumann *et al.*, 2018]	$128k$/$38k$	$103k$/$8k$	$231k$	0.20
	KAIST [Hwang *et al.*, 2015]	$17k$/$17k$	$16k$/$12k$	$33k$	0.86
	NightSurveillance	$19k$/$26k$	$19k$/$26k$	$38k$	2.46

Table 1: The number of frames and pedestrian annotations in datasets.

2015] and the color dataset NightOwls [Neumann *et al.*, 2018]. Half of KAIST consists of thermal images, which are not feasible due to its difficulty to distinguish the clothes and identity of the pedestrian. NightOwls is a large-scale dataset for pedestrian detection at nighttime. All the images are fully annotated and contain additional visual attributes such as occlusion, pose and difficulty, as well as tracking information to identify the same pedestrian across multiple frames. However, KAIST and Nightowls were captured by the car cameras, which are mainly used for autonomous driving, focusing on pedestrians on the sidewalk. They are not suitable for surveillance applications.

2.3 Surveillance *vs.* Autonomous Driving

The surveillance dataset for pedestrian detection at nighttime is still vacant. The characteristic of the data captured from surveillance and from autonomous driving scenarios are quite different, mainly in the following aspects: 1) *Viewpoint.* In autonomous driving scenario, pedestrians are photographed by the car camera at parallel angles. In surveillance scenario, pedestrians are pictured by the camera at downward angles. Pedestrians are significantly different in the two scenarios (Figure 1). Hence, their details of attributes are also different, such as small scale and occlusion situations. 2) *Imbalanced illumination.* The distribution of lightness in autonomous driving scenario [Gan *et al.*, 2019] is mainly concentrated in the middle part of the images with the help of car lights (as shown in the NightOwls of Figure 2). In contrast, the lighting in surveillance is more chaotic due to the multiple light sources, leading to imbalanced illumination. Pedestrian's robustness is more sensitive to imbalanced illumination. 3) *Shooting time.* The shooting time of surveillance camera is long-term, which covers a variety of weather, including not only normal weather but also rainy weather. Nighttime often lasts from 17:00 to 5:00 the next day. Shooting time of car camera is often indirect, which is not long-term. It can avoid significant temperature differences, lousy weather, and equipment ageing. Therefore, the image of autonomous driving is clearer, while the image of the surveillance is more blurred.

3 NightSurveillance

The purpose of *NightSurveillance* dataset is to provide a new benchmark and to improve the performance of pedestrian detection at nighttime. *NightSurveillance* is the key catalyst for pedestrian detection in the security system. In this section, we describe the detail of *NightSurveillance* data.

Data Collection. We collected 20G data from cameras in the real campus surveillance, where resolution of image is 1920×1080. The dataset contained both blurred and sharp images. The recordings were collected from 20 cameras during a period of 17:00 to 5:00 the next day. The dataset has covered two common weather: sunny and rainy day. Similar to existing datasets, the attributes (Table 3) of the *NightSurveillance* have been classified into several groups to allow more fine-grained evaluation using different settings.

Bounding box annotation. The *NightSurveillance* dataset already provides bounding boxes level for pedestrians. For annotations, the coordinates of the upper left and lower right of the pedestrian in image have been recorded. The whole annotation process consists of several stages, as follows: 1) *Keyframe extraction.* An hour of video is randomly taken from each camera. For each video, we use FFmpeg to extract eight keyframes per second, resulting in a total of 576K keyframes. 2) *Annotating bounding boxes on keyframes.* We used a pedestrian detector (Faster R-CNN [Ren *et al.*, 2017] pre-trained on MSCOCO [Lin *et al.*, 2014]) to detect the pedestrian in the keyframes, and 30k bounding boxes can be obtained. 3) *Manual correction.* To ensure the accuracy of the intercepted bounding boxes, we added a manual assisted verification phase using colabeler tool. Six volunteers are invited to check and correct the bounding boxes for keyframes. Since fewer pedestrians are active at night, there are a large number of frames with no pedestrians at nighttime than in the daytime. To reduce these invalid frames, we have removed most frames with no pedestrians. Then, we split the rest data into train and test portion according to the ratio of 1:1, and the overall distribution is shown in Table 1.

Data Diversity. We will explain the data diversity of the *NightSurveillance* proposed in this paper, including data size, occlusion, scale, lighting, blur, rain, and attribute statistics. 1) *Data Size.* We count the number of frames and bounding boxes in the existing datasets and *NightSurveillance* (as shown in Table 1). It can be seen that most of the data is in the daytime, and the datasets at night are less, only including KAIST and NightOlws. Compared with the existing night data, our night dataset has two advantages: the number

Setting	#Occlusion	#Scale			#Lighting			#Blur	#Rainy	#All
		#Small	#Medium	#Large	#Low	#Medium	#High			
Train	$12k$(25%)	$22k$(47%)	$12k$(26%)	$13k$(28%)	$9k$(20%)	$30k$(63%)	$8k$(17%)	$1k$(2%)	$2k$(4%)	$47k$
Test	$12k$(26%)	$21k$(46%)	$12k$(26%)	$13k$(28%)	$8k$(17%)	$30k$(65%)	$8k$(18%)	$1k$(2%)	$2k$(4%)	$46k$
All	$24k$(26%)	$43k$(46%)	$24k$(26%)	$26k$(28%)	$17k$(18%)	$60k$(65%)	$16k$(17%)	$2k$(2%)	$4k$(4%)	$93k$

Table 2: The proportion of pedestrians with different settings in *NightSurveillance* dataset.

of night images and the average number of pedestrians per image. The number of nights frames in KAIST is 33k, and that in our dataset is 38k. The number of nights frames in our dataset is 1.2 times of that in KAIST dataset. The average number of objects per frame is lower, because the natural streets are less busy at night. The average number of pedestrians per image in NightOlws is 0.2, while that in our dataset is 2.4. The occupancy rate in our dataset is 12 times of that in NightOlws.

2) *Occlusion.* Occlusion is a vital research topic. To elude from the camera, suspects often hide behind other things or pedestrians, leading to occlusion. These suspects are notoriously hard to detect due to the substantially various appearance in the intricate occlusion, especially at nighttime. We have counted the number of occlusion pedestrians in our dataset, and the proportion distribution in each portion is shown in Table 2. It can be seen that the number of occlusion pedestrians covers a quarter of all pedestrians and provides sufficient samples for the occlusion topic.

3) *Data Scale.* The multi-scale has always been a critical topic for pedestrian detection, especially in small scale pedestrian. The small scale pedestrian has less information than large scale pedestrian. These pedestrian are easy to be confused with the backgrounds, leading to lower performance. It is the main part of the false negatives. The pedestrians have been divided into Small ($<$90), Medium (90-150), and Big ($>$ 150) based on the height (pixel) of the pedestrian in surveillance. The majority of this dataset contains small scale, because of the far distance from the camera to pedestrians in surveillance. The proportion of each part is shown in the Table 2, where the small-scale pedestrians account for a large percentage (approximately 50%).

4) *Illumination.* The main difference between day and night is the lightness. The light source in the day comes from natural light, while the light source in the night comes from the street lamp and vehicle lamp. Pedestrian is more sensitive to different light. It is harmful for pedestrians faced with both weak light and strong light. According to the diversity of lightness in *NightSurveillance*, we divide the data into three levels: Low ($<$80), Medium (80-130), and High ($>$130) based on the lightness of pedestrian images. We have calculated the histogram of the mean lightness of images in the *NightSurveillance*, as shown in right of Figure 2. The corresponding proportional distribution is shown in Table 2.

5) *Attributes.* Compared with the existing datasets, the most significant difference in our dataset is a downward angle from surveillance. The current datasets have been captured with the aid of a car. The viewpoint between the camera and the pedestrian is parallel, leading to complete contour of pedestrians. By observing the real surveillance scene, we find that the complexity mainly comes from many aspects, such as occlusion, small-scale, blur, rain, and lighting. The best standard to evaluate a dataset is whether it covers the real scene as much as possible. Our dataset has covered all of the above aspects. The statistics of the attributes of the existing datasets and our dataset are shown in Table 3.

Dataset	ImageSize	Data Diversity				
		Occlusion	Scale	Blur	Rainy	Lighting
KITTI	1392×512	✓				
Caltech	640×480	✓	✓			
CityPersons	2048×1024	✓	✓			
KAIST	640×480		✓			
NightOwls	1024×640	✓	✓			✓
NightSurveillance	1920×1080	✓	✓	✓	✓	✓

Table 3: Comparison of the annotation attributes.

4 Experiments

4.1 Experiments Settings

In this section, we compare the overall performance of classical pedestrian detection methods on the existing datasets and our dataset. These methods include ACF [Dollar *et al.*, 2014]), RPN+BF [Zhang *et al.*, 2016a], Vanilla Faster R-CNN [Ren *et al.*, 2017], Adapted Faster R-CNN [Zhang *et al.*, 2017]), SDS R-CNN detector [Brazil *et al.*, 2017], and S3D [Wang *et al.*, 2019] detectors.

4.2 Performance Study

Table 4 shows the evaluation results. From the results, we can determine that there is a gap between the performance of classical detectors on existing datasets and our dataset. The existing classical methods have better performance on the daytime datasets (KITTI, Caltech, and CityPersons). The mAP of Adapted Faster R-CNN is 66.73% on KITTI dataset. The miss rate of RPN+BF is 7.31% on CityPersons dataset. These methods have achieved satisfactory results on the daytime dataset. For Nightowls dataset, the performance tends to decline. The best miss rate is 14.32% for S3D. The main reason is that the light from night data is weak, which interferes the detection process. For our dataset, the performance is further terrible, while the miss rate is up to 21.73%, which is the best performance of the classical detector. Therefore, the *NightSurveillance* is more difficult and more challenging than the existing datasets of autonomous driving.

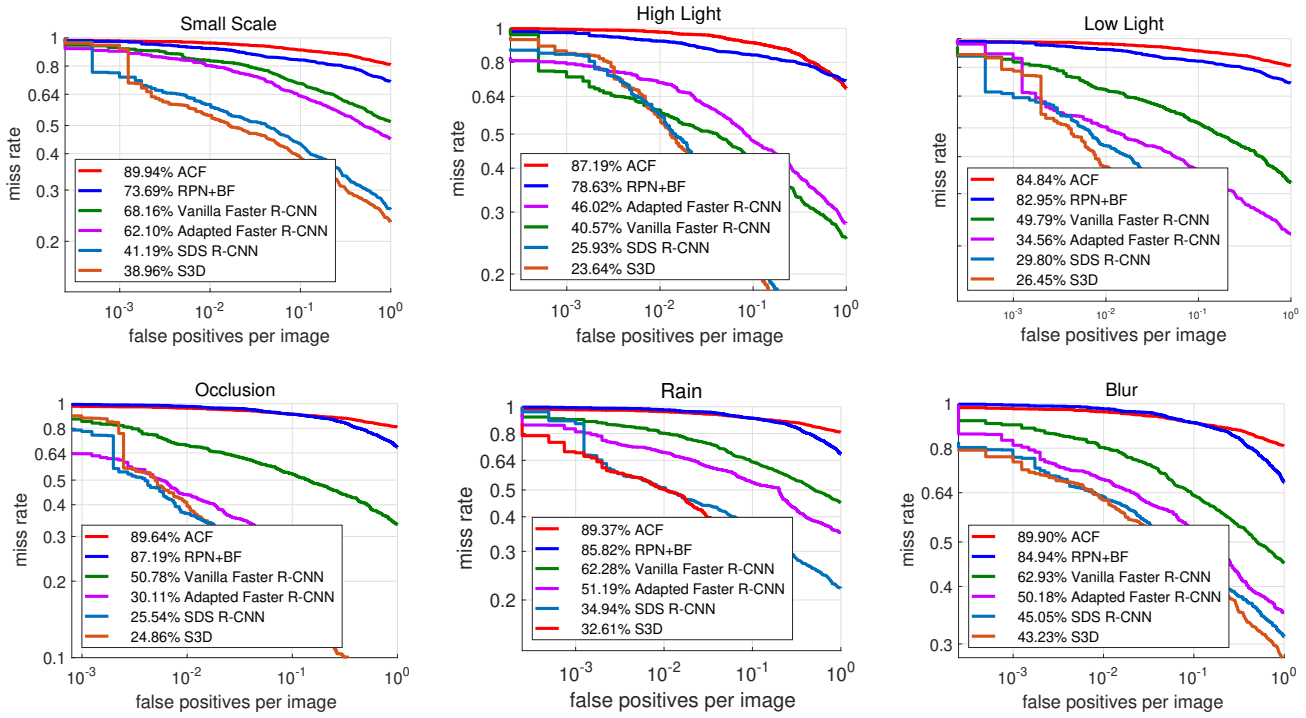

Figure 3: Miss rates versus false positives of existing classical algorithms on different settings of *NightSurveillance* dataset. Lower curve means better performance, the legend denotes average miss rate.

Methods	KITTI	Caltech	CityPersons	NightOwls	*NightSurveillance*
	mAP (%)	Average Miss Rate (%)			
ACF [Dollar *et al.*, 2014]	47.29	27.63	33.10	51.68	89.34
RPN+BF [Zhang *et al.*, 2016a]	61.29	9.58	7.31	23.26	86.34
Vanilla Faster [Ren *et al.*, 2017] R-CNN	65.91	20.98	23.46	20.00	26.55
Adapted Faster R-CNN [Zhang *et al.*, 2017]	66.72	10.27	12.81	18.81	24.84
SDS R-CNN [Brazil *et al.*, 2017]	63.05	7.36	13.26	17.80	23.62
S3D [Wang *et al.*, 2019]	65.60	9.28	11.24	14.32	21.73

Table 4: Comparison of state-of-the-art pedestrian detection methods trained and tested on the corresponding dataset with the protocol of Average Miss Rate and mAP.

To explore the reason that the performance of these detectors is unsatisfactory on our dataset, we have made further exploration. We have divided the data into different portions and observed the performance of detectors in each portion (scale, light, occlusion, rain, and blur). The performance is shown in the Figure 3.

Scales. We evaluated the performance of all detectors on the *NightSurveillance* dataset, depending on different ground truth attributes, in line with the standard evaluation introduced by [Dollar *et al.*, 2012]. We show that the methods are not as sensitive to aspect ratios, but they are very sensitive to the scale of pedestrians. The deep learning methods benefit from the amount of training data and for the medium and big scales. This miss rate is comparable to daylight datasets, however for the small pedestrians in the small scale , the miss rate rises to 38.96% and dramatically high, and the accuracy of deep learning methods is also unsatisfactory.

Light. We also compare the performance based on the image lightness. The error is not lower for brighter images than the low ones. In contrast, the miss rate of medium lightness is lower. This is caused by overexposure and less lightness in image, which makes the detection very challenging. The core of detection is to locate the position of the pedestrian in the image, so the lightness of the pedestrian's patch has a more direct effect on pedestrian detection. Note that, in this evaluation, we counted the lightness of pedestrians' image patch, instead of image lightness.

Occlusion. For this experiment, we test the performance of the portion with the occlusion labels. The evaluation results of the classical detectors are shows in Figure 3. The best performance is the miss rate of 24.86%. The number of pedestrians in night scenes tends to be relatively small. While for safety awareness, they are often walking together. In the field of security, suspects will also hide behind other pedestri-

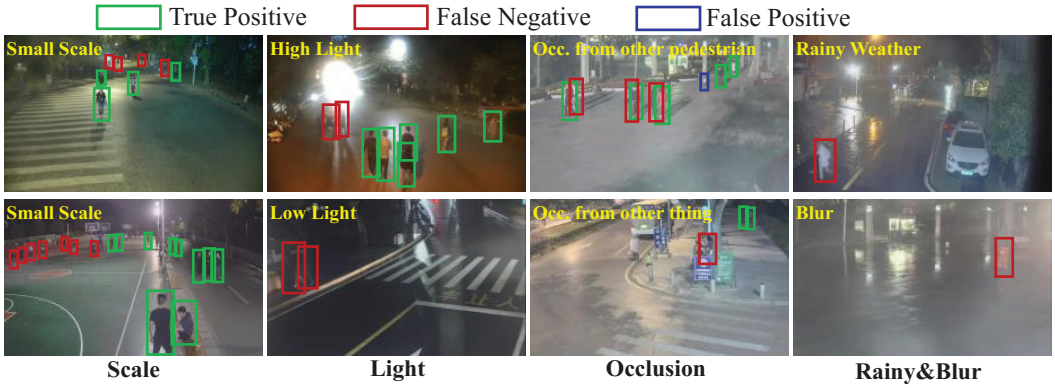

Figure 4: Subjective results on typical samples with different settings of *NightSurveillance* dataset.

ans or other things. All of the above will cause the occlusion of the surveillance scene at night. Current performance is not satisfied with the actual demand. Performance optimization under occlusion conditions is crucial for surveillance security applications.

Rain. Figure 3 also shows the performance of the existing classical detectors on the rainy portion of our dataset, which tends to worsen. The lowest missing rate is 32.61% from S3D detector. The most common weather include sunny, overcast, and rainy day. It is not much difference between the first two types. Another common weather is rain. Anomalous events at night do not stop because of weather changes. The rain will affect the imaging quality of the camera in surveillance due to raindrops. Besides, from the perspective of pedestrians, foreign umbrellas will also interfere with pedestrian discrimination. All the above situations have brought challenges to pedestrian detection on rainy weather at nighttime. The current methods are not directly applied to practice, and it is an urgent problem to be solved.

Blur. Figure 3 also shows the performance of the blur portion of our dataset, and the existing classical detectors are unfortunate. The lowest missing rate is 43.23% from S3D detector, which is far from practical application. The shooting of the surveillance camera is uninterrupted, including day and night. It is easy to bring about the equipment ageing situation due to the long-term shooting, as a result of the camera imaging quality. The large temperature difference between day and night leads to increased fog, leading to a seriously blurred situation. On the other hand, the quality of images also reduces when we sample images from the second day of the night in surveillance due to the less light. The current methods cannot be directly applied to practice, and the corresponding optimization techniques are needed to overcome the above problems.

4.3 Visualization and Analysis

We observed all the detection results from S3D detector, and selected parts for display (as shown in Figure 4). We also made further thoughts on the issues in each setting. For scale issues, the failure samples mostly come from pedestrians with the far distance from the camera. We can optimize with the technology of local magnification. For lightness, the failure samples mostly come from exposure or low light. We can use the technology of lighting compensation to optimize. We can also use semantic segmentation [Xu *et al.*, 2019] to locate the visualization area or learn two detectors for nighttime to optimize occlusion. For rain or blur, we could use the image enhancement technology to make up for it. Moreover, we can also reduce the background interference through the background modelling technology.

5 Conclusion

Although most of the existing pedestrian detection methods have excellent performance in the daytime datasets, they can not be directly applied to the nighttime scene. The current night pedestrian detection dataset is still shot by a car camera, which is suitable for autonomous driving scenario. The dataset of surveillance scenario is still vacant. There is a huge difference between autonomous driving and surveillance, including viewpoint and illumination. In this paper, we have introduced a novel comprehensive pedestrian dataset to encourage research on night images for surveillance. *NightSurveillance* has covered the complicated situation in the existing real scenario, including small scale, unbalanced illumination, occlusion, blur, etc. We believe that the *NightSurveillance* dataset and benchmark for pedestrian detection can provide precious data support to stimulate more potential detection methods.

Clarity of privacy. We should clarify that there is no privacy issue raised by the NightSuveillance dataset. This dataset has been collected from the actual scene. Previous work often blurs human faces manually. Our dataset does not contain clear facial identity information and is built just for research purpose, which does not have this kind of issue.

Acknowledgments

This work is supported by National Key R&D Program of China (No.2017YFC0803700), National Nature Science Foundation of China (No. U1611461, 61876135, 61801335, U1903214), the supercomputing system in the Supercomputing Center of Wuhan University, Hubei Province Technological Innovation Major Project (2018AAA062, 2018CFA024, 2019CFB472), and the Grant-in-Aid for JSPS Fellows 18F18378.

References

[Brazil *et al.*, 2017] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection & segmentation. In *ICCV*, 2017.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[Dollar *et al.*, 2012] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 2012.

[Dollar *et al.*, 2014] Piotr Dollar, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *TPAMI*, 2014.

[Enzweiler and Gavrila, 2009] Markus Enzweiler and Dariu M Gavrila. Monocular pedestrian detection: Survey and experiments. *TPAMI*, 2009.

[Ess *et al.*, 2008] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008.

[Gan *et al.*, 2016] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, 2016.

[Gan *et al.*, 2019] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *ICCV*, 2019.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[Hwang *et al.*, 2015] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[Liu *et al.*, 2017] Wu Liu, Chenggang Clarence Yan, Jiangyu Liu, and Huadong Ma. Deep learning based basketball video analysis for intelligent arena application. *MTAP*, 2017.

[Liu *et al.*, 2018] Wu Liu, Cheng Zhang, Huadong Ma, and Shuangqun Li. Learning efficient spatial-temporal gait features with deep learning for human identification. *Neuroinformatics*, 2018.

[Liu *et al.*, 2019] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *CVPR*, 2019.

[Liu *et al.*, 2020] Kun Liu, Wu Liu, Huadong Ma, Mingkui Tan, and Chuang Gan. A real-time action representation with temporal encoding and deep compression. *TCSVT*, 2020.

[Neumann *et al.*, 2018] Lukáš Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, et al. Nightowls: A pedestrians at night dataset. In *ACCV*, 2018.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *TPAMI*, 2017.

[Sun *et al.*, 2019] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019.

[Wang *et al.*, 2019] Xiao Wang, Chao Liang, Chen Chen, Jun Chen, Zheng Wang, Zhen Han, and Chunxia Xiao. S3d: Scalable pedestrian detection via score scale surface discrimination. *TCSVT*, 2019.

[Wanigasekara *et al.*, 2019] Nirandika Wanigasekara, Yuxuan Liang, Siong Thye Goh, Ye Liu, Joseph Jay Williams, and David S Rosenblum. Learning multi-objective rewards and user utility function in contextual bandits for personalized ranking. In *IJCAI*, 2019.

[Wojek *et al.*, 2009] Christian Wojek, Stefan Walk, and Bernt Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, 2009.

[Xu *et al.*, 2019] Yonghao Xu, Bo Du, Lefei Zhang, Qian Zhang, Guoli Wang, and Liangpei Zhang. Self-ensembling attention networks: Addressing domain shift for semantic segmentation. In *AAAI*, 2019.

[Zeng *et al.*, 2020] Zelong Zeng, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Illumination-adaptive person re-identification. *TMM*, 2020.

[Zhang *et al.*, 2015] Lefei Zhang, Liangpei Zhang, Dacheng Tao, and Bo Du. A sparse and discriminative tensor to vector projection for human gait feature representation. *Signal Processing*, 2015.

[Zhang *et al.*, 2016a] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster R-CNN doing well for pedestrian detection? In *ECCV*, 2016.

[Zhang *et al.*, 2016b] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *CVPR*, 2016.

[Zhang *et al.*, 2017] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017.