NightOwls: A Pedestrians at Night Dataset

Lukáš Neumann[1⋆], Michelle Karg[2*], Shanshan Zhang[3*], Christian Scharfenberger[2], Eric Piegert[2], Sarah Mistr[2], Olga Prokofyeva[2], Robert Thiel[2], Andrea Vedaldi[1], Andrew Zisserman[1], and Bernt Schiele[4]

[1] Department of Engineering Science, University of Oxford
[2] Continental Corporation, BU Advanced Driver Assistance Systems
[3] Nanjing University of Science and Technology
[4] Max Planck Institute for Informatics
http://www.nightowls-dataset.org/

Abstract. We introduce a comprehensive public dataset, *NightOwls*, for pedestrian detection at night. In comparison to daytime conditions, pedestrian detection at night is more challenging due to variable and low illumination, reflections, blur, and changing contrast.
NightOwls consists of 279k frames in 40 sequences recorded at night across 3 countries by an industry-standard camera, including different seasons and weather conditions. All the frames are fully annotated and contain additional object attributes such as occlusion, pose and difficulty, as well as tracking information to identify the same object across multiple frames. A large number of background frames for evaluating the robustness of detectors is included, a validation set for local hyper-parameter tuning, as well as a testing set for central evaluation on a submission server is provided.
As a baseline for pedestrian detection at night time, we compare the performance of ACF, Checkerboards, Faster R-CNN, RPN+BF, and SDS-RCNN. In particular, we demonstrate that state-of-the-art pedestrian detectors do not perform well at night, even when specifically trained on night data, and we show there is a clear gap in accuracy between day and night detections. We believe that the availability of a comprehensive night dataset may further advance the research of pedestrian detection, as well as object detection and tracking at night in general.

1 Introduction

Detecting and tracking people is one of the most important applied problems in computer vision. Significant applications such as entertainment, surveillance, robotics, and assisted and automated driving, are all centered around people. They thus require highly-reliable people detectors that can work in a variety of indoor and outdoor scenarios and are robust to challenging visual effects such as variable appearance, inhomogeneous illumination, low resolution, occlusions and limited field of view.

⋆ equal contribution

While recent progress in object detection has been substantial, current systems may still fail to measure up to the demands of such requirements, particularly when, as in autonomous driving and surveillance, detecting people with high reliability is paramount for safety. Unfortunately, current benchmarks are insufficient to assess such limitations in a reliable manner, let alone support further research to address them.

In order to fill this gap, we introduce *NightOwls*, a new dataset to assess the limitations of state-of-the-art pedestrian detectors when used in extreme but realistic conditions. We focus in particular on detection at nighttime, a problem largely underrepresented in the literature, but which is very important in many applications - in assistive driving, in surveillance monitoring, or in autonomous driving as a key input to the sensor fusion. Vision sensors also benefit from the advantage of high-quality shape and color information, human interpretability of the sensor output and low energy consumption (passive sensor), which is not the case for other sensor modalities.

Our work is inspired by datasets such as PASCAL VOC [8], ImageNet [13] and MS-COCO [14], whose introduction kickstarted new waves of fundamental research in classification and detection, moving the field from bag-of-visual-words, to deformable parts model, and finally to deep convolutional neural networks. Benchmarks such as Caltech pedestrians [4, 5] had a similar impact in pedestrian detection.

For a dataset to be impactful, it must highlight important shortcomings in the current generation of algorithms. For pedestrian detection, the most frequently-used dataset, Caltech, is nearly saturated, with an average miss rate of 8.0% for state-of-the-art detectors [1, 22] compared to 83.0% average miss rate at the time of introduction [5]. This tremendous improvement suggests that the Caltech benchmark is almost "solved", at least if we take human performance, estimated at 5.6% by [22], as an upper bound.

While Caltech and similar benchmarks may be saturated, we cannot conclude that pedestrian detection is "solved" in general. A limitation with most datasets is that they focus on detection during the daytime. While this requires to cope with challenges such as occlusions and variable appearance, scale, and pose, doing so in poor lighting, and at nighttime in particular, is more challenging still. Empirically, we will show that current detectors are far below human performance in such conditions.

In order to do so, *NightOwls* is designed to be representative of the following challenges:

1. **Motion blur and image noise:** Imaging at night requires a trade-off between long exposure times and sensor gain, resulting in significant motion blur or noise.
2. **Reflections and high dynamics:** The variations in light intensity in night scenes, caused by inhomogeneous light sources and their reflections, may exceed the dynamic range of a camera, resulting in inhomogeneous illumination with under- and over-saturated areas.

Fig. 1: Sample images from the *NightOwls* dataset exhibiting the challenges of detection at night. Occlusion, low contrast, motion blur, image noise and inhomogeneous illumination (left) and different illumination and weather conditions (right).

3. **Large variation in contrast, reduced color information:** Inhomogeneous illumination induces large contrast variations in images. Detection is difficult in low-contrast regions and may result in loss of color information and in confusing foreground and background regions.
4. **Weather and seasons:** Weather and seasons cause other visual variations impacting the performance of detectors. While snow has the potential to illuminate a scene more homogeneously, rain can reduce the contrast dramatically and add reflections to road surfaces.

In addition to addressing these challenges, *NightOwls* has a number of additional desirable properties: (i) images are captured by an industry-standard camera for automotive, whereas other datasets often use generic cameras, (ii) full annotations for each frame are provided, in the standard MS-COCO and Caltech formats, (iii) multiple European cities and countries are represented, (iv) track identity information when an object is detected in multiple frames is provided, (v) a central evaluation server for results submission and comparison is available, and (vi) additional classes (cyclists, motorbikes) and attributes (pose, difficult) are annotated.

Empirically, we demonstrate that state-of-the-art pedestrian detection methods do not perform well on this dataset, even when specifically trained on night data, and we show the gap in accuracy between day and night detections is quite significant. While we primarily focus on detecting pedestrians, we also believe that the availability of a comprehensive night dataset may initiate further research in other domains, such as general object detection or tracking.

2 Related Work

Existing Datasets. Over the last decade, several datasets have been created for pedestrian detection. Early efforts include INRIA [2], ETH [7], TUD-Brussels

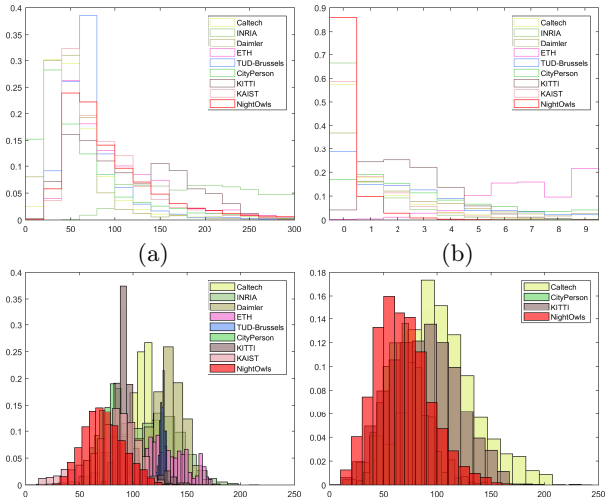

Table 1: Normalized histogram of pedestrian height in pixels (a), the number of objects per frame (b), the average image (c) and pedestrian patch (d) lightness of standard pedestrian datasets. Note that only datasets with an "occlusion" flag are shown for the patch statistics (d)

[17], and Daimler [6]. These datasets are now either too small (INRIA, ETH, TUD-Brussels) or only provide gray scale images (Daimler).

Recently, larger and richer datasets have been proposed and have become more popular, such as the Caltech [5], KITTI [10] and CityPersons [23] datasets. The Caltech dataset [4, 5] has been widely used as it provides a large number of annotations, including around 250, 000 frames and 185, 000 pedestrian bounding boxes. Yet, the diversity of the annotations is limited as the video was recorded in only 11 sessions within a single city, the alignment quality of the annotations is poor due to the interpolation implemented between neighboring frames, and it is only recorded at daytime. The noisy annotations were then further improved in [20].

The focus of the KITTI dataset is to encourage research in the field of a multi-sensor setup consisting of cameras, a laser scanner and GPS/IMU localization providing data for multiple tasks such stereo matching, optical flow, visual odometry/SLAM, object detection and 3D estimation [9, 10], but for pedestrian detection the dataset is relatively small.

The CityPersons dataset [23] consists of a large and diverse set of stereo video sequences recorded in streets from 27 cities in Germany and neighbouring countries. High quality bounding box annotations are provided for about 35k pedestrians in 5000 images. Additionally, fine pixel-level annotations of 30 visual classes are also available. The fine annotations include instance labels for persons and vehicles. However, the dataset does not have night or background images.

Also, it does not have driving sequences (it consists of individual images), and consequently it does not have examples of the same objects across multiple frames.

To the best of our knowledge, the KAIST dataset [12] is currently the only public dataset that contains some night images for pedestrian detection (5 out of 10 recordings are at night). The data was captured in one city in one season, which limits the diversity, the camera used for recording is a consumer-grade camera, which resulted in poor recording quality and considerable additional image noise, and the dataset does not provide occlusion labels which severely limits the ability to train on this dataset (see Section 4). The focus of the KAIST dataset is multi-spectral pedestrian detection which considers the data fusion from a thermal sensor and a RGB camera, as an attempt to overcome the issues of pedestrian detection at night. We note, that using just the thermal sensor for object detection may not be feasible because of its low spatial and dynamic resolution, the limited thermal footprint of people in clothes and their currently prohibitive cost for production vehicles.

The number of images and annotations in different datasets is summarised in Table 2, key statistics are compared in Table 1.

Pedestrian Detection at Night. Apart from KAIST, all the above datasets and the vast majority of work on pedestrian detection [1, 4, 5, 19, 21, 23] is focused on detection at daytime. Some early work attempted to solve the problem of object/pedestrian detection at night with the assistance from tracking methods [11, 18] or stereo images [15].

However, to our knowledge pedestrian or object detection at night has not attracted much attention in the research community, despite its importance for robust vision applications. We suspect the main reason is the lack of publicly available data for such research.

3 Dataset

In this section, we describe the data capture procedure, the annotation protocol, our design choices and the statistics of the dataset.

Data Recording. The dataset has been recorded in several cities across Europe with a forward-looking industry-standard camera, using windshield mounting identical to professional mounts in production vehicles. The data was collected at dawn and nighttime throughout the whole year and under different weather conditions (see Fig. 1). In total, 40 individual recordings were captured and then split into the training, validation and test sets, maintaining uniform distribution of key parameters such as weather and pedestrian pose/height difficulty.

Image Quality and Size. Research datasets [5,12] are often recorded with consumer camera equipment, which results in high level of image noise and limited dynamic range. To provide a night dataset with realistic variations in contrast and blurriness, the dataset was captured by an industry-standard camera (image resolution 1024×640), very similar one to the ones used in production vehicles.

Table 2: Image and pedestrian annotations counts in pedestrian detection datasets.

Dataset	Training				Validation				Test				All	
	Images	Pedestrian Bboxes	Pedestrian Tracks	Background Images	Images	Pedestrian Bboxes	Pedestrian Tracks	Background Images	Images	Pedestrian Bboxes	Pedestrian Tracks	Background Images	Images	Objects / Frame
Caltech [5]	**128k**	**153k**	**1k**	67k	-	-	-	-	**121k**	**132k**	**869**	61k	250k	1.14
INRIA [2]	2k	1k	0	1k	-	-	-	-	288	589	0	0	2k	0.86
Daimler [6]	22k	14k	0	15k	-	-	-	-	-	-	-	-	22k	0.65
ETH [7]	2k	14k	0	5	-	-	-	-	-	-	-	-	2k	7.85
TUD [17]	508	1k	0	145	-	-	-	-	-	-	-	-	508	2.95
KITTI [10]	7k	4k	0	6k	-	-	-	-	-	-	-	-	7k	0.60
KAIST [12]	50k	41k	495	32k	-	-	-	-	45k	45k	675	26k	95k	0.90
night subset	*17k*	*17k*	*141*	*10k*	-	-	-	-	*16k*	*12k*	*156*	*10k*	*33k*	*0.86*
CityPersons	3k	17k	-	672	500	3k	-	102	1.5k	14k	-	249	5k	7.00
NightOwls	128k	38k	1657	**105k**	**51k**	**9k**	**262**	**45k**	103k	8k	196	**97k**	**281k**	0.20

Table 3: Pedestrian attributes statistics.

	Occlusion	Pose		Height			All
		Sideways	Frontal	Far	Medium	Near	
Train	5k [20%]	9k [35%]	18k [64%]	16k [58%]	7k [24%]	5k [18%]	27k[66%]
Vali.	1k [13%]	2k [35%]	4k [64%]	3k [51%]	2k [29%]	1k [19%]	7k [16%]
Test	2k [25%]	2k [25%]	6k [75%]	5k [62%]	2k [26%]	1k [12%]	8k [18%]
All	8k [20%]	14k [33%]	28k [67%]	24k [58%]	11k [25%]	7k [17%]	**42k**

The dataset includes both blurred and sharp images and the quality realistically depends on the scene illumination and the vehicle speed.

Annotation. The frame-rate is 15fps and every frame was manually annotated. Every pedestrian, cyclists and motorcyclist (higher than 50px) is annotated with a bounding box, alongside with three attributes: occlusion, difficult (low contrast or unusual posture) and pose. People on posters, sculptures and groups where individuals are hard to separate are marked as "ignore". We note that compared to the existing datasets, the the average number of objects per frame is lower, because naturally streets are less busy at night (see Table 1).

As a result, the dataset contains $279k$ fully annotated frames with $42,273$ pedestrians, where $32k$ frames contain at least one annotated object and the remaining $247k$ are the background images. The annotations are provided in two standard MS-COCO [14] and Caltech (VBB) [4] formats, so that the new dataset can be plugged in to existing frameworks without any extra effort.

Fig. 2: Sample pedestrian, cyclist and motorcyclist annotations from the *NightOwls* dataset, including pose attribute.

Similarly to the Caltech dataset [4], the attributes are classified into several groups to allow more fine-grained evaluation using different data dimensions.

The pedestrian height is divided into *Far*, *Medium* and *Near* (see Table 3), based on the distance required to trigger automated breaking of a moving vehicle at different speeds. Using the pinhole camera model, the camera calibration parameters and average person height of (1.6m, 1.8m), the annotation height h is divided as

	$h \leq 90$	*Far*	braking distance at $\sim$ 50km/h
	$90 \leq h \leq 150$	*Medium*	braking distance at $\sim$ 40km/h
	$h \geq 150$	*Near*	braking distance at $\sim$ 30km/h

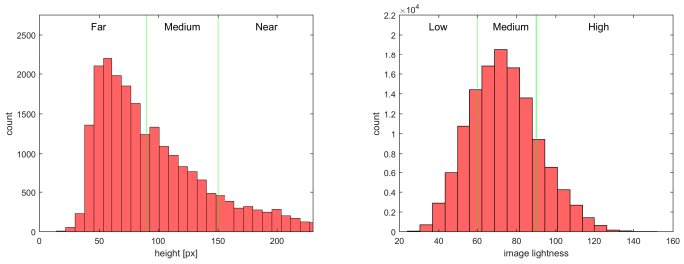

Fig. 3: Histogram of pedestrian scale (left) and image lightness (right) in the *NightOwls* dataset and the corresponding attribute categorisation

Table 4: Comparison of the annotation attributes.

Dataset	Year	Image Size	Attributes				Images				Annotations		
			Occlusion	Difficult	Pose	Tracking Id	Night	Dusk/Dawn	# Cities	Different Seasons	Pedestrian	Cyclist	Motorcyclist
Caltech [5]	2009	640x480	✓	x	x	✓	x	x	1	x	✓	x	x
Kitti [10]	2012	1392x512	✓	x	✓	x	x	x	1	x	✓	✓	✓
KAIST [12]	2015	640x480	x	✓	x	✓	✓	✓	1	x	✓	Driver	
CityPersons [23]	2017	2048x1024	✓	x	x	x	x	x	27	✓	✓	Rider	
NightOwls	2018	1024x640	✓	✓	✓	✓	✓	✓	7	✓	✓	✓	✓

We note that a majority of the pedestrians are categorized as *Far* (see Fig. 3 left), which is due to the exhaustive labeling process of every frame. Similarly, we also categorize image lightness as *Low*, *Medium* and *High*, based on the histogram of mean image lightness (see Section 4 and Fig. 3 right).

The pose is annotated as left, right, front and back, but we refer to them as *Frontal* (front, back) and *Sideways* (left, right). We note that there is a bias towards the *Frontal* pose in the data (see Table 3), which is given by the fact how people/cyclists typically move on and alongside the road.

Data Diversity. To achieve a high data diversity, which is desired for the generalization ability of detection algorithms, the recordings were collected in 7 cities across 3 countries (Germany, Netherlands, UK) during a period of five months. The dataset captures different weather conditions durin autumn, winter and spring, including rain and snow which change the lighting of the scene and add additional reflections..

Background Images. False positive rate is a major concern for real-world applications, because false alarms of safety-critical systems are not acceptable in driving scenarios. Moreover in these applications, the number of frames without

any object of interest is significantly higher than the number of frames with it, which increases the chance of false positives even further.

In order to support the research of robust detectors with low false positive rates and to reliably estimate detector precision, $247k$ background images are included in the dataset. For night images, especially regions with low illumination or reflections are typically prone to such false positives.

Temporal Tracking. Most methods focus on detection from a single frame, which is inherently more prone to both false positive as well as false negative errors. In order to enable research of more robust multi-frame detection methods, the dataset includes temporal tracking annotations, so that the same object can be identified across different frames.

Validation and Testing Set. Similarly to the recent large-scale datasets such as MS-COCO [14] or CityPersons [23], we explicitly split the data for evaluation into a validation and a testing set. We publish images for both sets, but only the annotations for the validation set are published - the testing set annotations are then only to be used by the evaluation server (see below). Both sets have similar data statistics and the validation set is sufficiently large, so that it can be used by the researchers for local evaluation and hyper-parameter tuning. An additional benefit of a common validation set is that the hyper-parameter experiments become comparable between different methods.

Evaluation Server. A central submission server is provided for dataset download and evaluation. The submissions of detection results (JSON format) are automatically evaluated on the testing set and a leader board is presented, so that all detection methods are evaluated in a single place. The submissions are limited to one submission a day to reduce the possibility of over-fitting to the testing set. Additionally, because the testing set is sufficiently large, we only publish performance on one subset on the leader board, whilst the performance of the second sequestered subset will remain private - if there is a significant discrepancy in the accuracy on the both subsets, this points towards over-fitting or training on the testing data.

4 Experiments

Methods. We have evaluated 6 recently published pedestrian detection algorithms on the existing datasets, as well as the newly introduced dataset:

ACF [3] Our experiments are based on the open source release of ACF[5]. One minor change we made is to use a larger model size (60×120 instead of 30×60 pixels), which shows to improve the vanilla ACF on several benchmarks, e.g. Caltech, KITTI. All other parameters are kept identical to the vanilla version.

Checkerboards [21] In contrast to ACF, the Checkerboards detector applies more filters with various sizes on top of the HOG+LUV channels in order to

[5] `https://github.com/pdollar/toolbox`

Table 5: Comparison of state-of-the-art pedestrian detection methods trained and tested on the corresponding dataset. Average Miss Rate (MR) or mean Average Precision (mAP) shown, as per the dataset protocol, using Reasonable subset

Method *metric*	Caltech *MR*	KITTI *mAP*	CityPersons *MR*	**ours** *MR*
ACF [3]	27.63%	47.29%	33.10%	51.68%
Checkerboards [21]	18.50%	56.75%	31.10%	39.67%
Vanilla Faster R-CNN [16]	20.98%	65.91%	23.46%	20.00%
Adapted Faster R-CNN [23]	10.27%	66.72%	12.81%	18.81%
RPN+BF [19]	9.58%	61.29%		23.26%
SDS-RCNN [1]	7.36%	63.05%	13.26%	17.80%

extract more representative features. We used the open source release[6] without tuning any parameters.

Vanilla Faster R-CNN [16] We reimplemented vanilla Faster R-CNN using the open source code[7]. We only changed the scales and aspect ratios in RPN network. We used a uniform scale step of 1.3, allowing the anchor boxes to cover the image height. Instead of the default multiple aspect ratios, we used only one (width/height=0.41), which is consistent with our evaluation protocol. For training, we started with the VGG16 network pretrained on ImageNet and trained on our dataset for 100k iterations (LR = 10^{-3} for 60k and LR = 10^{-4} for another 40k).

Adapted Faster R-CNN [23] We followed the experimental findings from [23], and made corresponding modifications to vanilla Faster R-CNN for better performance. We started with the adapted Faster R-CNN model pretrained on the CityPersons dataset and then trained on our dataset for 100k iterations (LR = 10^{-3} for 60k and LR = 10^{-4} for another 40k).

RPN+BF [19] We followed the training procedure described by the authors, starting with the VGG16 network pretrained on ImageNet [13] and training the RPN network for 80k iterations (LR = 10^{-3}), followed by training the whole RPN+BF network for 80k iterations.

SDS-RCNN [1] Similarly to the previous method, we started with the VGG16 network pretrained on ImageNet and trained the RPN network for 120k iterations (LR = 10^{-3}), followed by 120k iterations for both RPN+BCN (full SDS-RCNN) network, using vanilla SGD.

Each method was trained on the *training* subset and evaluated on the validation subset (where available, otherwise the testing subset) of the dataset, keeping the training meta-parameters such as the learning rate or the number of epochs identical for the given method between different datasets. We however calculated mean image color and subtracted it as a preprocessing step for each dataset individually - this value was same for all the methods. We followed the standard

[6] `https://bitbucket.org/shanshanzhang/code_filteredchannelfeatures`
[7] `https://github.com/rbgirshick/py-faster-rcnn`

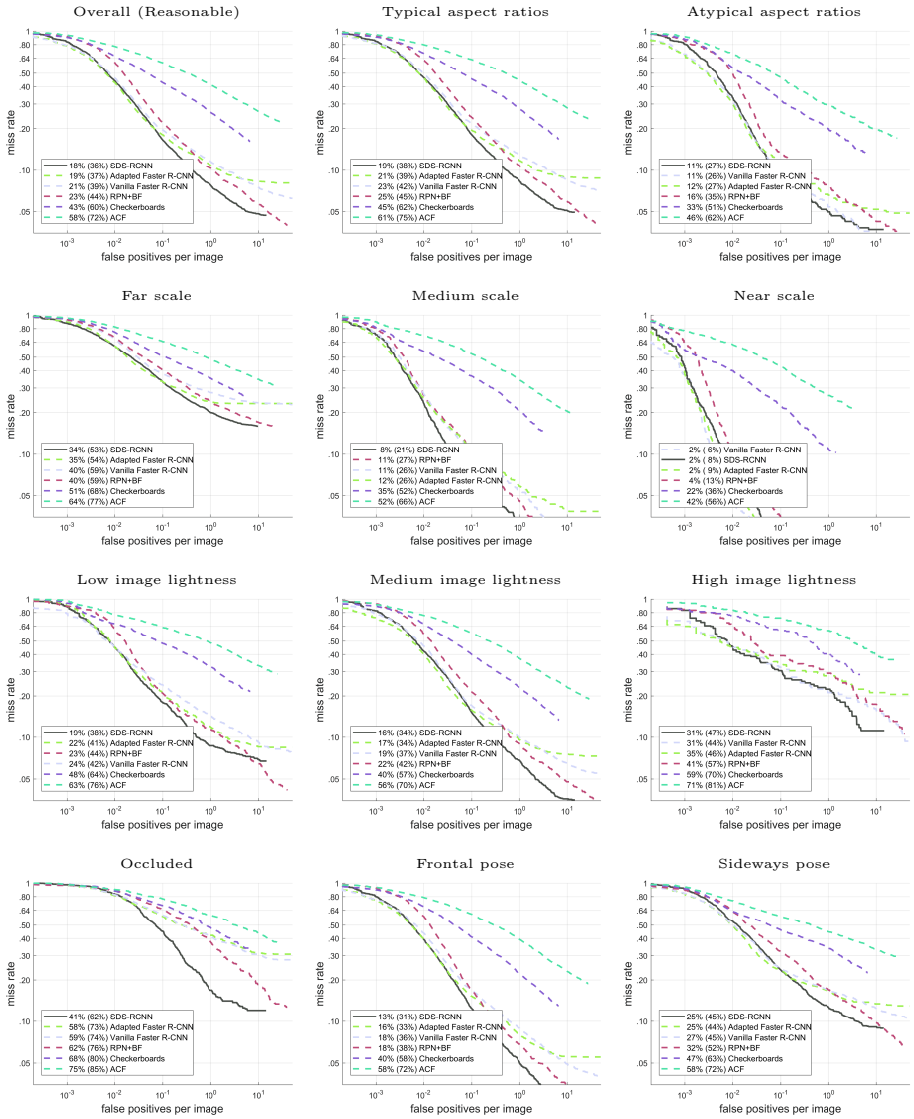

Fig. 4: Miss rates versus false positives of the recent pedestrian detection methods on the *NightOwls* dataset. Lower curve means better performance, the legend denotes average miss rate MR^{-2} [5] (MR^{-4} as in [22] in the parentheses)

average Miss Rate (MR) metric [4, 5] across all datasets, with the exception of the KITTI dataset, where the mean average precision (mAP) is typically used.

Comparison with Other Datasets. Using the Reasonable subset [4], the SDS-RCNN [1] detector, which is the state-of-the-art method on the Caltech

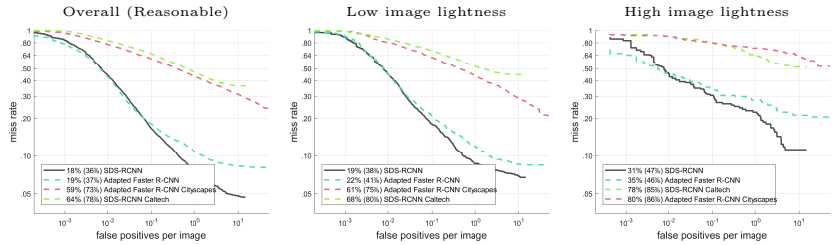

Fig. 5: Specifics of the night data. Comparing accuracy of methods trained on Caltech/CityPersons with methods trained directly on the *NightOwls* dataset

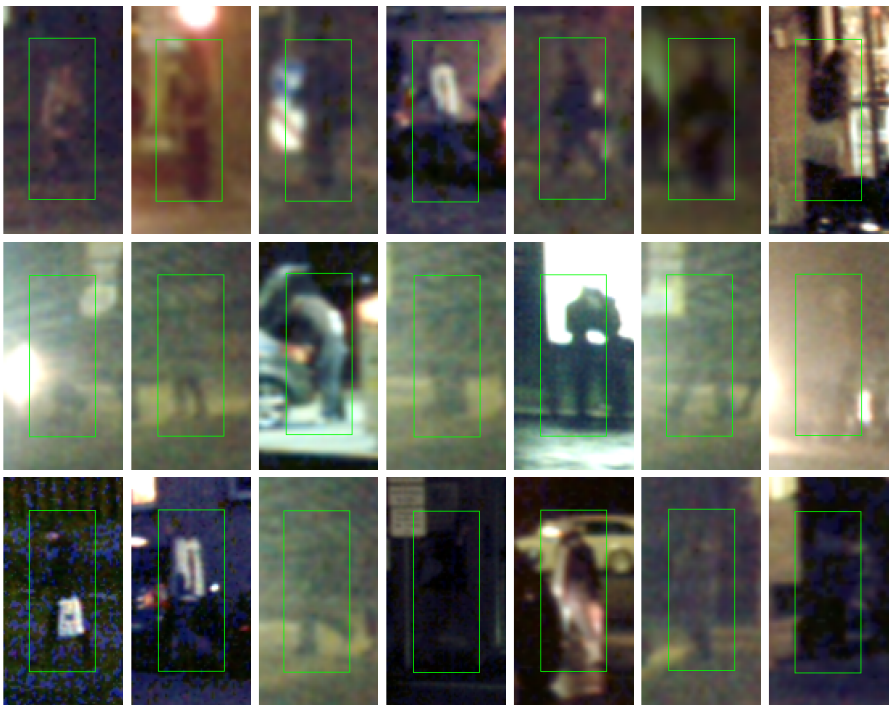

Fig. 6: Sample images of pedestrians missed by all the detection methods. *Far* scale (top row), *High image lightness* (middle row) and *Sideways pose* (bottom row)

dataset, also achieved the lowest average miss rate for our dataset (see Table 5), however the error is still 2.5times higher than for the Caltech and 50% higher than for the CityPersons dataset, which suggests that the proposed dataset is more challenging than the existing datasets. The gap in the miss rate between the vanilla Faster R-CNN and the improved version (SDS-RCNN) is also much smaller for our dataset, which suggests that the additional information brought

by the instance segmentation in SDS-RCNN [1] is not as helpful for night scenarios.

We also train the Adapted Faster R-CNN detector [23] on the KAIST dataset [12], which is the only existing dataset with some nighttime images, and we compare the accuracy on both datasets (see Table 6). We show that on the KAIST test set, the model trained on the *NightOwls* dataset actually outperforms the model trained on the KAIST training set, which is most likely due to the problems with KAIST image and annotations quality (see Section 2). Note that because KAIST does not have an occlusion flag, we did not use the flag for either of the datasets in the above experiment, to make to comparison fair.

Aspect Ratio & Scale. We evaluated the performance of all the methods trained on the *NightOwls* dataset, depending on different ground truth attributes, in line with the standard evaluation introduced by Dollar et al. [4]. We show that the methods are not as sensitive to aspect ratios (Fig. 4 - top row), but they are very sensitive to the size of pedestrians (Fig. 4 - 2$^{\text{nd}}$ row). The deep-learning methods clearly benefit from the amount of training data and for the *Medium* and *Near* scales their error rate is comparable to daylight datasets, however for the small pedestrians in the *Far* scale ($h < 90$px), the miss rate rises dramatically and the accuracy of deep-learning methods is close to the traditional ones (see Fig. 6 - top row for sample images).

Illumination. We also compare the performance based on the average image lightness, where the lightness is luma L[8] in the HSL colour space (Fig. 4 - 2$^{\text{3rd}}$ row). Perhaps counterintuitively, the error is higher for brighter images than the darker ones - this is caused by camera overexposure (see Fig. 6 - middle row), which makes the detection very challenging. Note that the evaluation based on pedestrian image patch lightness as opposed to whole image lightness and different lightness definitions give very similar results, hence we only include them in the supplementary material.

Pose. In contrast to the most commonly used datasets, we can also evaluate the detections based on the pedestrian pose - this clearly shows that all methods perform significantly better for people facing towards or away from the camera (*frontal pose*), than for pedestrians facing sideways (Fig. 4 - bottom row). We suggest this is due to higher ambiguity of the sideway pose, where there is a higher chance of confusion with other objects when a person is viewed from a side than from the front, but generally also due to lower number of pixels and therefore lower amount of information captured in the image for sideway poses (see Fig. 6 - bottom row).

Night Data Specifics. In order to evaluate how specific the night data is, we also run the state-of-the-art SDS-RCNN detector [1] trained on the Caltech dataset, which has a similar number of images, but it's exclusively captured in daytime. The model has an average miss rate of 7.36% on Caltech, but 63.99% on our dataset (the image mean subtracted as a pre-processing step was updated

[8] $L = 0.299R + 0.587G + 0.114B$

Table 6: Comparison of training and testing the Adapted Faster R-CNN detector on the KAIST-night and NightOwls datasets. The model trained on NightOwls performs better on both testing datasets. All numbers are MR on the Reasonable subset.

Test \ Train	KAIST-night	NightOwls
KAIST-night	65%	**63%**
NightOwls	57%	**19%**

accordingly to make sure the image data is always centered around zero). Similarly, the Adapted Faster R-CNN model [23] trained on the CityPersons dataset has a miss rate of 59.05% (see Fig. 5). These results confirm the expectation that pedestrian detectors trained on daytime data do not work well at night and training specifically on night data as in the previous sections is required.

5 Conclusion

In this paper, we have introduced a novel comprehensive pedestrian dataset *NightOwls* to encourage research on night images. Recent benchmarks for pedestrian detection and - in general, for object detection in computer vision - have predominantly focused on images collected at daytime. Even though detection at nighttime is a more challenging task because of low illumination, changing contrast, and less color information, studies on nighttime data are underrepresented, rely on study-specific data, and are limited to individual case studies lacking official benchmarks. We believe that by introducing a comprehensive dataset and benchmark for pedestrian detection at night, cutting-edge research on the challenges of nighttime vision can be stimulated.

References

1. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. In: ICCV 2017 (2017)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
3. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. PAMI (2014)
4. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR (June 2009)
5. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. PAMI **34** (2012)
6. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. PAMI (2009)
7. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR (2008)

8. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
9. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 3354–3361. IEEE (2012)
11. Huang, K., Wang, L., Tan, T., Maybank, S.: A real-time object detecting and tracking system for outdoor night surveillance. Pattern Recognition **41**(1), 432–444 (2008)
12. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1037–1045 (2015)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
15. Liu, X., Fujimura, K.: Pedestrian detection using stereo night vision. IEEE Transactions on Vehicular Technology **53**(6), 1657–1665 (2004)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
17. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: CVPR (2009)
18. Xu, F., Liu, X., Fujimura, K.: Pedestrian detection and tracking with night vision. IEEE Transactions on Intelligent Transportation Systems **6**(1), 63–71 (2005)
19. Zhang, L., Lin, L., Liang, X., He, K.: Is faster r-cnn doing well for pedestrian detection? In: European Conference on Computer Vision. pp. 443–457. Springer (2016)
20. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR (2016)
21. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: CVPR (2015)
22. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1259–1267 (2016)
23. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR (2017)