2021 Fifth IEEE International Conference on Robotic Computing (IRC) | 978-1-6654-3416-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/IRC52146.2021.00024

A Computationally Effective Pedestrian Detection using Constrained Fusion with Body Parts for Autonomous Driving

Muhammad Mobaidul Islam, Abdullah Al Redwan Newaz, Renran Tian, Abdollah Homaifar, and Ali Karimoddini*

***Abstract*—This paper addresses the problem of detecting pedestrians using an enhanced object detection method. In particular, the paper considers the occluded pedestrian detection problem in autonomous driving scenarios where the balance of performance between accuracy and speed is crucial. Existing works focus on learning representations of unique persons independent of body parts semantics. To achieve a real-time performance along with robust detection, we introduce a body parts based pedestrian detection architecture where body parts are fused through a computationally effective constraint optimization technique. We demonstrate that our method significantly improves detection accuracy while adding negligible runtime overhead. We evaluate our method using a real-world dataset. Experimental results show that the proposed method outperforms existing pedestrian detection methods.**

I. INTRODUCTION

Throughout the last decade, we have seen notable progress in vision-based pedestrian detection using deep learning techniques. Real-time accurate pedestrian detection is a key factor to ensure the safe operation of autonomous vehicles [1]. However, accurate and robust pedestrian detection in autonomous driving is a notoriously hard task. Pedestrians may appear in an image with different articulations of body parts and various poses. Besides, different illumination levels in the environment and various sizes and aspect ratios of pedestrians in an image make it challenging to detect pedestrians accurately. In particular, partial occlusion of pedestrians in urban settings is a major challenge for detecting pedestrians.

A common approach to handle the occlusion in complex scenarios is either to use a separate detection module delegated to occluded pedestrians [2], [3] or to use a multi-shot image detector that has an occlusion-aware region of interest [4]. Besides, some efforts have been made to improve the pedestrian detection performances using body parts in recent works [5], [6]. Most of these methods struggle to balance the tradeoff between accuracy and speed when adopted for autonomous driving applications.

In the cases where a generic object detection model is used to detect pedestrians, the decisions are made based on the full-body features of pedestrians. However, in many scenarios, the full-body features of pedestrians may not be available. Consider the sudden pedestrian crossing scenario in Fig. 1, which has always been a safety issue for autonomous driving. We argue that an autonomous vehicle can avoid such vehicle-pedestrian crashes if it could detect occluded pedestrians in real time. To do that, given an image frame or a video stream, a pedestrian detection method needs to identify which of a known set of body parts might be present and their corresponding positions within the image frame. If we train a generic object detector on the body parts dataset, it can detect body parts but fails to establish a semantic relationship among body parts. Hence even if body parts are detected, a generic object detector cannot infer pedestrians from the body parts directly, e.g., unable to separate multiple pedestrians in a partially occluded scene.

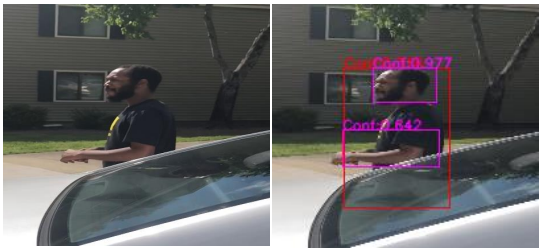

Fig. 1: The left side result shows the performance of the conventional pedestrian detector that is unable to detect a partially occluded pedestrian due to the lack of full-body features. On the right side, results show that our proposed BP-SSD can detect the same pedestrian utilizing the semantic relationship of their body parts.

In this paper, we develop an improved pedestrian detector modifying the single-shot multi-box object detector [7]. Unlike other object detection methods, the Single Shot Detection (SSD) method takes only one single shot to detect multiple objects in the image frame [7], resulting in better runtime efficiency. Our proposed has body part-based single-shot detection (BP-SSD) architecture that can detect pedestrians efficiently and robustly. To achieve this, we first consider different body parts as separate object classes and train our BP-SSD with body parts and full-body labels. Then, we propose a constraint optimizer to effectively select bounding boxes for pedestrians. In the cases, where some body parts are detected at high confidence while full-body bounding boxes are detected at low confidence, the constraint optimizer finds the best bounding boxes for the pedestrians utilizing the semantic relationship of body parts information.

This work is supported in part by the North Carolina Department of Transportation awards RP2019-28, RP2022-063, and TCE2020-03, the National Science Foundation awards 2018879 and 2000320, and by Ford Motors.

M. M. Islam, A. A. Redwan Newaz, A. Homaifar, and A. Karimoddini are with the Department of Electrical and Computer Engineering, North Carolina Agricultural and Technical State University, Greensboro, NC 27411 USA. R. Tian is with Department of Computer Information & Graphics Technology, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202.

*Corresponding author: Ali Karimoddini akarimod@ncat.edu.

978-1-6654-3416-4/21/$31.00 ©2021 IEEE
DOI 10.1109/IRC52146.2021.00024

In summary, the contributions of this work are as follows: 1) *we develop BP-SSD that can robustly detect body parts along with full-body pedestrians,* 2) *we propose a constraint optimization technique that can utilize semantic relationship of body parts to detect pedestrians accurately,* In our earlier work [8], we showed how to detect pedestrians by classifying body parts. However, we did not specify body parts concerning a pedestrian. Here, we established a semantic relationship of body parts for a particular pedestrian.

II. Related work

Pedestrian detection from images is an important yet difficult task since pedestrians may appear with different sizes, deformed shapes, and partial occlusion. In the literature, we find a variety of solutions to improve pedestrian detection utilizing both generic deep learning-based methods and methods highlighting pedestrian detection in case of partial occlusion. In the literature, we observe that most of the deep learning-based pedestrian detection methods are based on either two stage detector [9], [10] or single stage detector [11]–[13]. Deep learning-based methods enjoy automatic feature extractions and achieve superior results when trained with a sufficient amount of labeled data. Canonically, the pedestrian detection problem is viewed as an object detection problem where pedestrians are treated as one class of data [9], [12], [13]. To improve the detection accuracy, recent works focus on fusing semantic segmentation information to model the pedestrian detection problem as a joint estimation problem. In [10], the semantic segmentation information is fused for joint supervision to improve the detection accuracy. In [11], a single-stage detector is fused with a semantic segmentation network for integrating pixel-wise semantic map information. The fundamental improvement in a single-stage detector comes from eliminating bounding box proposals and the feature resampling stage. Out of many single-stage detectors the single-shot multibox detector (SSD) [7] offers superior detection performances with a competitive runtime efficiency. Therefore, in this paper, we adopt the SSD for implementation purposes. Although SSD has good run-time efficiency, similar to most other single-stage detectors, its accuracy is lower than two-stage detectors. Therefore, in this work, we focus on improving the detection accuracy of the SSD while achieving a competitive runtime efficiency.

Another aspect of pedestrian detection problem is to handle partial occlusion where a pedestrian is blocked by other pedestrians or different objects. To handle the partially occlusion problem in pedestrian detection, body part-based models are explored in literature [5], [6], [14], [15]. A mixture-of-expert framework is developed to detect the partially occluded pedestrians using body parts components like head, torso, and leg in [14]. By constructing a body parts pool, a pedestrian detection model is proposed in [15]. To recall the lost body parts and the shifting problem, a part-level CNN using saliency and boundary box alignment is proposed in [6]. A probabilistic framework is introduced in [3] where a deformable part-based model of humans is used to compute the pedestrian visibility estimation. Apart from the part-based modeling, an attention network along with a two-state object detector [2], and a mask-guided attention network with a two-state object detector [16] are studied to handle the occlusion problem of pedestrian detection. Even though these methods can handle partial occlusion, these methods usually suffer from the high computation cost compared to single-stage detectors.

In this paper, we adopt the SSD architecture for generating pedestrian candidate boxes along with the body parts labels. We then utilize the body parts semantic relationship to handle complex scenarios such as partial occlusion, motion blur, and unusual body articulations. Finally, we propose a novel optimization technique to select pedestrian candidates subject to the body parts semantics.

III. Preliminaries

We develop a novel pedestrian detection method based on Single Shot MultiBox Detector (SSD) [7], which combines the classification and localization tasks in a single forward pass of the network.

SSD uses smooth L1-Norm, denoted by smooth_{L1} in eqn. (1), to calculate the localization loss, resulting in effective and robust bounding box prediction. Let x_{ij}^p be the binary decision variable for indicating the i^{th} default box from the p^{th} category matches to the j^{th} ground truth box. In our setup, p represents either full body or one of body parts. Let m be the default bounding box parameterized by center coordinates (cx, cy), width w, and height h such that $m \in \{cx, cy, w, h\}$. Also let Pos and Neg be the sets of default box indices that represent positive examples and negative examples, respectively. Comparing the difference between the predicted bounding box and the ground truth box, the localization loss of SSD is computed as follows:

$$\mathcal{L}_{loc} = \sum_{i \in Pos}^{N} \sum_{m} x_{ij}^p \text{smooth}_{L1}(\ell_i^m - \hat{g}_j^m) \tag{1}$$

where $\hat{g}_j^m$ represents the regression loss between the j^{th} ground truth box and the m^{th} parameterized default box, and ℓ_i^m represents the i^{th} predicted bounding box with parameter m. On the other hand, for each predicted bounding box, a set of class predictions are computed utilizing the base network, for every possible class in the dataset. Based on this information, the classification loss can be calculated as:

$$\begin{gathered} \mathcal{L}_{cla} = -\sum_{i \in Pos}^{N} x_{ij}^p log(\hat{c}_i^p) - \sum_{i \in Neg} \log(\hat{c}_i^p) \\ \text{where } \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)} \end{gathered} \tag{2}$$

where $\hat{c}_i^p$ represents the categorical loss from i^{th} default box of the p^{th} category. Having the information about the bounding boxes and the scores as the outputs of SSD, a non-maximum suppression step is used as a post-processing step to produce the final detection, which helps to identify the highest probable object within a single bounding box.

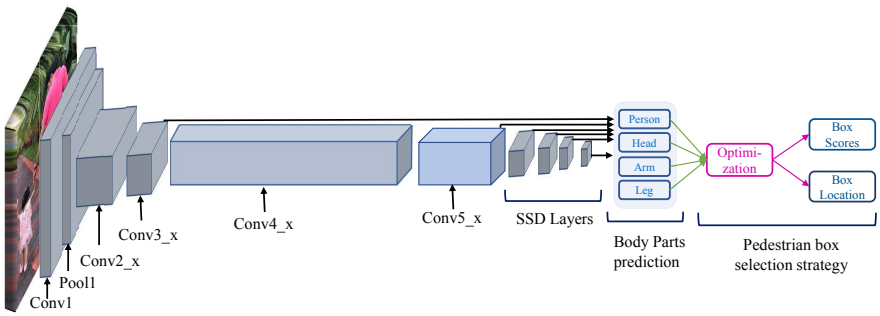

Fig. 2: Architecture of the BP-SSD network. We augment the BP-SSD network with a constraint optimizer to detect pedestrians accurately as well as robustly.

IV. Proposed Methodology

In this section, we will explain our proposed methodology. Although our framework can use any single-stage detector, we adopt the SSD [7] for this work. At the high level, we augment an original SSD network architecture along with body parts semantic and a constraint optimization strategy to detect pedestrians accurately. Unlike the original SSD, we are interested in a limited amount of classes to efficiently and robustly detect pedestrians. Fig. 2 shows our proposed BP-SSD architecture. We modified the classification layer of the SSD network for body parts proposal generation. Further, after the classification layer, an optimization module is added to the network to select the pedestrian candidates while considering the body part's semantic relationship.

A. Body Parts Semantics

We develop a constraint optimization solver that utilizes the semantic relationship between detected body parts and normal structural descriptions of human bodies to predict the number of pedestrians for an input image. To formally describe the proposed method, consider a detected body part b from a set of body parts $\mathbf{B}$ whose confidence score c_b. We define an indicator function I whose value for the detected body part b is 1 if c_b is greater than or equal to a given threshold value θ.

Here, we use the indicator function I to establish the semantic relationship from detected body parts. For instance, if our BP-SSD detects 2 heads, 3 legs, and 1 arm in an image with confidence scores equal or above a threshold value, we can then expect that there exist at least 2 pedestrians in this scene because 2 heads are detected with high confidence. To capture this idea, given N number of bounding boxes for b part, we first define the expected b part per person as e_b, which is a constant number. Then, we compute a rounding ratio, r_b, between the detected body parts and expected b part per person as $r_b = ceil\left(\frac{\sum_{i=1}^{N} I_{\mathbf{B}}(b_i)}{e_b}\right)$, where $ceil$ function rounds a number up to the next largest integer.

To infer the expected number of person boxes, W, for an input image, we can simply take the maximum value over all possible values of r_b as $W = \max\left(\{r_b | b \in \mathbf{B}\}\right)$. Here, we expect the number of pedestrians to be the same as the ground truth by establishing the semantic relationship from detected body parts. One important observation is that in challenging conditions (i.e., partial occlusion, different body articulation, complex background, etc.), many accurate bounding box proposals of BP-SSD for the person class are ignored due to the lack of confidence scores on full-body features. If we boost the confidence score of these proposals utilizing the semantic relationship from body parts, we can reduce the number of miss detection as well. In the next subsection, we will elaborate our explanation on the selection of multiple bounding boxes based on body parts constraints.

B. Bounding Box Selection Strategy

To detect a pedestrian with high confidence, the SSD needs to predict over the full-body features. However, in many cases, the full-body feature is not available due to partial occlusions or different body articulations. In such cases, the SSD fails to detect pedestrians as their corresponding confidence scores are low. To overcome this problem, given a set of default boxes for the person class, each with a confidence score and a weight, we determine the number of boxes includes in a scene so that the total weight is less than or equal to a given limit and the total confidence score is as large as possible. Formally, given a set of N default boxes numbered from 1 to N for an image, each with a weight w_i and a confidence score c_i, along with a maximum weight capacity of W, we formulate this problem as a constrained optimization problem as follows:

$$\max \sum_{i=1}^{N} c_i x_i, \text{ s.t. } \sum_{i=1}^{N} w_i x_i \leq W, \; x_i \in \{0,1\} \qquad (3)$$

where x_i represents the decision variable to be chosen optimally for selecting bounding boxes of pedestrians. In this setting, $x_i = 1$ represents the case that the i^{th} bounding box of proposal default boxes is chosen, and conversely, $x_i = 0$ represents the case that the i^{th} bounding box of proposal default boxes is not chosen. While it is straightforward to obtain confidence scores for bounding boxes directly from the predictions of SSD, the weight of each bounding box needs to be carefully defined so that it can lead to accurate detection.

In general, w_i is between 0 to 1. Here, for simplicity and reducing the search space, we choose $w_i = 1$ which means that each pedestrian can belong only to one bounding

Model	FP	TP	FN	Accuracy	Precision	Recall
SSD	12	105	75	0.546	0.897	0.583
BP-SSD	12	117	63	0.609	0.900	0.650

TABLE I: Evaluation Metrics

box at a time. The maximum weight capacity W can be determined from the body parts semantics. Alternatively, we can set the weight of pedestrians' bounding boxes based on the Generalized Intersection over Union (GIoU). GIoU provides a metric to evaluate how close the pedestrians' bounding boxes are to the body parts' bounding boxes. Thus, a pedestrian's bounding box which is close to the detected body parts' boxes has a lower weight compared to the far away bounding boxes. However, the computation of GIoU adds additional time complexity, resulting in sacrificing the run-time efficiency.

V. Experiment Results

We benchmark our results on a 64-bit Ubuntu 18.04 server that has an Intel(R) Core(TM) i9-7900XCPU @ 3.30GHz with 64 GB memory. We also use an NVIDIA GeForce RTX 2080 GPU with 8 GB memory. Although there are some publicly available datasets for human body-parts [17], these datasets do not consider pedestrian detection problems in autonomous driving scenarios. Therefore, to address this problem we collect a pedestrian dataset along with the body-parts label information. Particularly, we label three categories of body parts: head, arm, and leg. We collect images from different streets in the downtown area of Greensboro, North Carolina, USA. Our dataset has a total of 2367 images with 8143 body parts labels, which is rich enough to train the developed BP-SSD network.

We trained the proposed BP-SSD network on body parts and pedestrian labeled images. We showed the change of loss function with the number of steps in Fig. 3. In Fig. 3, the red and blue line shows localization and classification losses, respectively. The brown line shows the total loss, which is the sum of classification loss and localization loss. It is obvious that both classification and localization losses are gradually declining with the increase of number of steps. As a result, the network learns latent representations of the full body along with other body parts.

We evaluate the performance of our proposed BP-SSD with body parts semantic method in terms of Miss-Rate-vs-FPPI-curve in Fig. 4 and evaluation metrics in TABLE I. Currently, our result comparison is limited to the original SSD model. However, the presented constraint optimization method can be easily adopted to any Deep Learning-based object detectors. While testing on a relatively challenging dataset, the proposed BP-SSD with constraint optimizer technique obtains a miss rate of 35% whereas the SSD has a 41.7% miss rate. Thus, the proposed BP-SSD achieves a significantly lower miss rate in contrast to the SSD. TABLE I illustrates that our proposed BP-SSD method has better accuracy, precision, and recall compared to the SSD. As it can be seen in this table, the SSD shows the accuracy of 0.546, precision of 0.897, and recall of

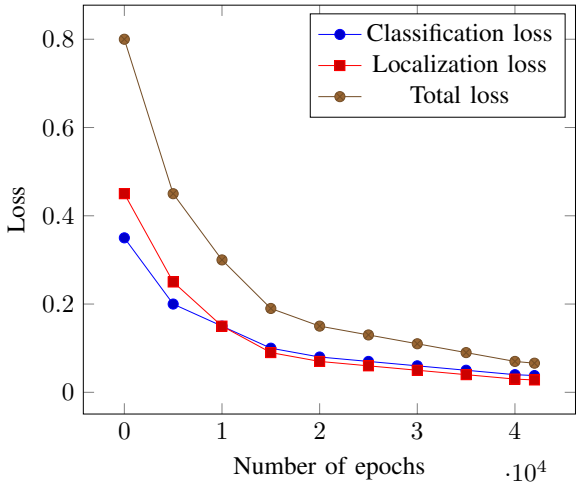

Fig. 3: Losses during training, run across 42×10^3 epochs. The red, green and blue lines represent total loss, localization loss and classification loss, respectively.

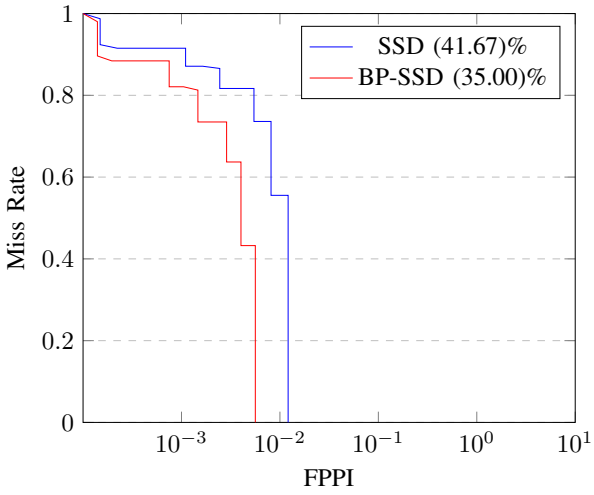

Fig. 4: The comparison of body parts information based technique on SSD model. Applying the body parts information based technique with SSD model get a lower miss rate compare to SSD model without body parts information

0.583, whereas we obtain the accuracy of 0.609, precision of 0.900 and, recall of 0.650 for the proposed BP-SSD. These results demonstrate that the proposed BP-SSD method can detect pedestrians more accurately and precisely.

Fig. 5 illustrates the qualitative improvements of our proposed method over the SSD model. The top row of Fig. 5 shows the detection results of the SSD model and the bottom row shows the detection results of the BP-SSD model. From the top row, we observe that the SSD comes with low confidence scores in some complex scenarios, resulting in miss detection. On the other hand, our proposed BP-SSD can detect pedestrians more accurately by detecting their body parts and using the constraint optimizer to select

(a) (b) (c) (d) (e)

Fig. 5: **Qualitative results comparison:** On the top row, the magenta colored bounding boxes represent the SSD predictions. On the bottom row, the magenta and the red colored bounding boxes represent the BP-SSD predictions. The top row shows the detection results of the SSD model where the SSD finds it difficult to detect pedestrians due to multiple pedestrians in close proximity (a), partially occluded pedestrian (b), complex background (c), deformed shape of pedestrian (d), different articulation of pedestrians (e). The bottom row shows the results of the BP-SSD model for the corresponding images.

appropriate bounding boxes for pedestrians by establishing a semantic relationship based on these body parts. Thus, our proposed BP-SSD outperforms the SSD in complex scenarios like multiple pedestrians in close proximity Fig. 5(a), partially occluded pedestrian Fig. 5(b), complex background Fig. 5(c), deformed shape of pedestrian Fig. 5(d), different articulation of pedestrians Fig. 5(e).

Since the proposed BP-SSD along with optimizer architecture is simple and computationally efficient, we observe a negligible run-time overhead in contrast to the SSD. The BP-SSD obtains a run-time of 44.96 Frame Per Seconds (FPS) which around 1 FPS slower than original SSD.

VI. Conclusion

In this paper, we addressed the problem of real-time accurate pedestrian detection in autonomous driving scenarios. A body parts-based single shot detector (BP-SSD) is developed to accurately and robustly detect pedestrians in real-time. The proposed method utilizes the semantic relationship of body parts to robustly detect pedestrians in partially occluded scenarios. When combined with our proposed constraint optimization strategy for selecting bounding boxes, the BP-SSD can accurately detect multiple pedestrians. Experimental results show that the BP-SSD along with the proposed constraint optimization technique outperforms the SSD-based pedestrian detection method in terms of accuracy, precision, recall, and miss rate. Due to the simple architecture of BP-SSD, we also observe a negligible runtime performance overhead in contrast to the SSD.

References

[1] M. M. Islam, A. A. R. Newaz, and A. Karimoddini, "A pedestrian detection and tracking framework for autonomous cars: Efficient fusion of camera and lidar data," *arXiv preprint arXiv:2108.12375*, 2021.

[2] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6995–7003.

[3] W. Ouyang, X. Zeng, and X. Wang, "Partial occlusion handling in pedestrian detection with a deep model," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 2123–2137, 2015.

[4] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: detecting pedestrians in a crowd," in *European Conference on Computer Vision*, 2018, pp. 637–653.

[5] S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou, "Pedestrian detection via body part semantic and contextual information with DNN," *IEEE Transactions on Multimedia*, pp. 3148–3159, 2018.

[6] I. Yun, C. Jung, X. Wang, A. O. Hero, and J. K. Kim, "Part-level convolutional neural networks for pedestrian detection using saliency and boundary box alignment," *IEEE Access*, pp. 23 027–23 037, 2019.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[8] M. M. Islam, A. A. R. Newaz, B. Gokaraju, and A. Karimoddini, "Pedestrian detection for autonomous cars: Occlusion handling by classifying body parts," in *International Conference on Systems, Man, and Cybernetics*. IEEE, 2020, pp. 1433–1438.

[9] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *European conference on computer vision*. Springer, 2016, pp. 443–457.

[10] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *IEEE International Conference on Computer Vision*, 2017, pp. 4950–4959.

[11] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Winter conference on applications of computer vision*. IEEE, 2017, pp. 953–961.

[12] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 5420–5428.

[13] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *European Conference on Computer Vision*, 2018, pp. 618–634.

[14] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 990–997.

[15] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *IEEE international conference on computer vision*, 2015, pp. 1904–1912.

[16] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4967–4975.

[17] L. Yang, Q. Song, Z. Wang, M. Hu, and C. Liu, "Hier R-CNN: Instance-level human parts detection and a new benchmark," *IEEE Transactions on Image Processing*, vol. 30, pp. 39–54, 2020.