5th International Conference on Advanced Technologies
For Signal and Image Processing, ATSIP' 2020
September 02-05, 2020, Sfax, Tunisia

Deep CNN-based Pedestrian Detection for Intelligent Infrastructure

Bilel TARCHOUN[xi], Imen JEGHAM*, Anouar BEN KHALIFA†, Ihsen ALOUANI‡, Mohamed Ali MAHJOUB§
*Université de Sousse, Institut Supérieur d'Informatique et des Techniques de Communication de H. Sousse,
LATIS-Laboratory of Advanced Technology and Intelligent Systems, 4011, Sousse, Tunisie;
†[xi]§Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse,
LATIS-Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisie;
‡IEMN-DOAE, Université polytechnique Hauts-de-France, Valenciennes, France
Email: [xi]tarchoun.bilel@yahoo.com, *imen.jegham@isitc.u-sousse.tn, †anouar.benkhalifa@eniso.rnu.tn,
‡ihsen.alouani@uphf.fr, §mohamedali.mahjoub@eniso.rnu.tn

***Abstract*—Autonomous driving systems and driver assistance systems are becoming the center of attention in transport technology. Given its safety criticality, pedestrian detection is a highly important task. Transport oriented intelligent systems use embedded sensors for the detection task. However, vehicle side detection is starting to show its limitations especially when dealing with certain challenges such as occlusions. In this paper, we propose an infrastructure side perception system that has a bird's eye view. We introduce a new deep pedestrian detector that can use the detection results to warn nearby vehicles of the presence of pedestrians on the road. The results show that our proposed system is able to detect pedestrians in most conditions with 70.41% precision and 69.17% recall.**

***Index Terms*—Pedestrian detection, transfer learning, Faster R-CNN, intelligent infrastructure, Intelligent transportation systems**

I. Introduction

Today, the trends in transportation technology are focused on developing Intelligent Transportations Systems (ITS) that are able to communicate with each other and with intelligent infrastructure elements [1]. This communication is made possible thanks to the integration of sensors in vehicles and infrastructure elements and the development of vehicular communications [2]. These Infrastructure to Vehicle (I2V) communications are used to send sensor data to other vehicles, and are an integral part of establishing intelligent transportation systems [3]. Some applications that use vehicular communications have been proposed in the context of intelligent transportation, such as intersection safety [4], [5], traffic lights and signs detection [6], visibility estimation [7], [8], road monitoring [3], [9], [10], pedestrian and cyclist detection [11], [12], etc. In this context, we focus in particular on pedestrian detection, which is a cornerstone of any ITS. Pedestrian detection is the subject of a large domain of research with many real-world applications such as security, safety and robotics [13]–[16]. The interest in this domain has further grown in recent years thanks to the rise of driver assistance systems and developments in computer vision [17], [18].

In the context of pedestrian detection on the road, a large amount of existing pedestrian detection systems focus on the vehicle view [19]–[21]. While these systems have obtained good results, they are reaching their limits because of challenges such as illumination problems, complex backgrounds, the presence of shadows, variations of pedestrian appearances and especially occlusions [22]. These limitations coupled with the increasing availability of infrastructure side cameras may lead to an alternative way to detect pedestrians that may be difficult to detect from the vehicle view. The biggest advantage of infrastructure side cameras is their bird's eye view combined with the larger fields of view, giving a much more comprehensive perspective on the situation.

In parallel, deep learning based detectors have been extremely popular in the last few years, with many convolutional neural network based detectors developed for pedestrian detection applications achieving substantial results over handcrafted feature based detectors [13], [23].

In this paper, we propose to exploit infrastructure side cameras pedestrian detection on the road. These detection results can then be exploited for a multitude of applications, such as warning nearby vehicles of the presence of potentially hidden pedestrians. The proposed method is based on Faster R-CNN with transfer learning based on various pre-trained architectures. Our detector will be designed for use on images acquired from an infrastructure side camera, exploiting the advantages brought by these cameras.

The main contributions of this paper are the following:

- We propose a deep pedestrian detector designed based on infrastructure side cameras that provide a bird's eye view.
- We give the main challenges encountered in pedestrian detection and some proposed methods to handle these challenges.

The paper is organized as follows. Section 2 introduces the common challenges encountered in the task of pedestrian detection and some related works in solving these challenges. Section 3 presents our proposed pedestrian detection method. In section 4, we evaluate our detector's performance and discuss the results. And finally, section 5 concludes the paper.

978-1-7281-7513-3/20/$31.00 ©2020 IEEE

II. CHALLENGES AND RELATED WORK

Pedestrian detection constitutes a significant part of research in the area of computer vision. In the process of studying these detectors, researchers encounter many challenges. In this section, we discuss pedestrian detection challenges as well as some proposed approaches in handling these challenges.

- **Illumination problems:** Bad lighting conditions can cause many problems in pedestrian detection applications. Pedestrians are very hard to detect in a dark environment, and a detector can be affected by a light source shining directly on the camera (Fig.1). This problem is accentuated in an uncontrolled environment, since lighting conditions are always changing due to time of day, weather or temporary occlusions. Xu *et al.* [24] propose to train a deep convolutional network on RGB and infrared images to learn features robust to bad illumination conditions, and then apply the trained detector on RGB images, which improves detection rates in images with bad illumination conditions.

Fig. 1. Illustration of illumination problems [25].

- **Appearance variations:** Pedestrians generally have different shapes since humans have extremely varied body types and the view angle and distance from the camera can cause significant variations in pedestrian appearances as shown in Fig.2 Pedestrians also wear different clothes or accessories and stand in different poses, contributing to these variations. To improve detection accuracy for pedestrian that are far from the camera, Zhang *et al.* [26] introduce a scale aware localization policy to an R-CNN that uses features extracted from multiple levels in the R-CNN as well as the initial region proposals and combines them with an active detector that uses spatial and temporal contextual information to perform a series of coordinate transformations which lead to the choice of the proper feature extraction layer for detection.

Fig. 2. Illustration of pedestrian appearance variations [25].

- **Complex backgrounds:** In a real world scenario, pedestrian detection happens on an uncontrolled background where many static objects are present and dynamic objects may appear and disappear (Fig.3). The presence of vertical objects such as trees and poles can be the cause false positives. In [27], the authors handle these false negatives in single stage object detectors by adding multiple classification layers that divide the input image into a grid and calculate the likeliness of each grid element to be a part of a pedestrian. These scores are used to adjust final detection scores of each detection candidate.

Fig. 3. Samples of complex backgrounds [25].

- **Shadows:** Shadows are a problem in pedestrian detection since in the right lighting conditions, shadows can have a similar shape to a pedestrian as shown in Fig. 4, which can lead to an increased amount of false positives. In [28], the authors propose to detect pedestrians by combining background subtraction with improved HOG features, to deal with the problem of shadows, the input image is converted into the HSI color space and the disparity between the pixel hues and intensities and the background is calculated to detect shadow areas and remove them from the ROIs.

Fig. 4. Pedestrian shadows [25].

- **Non-rigid deformations:** Non rigid deformations are when a part of an object moves independently of the rest of the object as shown in Fig. 5. An example of this is when a pedestrian bends down to pick up an object of tie his shoelaces. To deal with this challenge, Ouyang *et al.* [29] add a deformation layer in their CNN. This deformation layer contains a deformable parts model that is integrated into the process of training the CNN, thus improving its accuracy. This addition makes the proposed

CNN able to learn discriminative features for each part of the body, therefore making the network able to handle body deformations.

Fig. 5. Examples of non-rigid deformations [25].

- **Occlusions:** Occlusions are the presence of an object or another person that hides the detection target in part or in full (Fig. 6). Occlusions are a major problem in pedestrian detection applications especially in cluttered environments since they often hide important information used by the detectors. In [27], Noh *et al.* propose to augment single stage object detectors (such as YOLOv2) with a module that divides each bounding box into 6 × 3 sub boxes and associate a score to each sub-box. These scores are then used to refine the initial detection score.

Fig. 6. Occluded pedestrians [25].

Some of these challenges cannot be fully overcome by only using data from the vehicle mounted camera, especially in the case of fully occluded pedestrians since there is no data about these pedestrians to exploit. The use of an infrastructure side camera can be an alternative to vehicle-side pedestrian detection with its larger field of view and the results can be relayed to nearby vehicles. This will be the basis of our proposed approach detailed in the next section.

III. PROPOSED METHOD

The motivation of this paper is to propose a pedestrian detection system that performs well in real world conditions and can be applied in an I2V context. We use a roadside unit that contains a camera, a processing unit and wireless communication equipment that can warn nearby vehicles of the presence of potentially hidden pedestrians. Fig. 7 presents a sample application of our method. A passenger crossing the street is occluded by a tree in the vehicle driver's view, but the infrastructure side camera can clearly see him.

Our proposed method is based on Faster R-CNN and composed of two steps as show in Fig. 8. In the first step, we train

Fig. 7. Example of a situation encountered in our context.

our detector using a transfer learning approach. In the next step we perform pedestrian detection on the infrastructure's camera images using our trained detector. We divide our database in two sets: the training (1) and validation (2) databases, Where F_i is the image number *i* in the database and $GT_i = \{gt_k\}_{k=1}^{|GT_i|}$ is the set of ground truth bounding box coordinates in the image F_i with $gt_k = [x_k^{gt}, y_k^{gt}, w_k^{gt}, h_k^{gt}] \in \mathbb{R}^4$ denoting the top left corner, width and height of the ground truth bounding box respectively. :

$$D_{train} = \{(F, GT)_i\}_{i=1}^{|D_{train}|} \tag{1}$$

$$D_{test} = \{(F, GT)_i\}_{i=1}^{|D_{test}|} \tag{2}$$

A. Transfer Learning

In our method, we adopt a Faster R-CNN architecture as the basis of our detector. This choice is motivated by the good performance to speed compromise found in these detectors [30]. To train our detector, we choose to adapt an existing pre-trained network to our application by using transfer learning techniques. This choice is justified by the advantages of transfer learning. We can exploit the strength of a detector trained on a large database while only requiring a smaller amount of training data and training time in order to learn the specific characteristics of our application.

Fine-tuning a Faster R-CNN architecture is done in a four step process [31] using the D_{train} database as training data:

- The region proposal network (RPN) of the faster R-CNN is re-trained with the CNN feature extraction layer weights locked.
- With the new RPN weights, re-train the CNN feature extraction layers
- Fine-tune the RPN using the new weights of the CNN feature extraction layers. re-trained in step 2 (These weights are kept fixed).
- Fine-tune the feature extraction layers using the final RPN weights obtained from the previous step.

After these steps, our pedestrian detector is ready for the pedestrian detection task.

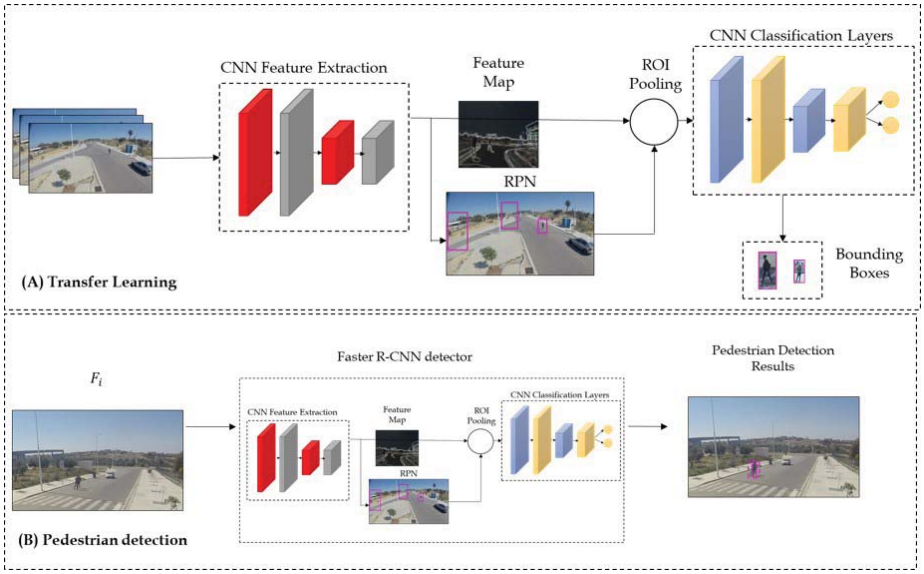

Fig. 8. Illustration of the proposed method.

B. Pedestrian Detection

The infrastructure side camera will capture an image F_t at a time t. This frame is then transferred to the processing unit which applies our detector trained in the previous step to detect pedestrians. The Faster R-CNN detector executes two tasks in order to find the pedestrians within the image: The feature extraction layers generate the convolutional feature maps of the input image, and the region proposal network suggests regions of interest (ROI) where pedestrians are potentially present. These ROIs are proposed around anchor points in the image according to different scales and aspect ratios chosen by the RPN. The results of these two tasks (the feature maps and the ROIs) are finally pooled together to obtain the final detection results.

After applying our trained detector to the image F_t, we obtain a set of detected pedestrian bounding boxes $BBox_t = \{[x_i, y_i, w_i, h_i]\}_{i=1}^{|BBox_t|}$. These detection results can be exploited for many applications such as warning vehicles of potentially hidden pedestrians.

IV. EXPERIMENTS AND RESULTS

The experiments with the proposed model have been evaluated on a computer with an Intel i7-7700 HQ processor, 16 GB of RAM, and an NVIDIA GTX 1050Ti graphics card with 4 GB of VRAM. We use the I2V-MVPD database, the images are captured by a camera mounted on top of a light pole providing a bird's eye view of the scene. This dataset contains 4740 images divided into sixty-one videos of pedestrians and vehicles interacting on the road. Twenty video were used as training data, and the rest of the videos were used for validation. The videos depict a variety of scenarios such as one or multiple pedestrians crossing the road. The database contains also the common challenges found in pedestrian detection [25].

A. Transfer learning parameters

In our experiments, we evaluate five different pre-trained CNN architectures: Mobilenetv2, Googlenet, Resnet-18 VGG-19 and VGG-16. These models were initially pre-trained on the ImageNet database which is a large database with one thousand classes. To adapt these architectures to our application that needs two outputs (absence or presence of a person), we modify the classification layer of these architectures and replace the previous layers with new layers that fit our new classification layer. With these modifications, the architectures are ready to be fine-tuned. We use the training dataset combined with the following hyperparameters (Table I) for the training step.

B. Results and discussion

After training the detectors, we can proceed to the evaluation step using the validation dataset. We evaluate the detectors using four different metrics : Precision, Recall, Average Precision and detection time. The results are shown in Table II.

TABLE I
TRAINING PARAMETERS

Hyperparameter	Value
Number of epochs	15
Mini Batch Size	2
Learning Rate	10^{-3}
Number of Strongest Regions	1500
Number of Regions To Sample	64
Momentum	0.9

Our results show that the detector based on the Resnet-18 architecture has the best overall performance. However, we notice that detection time for all evaluated architectures is

TABLE II
DETECTION RESULTS

Architecture	Average Precision (%)	Precision (%)	Recall (%)	Time (s)
Mobilenetv2	55.92	70.91	67.62	0.485
Googlenet	41.01	79.26	46.12	0.500
VGG-16	54.63	45.27	62.73	0.452
VGG-19	41.44	61.38	50.19	0.577
Resnet-18	**60.88**	**70.42**	**69.17**	**0.411**

high, with approximately 2 frames per second, which is still far from real-time requirements.

While our detector performs well in most situations (Fig. 9), the detection process fails under certain conditions. For example, in large crowds some severely occluded pedestrians may not be correctly detected (Fig 10.a). Certain view angles may cause detection failure, such as the angle shown in Fig. 10.b. Finally, our proposed detector may not be able to detect pedestrians who are far from the camera as shown in Fig. 10.c.

Fig. 9. Sample of detection results.

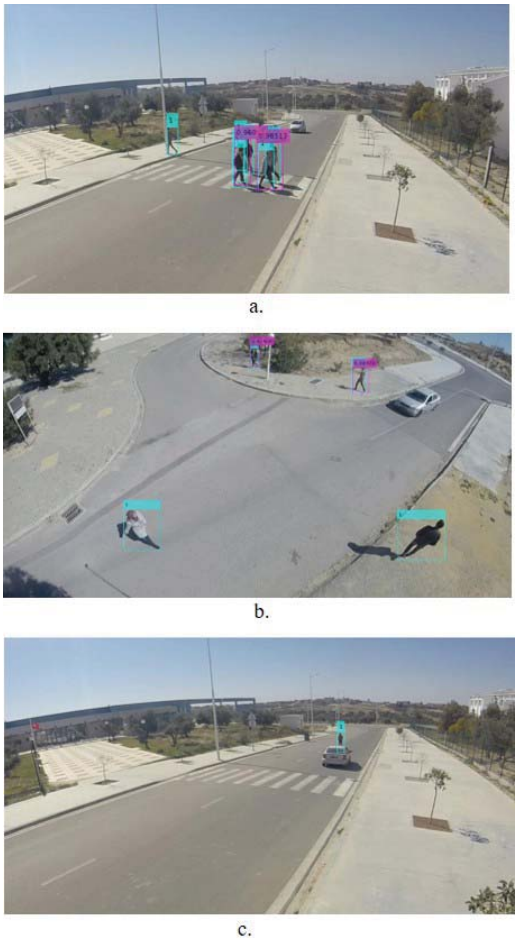

Fig. 10. Example of the limitations of the proposed method (a) Severe occlusions in crowds (b) View angle (c) Distant pedestrians.

V. CONCLUSION

In this paper, we propose a pedestrian detection system based on a camera mounted on the road infrastructure providing a bird's eye view. The presented system is based on a new deep pedestrian detector trained using transfer learning. Our proposed detector has proven to perform well in real world conditions. This work can be applied to support the development of cooperative ITS using I2V communications through many applications.

Future extensions of this work can be considered, such as combining this system with vehicle side pedestrian detection in order to create an intelligent and collaborative road side system.

REFERENCES

[1] C. Olaverri-Monreal, G. C. Krizek, F. Michaeler, R. Lorenz, and M. Pichler, "Collaborative approach for a safe driving distance using stereoscopic image processing," *Future Generation Computer Systems*, vol. 95, pp. 880 – 889, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17307124

[2] L. A. Garcia and V. R. Tomas, "A smart peri-urban i2v architecture for dynamic rerouting," *IEEE Intelligent Transportation Systems Magazine*, vol. 10, no. 2, pp. 69–79, Summer 2018.

[3] F. Jiménez, J. E. Naranjo, J. J. Anaya, F. Garcia, A. Ponz, and J. M. Armingol, "Advanced driver assistance system for road environments to improve safety and efficiency," *Transportation Research Procedia*, vol. 14, pp. 2245 – 2254, 2016, transport Research Arena TRA2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352146516302460

[4] S. Vishnu, U. Ramanadhan, N. Vasudevan, and A. Ramachandran, "Vehicular collision avoidance using video processing and vehicle-to-infrastructure communication," in *2015 International Conference on Connected Vehicles and Expo (ICCVE)*, Oct 2015, pp. 387–388.

[5] M. Goldhammer, E. Strigel, D. Meissner, U. Brunsmann, K. Doll, and K. Dietmayer, "Cooperative multi sensor network for traffic safety applications at intersections," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*, Sep. 2012, pp. 1178–1183.

[6] M. A. García-Garrido, M. Ocana, D. F. Llorca, E. Arroyo, J. Pozuelo, and M. Gavilán, "Complete vision-based traffic sign recognition supported by an i2v communication system," *Sensors*, vol. 12, no. 2, pp. 1148–1169, 2012.

[7] N. Hautiere and A. Boubezoul, "Combination of roadside and in-vehicle sensors for extensive visibility range monitoring," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, Sep. 2009, pp. 388–393.

[8] H. Chaabani, F. Kamoun, H. Bargaoui, F. Outay, and A.-U.-H. Yasar, "A neural network approach to visibility range estimation under foggy weather conditions," *Procedia Computer Science*, vol. 113, pp. 466 – 471, 2017, the 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN

2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917317131

[9] S. A. Taie and S. Taha, "A novel secured traffic monitoring system for vanet," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, March 2017, pp. 176–182.

[10] X. Zhao, K. Mu, F. Hui, and C. Prehofer, "A cooperative vehicle-infrastructure based urban driving environment perception method using a d-s theory-based credibility map," *Optik*, vol. 138, pp. 407 – 415, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0030402617303686

[11] G. Al-refai, M. Horani, and O. A. Rawashdeh, "A framework for background modeling using vehicle-to-infrastructure communication for improved candidate generation in pedestrian detection," in *2018 IEEE International Conference on Electro/Information Technology (EIT)*, May 2018, pp. 0729–0735.

[12] J. Zhao, H. Xu, J. Wu, Y. Zheng, and H. Liu, "Trajectory tracking and prediction of pedestrian's crossing intention using roadside lidar," *IET Intelligent Transport Systems*, vol. 13, no. 5, pp. 789–795, 2018.

[13] A. B. Khalifa, I. Alouani, M. A. Mahjoub, and N. E. B. Amara, "Pedestrian detection using a moving camera: A novel framework for foreground detection," *Cognitive Systems Research*, vol. 60, pp. 77 – 96, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389041719305212

[14] I. Jegham, A. Ben Khalifa, I. Alouani, and M. A. Mahjoub, "Mdad: A multimodal and multiview in-vehicle driver action dataset," in *Computer Analysis of Images and Patterns*. Cham: Springer International Publishing, 2019, pp. 518–529.

[15] A. Mimouna, A. B. Khalifa, and N. E. B. Amara, "Human action recognition using triaxial accelerometer data: selective approach," in *2018 15th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 2018, pp. 491–496.

[16] I. JEGHAM, A. BEN KHALIFA, I. ALOUANI, and M. A. MAHJOUB, "Safe driving : Driver action recognition using surf keypoints," in *2018 30th International Conference on Microelectronics (ICM)*, Dec 2018, pp. 60–63.

[17] P. Kaur and R. Sobti, "Current challenges in modelling advanced driver assistance systems: Future trends and advancements," in *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, Sep. 2017, pp. 236–240.

[18] W. Lejmi, M. A. Mahjoub, and A. Ben Khalifa, "Event detection in video sequences: Challenges and perspectives," in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, July 2017, pp. 682–690.

[19] I. Jegham and A. Ben Khalifa, "Pedestrian detection in poor weather conditions using moving camera," in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, Oct 2017, pp. 358–362.

[20] K. Chebli and A. B. Khalifa, "Pedestrian detection based on background compensation with block-matching algorithm," in *2018 15th International Multi-Conference on Systems, Signals Devices (SSD)*, March 2018, pp. 497–501.

[21] S. Dhifalah, "A novel framework for foreground estimation: Application to pedestrian detection using a moving camera," Ph.D. dissertation, University of Sousse, 2018.

[22] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub, "Vision-based human action recognition: An overview and real world challenges," *Forensic Science International: Digital Investigation*, vol. 32, p. 200901, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S174228761930283X

[23] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[25] A. B. KHALIFA, "I2v-mvpd," 2019, last accessed 16/01/2020. [Online]. Available: https://sites.google.com/site/benkhalifaanouar1/6-datasets

[26] X. Zhang, L. Cheng, B. Li, and H. Hu, "Too far to see? not really!—pedestrian detection with scale-aware localization policy," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3703–3715, Aug 2018.

[27] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] S. A. Chowdhury, M. M. S. Kowsar, and K. Deb, "Human detection utilizing adaptive background mixture models and improved histogram of oriented gradients," *ICT Express*, vol. 4, no. 4, pp. 216 – 220, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405959517301613

[29] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1874–1887, Aug 2018.

[30] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf