Pedestrian Detection Using MB-CSP Model and Boosted Identity Aware Non-Maximum Suppression

Ameen Abdelmutalab and Chunyan Wang

***Abstract*— Pedestrian detection is an important task in autonomous surveillance systems. Despite the rapid progress in pedestrian detection field, detecting occluded pedestrians remains a challenging task due to the great variations in occluded pedestrians appearance and the drastic loss of pedestrian information in some severe cases. In this paper, we tackle the occlusion problem by proposing a multi-branch pedestrian detection model based on center and scale prediction framework. The proposed model employs features extracted from full pedestrian's body as well as its upper, middle, and lower body parts using four detection branches. This multi-branch approach ensures that data representing the true pedestrian appearances, whether they are partially or completely visible, can dominate the final decision-making, minimizing the interference of non-pedestrian data in the detection. Furthermore, to implement the proposed model, the visibility of different pedestrian parts is appropriately annotated, which facilitates the training process. The final decision is made based on the four MB-CSP branches outputs, using a proposed fusing method, named Boosted Identity Aware-Non Maximum Suppression. On heavy occlusion settings, the proposed model resulted in the miss rates of 27.83%, 47.29% and 33.3% for Caltech-USA, Citypersons and EuroCity Persons datasets, respectively.**

***Index Terms*— Pedestrian detection, occluded pedestrians, multi-branch model, pedestrian body parts, part fusing.**

I. Introduction

HUMAN detection is an important research subject considering its many applications in surveillance, robotics, self-driving cars and video gaming. In particular, accurate pedestrian detection is a challenging task because of the great variations in humans pose and appearance. Besides, pedestrians are frequently occluded in reality, as a result of obstacles such as trees/cars (inter-class occlusion), or the presence of other pedestrians in the scene, usually in crowded areas (intra-class occlusion). Although recent advances in deep learning, supported by the availability of large-scale labelled datasets, helped to improve the performance of some pedestrian detectors, detecting occluded pedestrians remains a challenging task.

Manuscript received 22 December 2021; revised 1 June 2022; accepted 6 July 2022. Date of publication 16 August 2022; date of current version 5 December 2022. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, in part by WestGrid, and in part by Compute Canada. The Associate Editor for this article was Q. Ye. *(Corresponding author: Ameen Abdelmutalab.)*

The authors are with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3B 1M8, Canada (e-mail: ame_abde@encs.concordia.ca; chunyan@ece.concordia.ca).

Digital Object Identifier 10.1109/TITS.2022.3196854

Fig. 1. Detection results of a group of four pedestrians. (a) Using CSP model [12]. (b) Using the proposed MB-CSP model.

Part-based methods [1]–[5], have been developed to improve the detection of occluded pedestrians. Their common approach is to divide a pedestrian target into small parts, e.g., 3×6 grid [3]. By doing so, the data in these parts can be processed more precisely for feature extraction, and networks can be trained specifically to detect these parts, thus improving the detection of occluded pedestrians. It should be mentioned that the difficulty in detection is partially due to the relatively limited data representing occluded pedestrians. A good detector should be able to select features representing real pedestrians and reject those that are irrelevant to the targets, while not significantly increasing the overall model complexity.

Some algorithms target crowded pedestrians by defining new loss functions [6], [7], which lowers the distance between detection boxes belonging to the same pedestrian and increases the distance between boxes of adjacent pedestrians in crowded scenes. This step renders the detector less sensitive to the Non-Maximum Suppression (NMS) threshold in the post-processing stage.

Most pedestrian detectors generate candidate locations for potential pedestrians at different scales. This can be achieved using a sliding window method [8], which is usually slow and opposes the real time constraint in most pedestrian detection applications. Recent pedestrian detection approaches such as R-CNN [9] use region proposal model, where a separate network is designed to generate candidates locations, that are then processed via a classification network. Another approach is based on a single neural network that directly detects pedestrians locations without the need for region proposal model. This approach is used in algorithms such as YOLO [10], SSD [11] and CSP [12], and it is usually faster and reduces the complexity of the overall detector.

In this paper, a new model referred to as Multi Branch Central and Scale Prediction (MB-CSP) is introduced. The proposed model adopts part-based approach and is based

on Center and Scale Prediction (CSP) framework. MB-CSP model involves four detection branches that simultaneously detect upper, middle and lower body parts as well as full pedestrian box. The detection branches are trained with the data, in which visible parts of pedestrian samples are precisely annotated. Moreover, a branch for full-body detection is included in the proposed model in order to learn holistic pedestrian structure and facilitate parts learning process. The losses calculated in the branches are combined into a single loss that is then used to update the network parameters in a joint fashion. Finally, the detection outputs of the four proposed branches are fused using a novel proposed algorithm referred to as Boosted-Identity-Aware-Non-Maximum-Suppression (BIA-NMS).

In Fig. 1, pedestrian detection examples show the improvement in the detection of occluded pedestrians using the proposed MB-CSP model, compared to CSP detector [12]. Moreover, on heavy occlusion settings, the proposed model resulted in the miss rates of 27.83%, 47.29% and 33.3% for Caltech-USA, Citypersons and EuroCity Persons datasets, respectively.

The contribution of this paper can be summarized as follows:

1) Proposing an anchor-free multi-branch detection model with four detection branches. Although the branches are configured identically, each of them is learned, specifically, to detect various patterns of a particular kind of body parts.
2) Implementing new parts annotation that recognises upper, middle and lower pedestrian parts. Each part is considered positive if it is visible for the specific pedestrian.
3) Introducing Boosted Identity Aware Non Maximum Suppression (BIA-NMS) to fuse different branches outputs. In particular, BIA-NMS targets intra-class occlusion by suppressing duplicated detection boxes of a single pedestrian, while preserving boxes of adjacent pedestrians.

The rest of the paper is organised as follows: Section II highlights the main approaches of handling occlusion, and presents an overview of CSP framework. In Section III, a detailed description of the design of the proposed model and its composition is presented. Section IV is dedicated to the presentation of the experiments and the results. The conclusion of the work presented in this paper is found in Section V.

II. Related Work

The related work to the proposed model is presented in this section. First, different approaches to target occluded pedestrians are discussed, followed by an illustration of anchor-free detectors and the CSP framework.

A. Occlusion-Handling Algorithms

One of the most commonly used approaches to tackle occlusion is part-based approach [1], [2], [13], where the full pedestrian body is divided into multiple parts, usually based on different occlusion patterns. During occlusion, some body parts remain visible, hence detecting these parts is more convenient compared to detecting the full body with mixed features of the pedestrian and the barrier. Earlier part-based approaches used ensembles models, in which separate part detectors are used independently. This approach is not suitable for real-time processing, as the system complexity grows linearly with the addition of every part detector. Moreover, ensembles models ignore the correlation between different parts during learning, resulting in a non context-aware part detectors. Other methods build parts models using a joint framework [4], where different body parts are trained collaboratively using a single convolutional neural network (CNN). This approach reduces the complexity presented in ensemble models, however it lacks accurate parts annotation. In [5], authors introduced multi label learning with separate labels assigned to different body parts, however their approach uses part pool of 20 parts, and requires region proposal network (RPN), which adds to the complexity of the final detector. Authors in [14] introduced occlusion-handling algorithm based on full and visible body information, however, their definition of visible body is rather broad, since it includes different parts of the body based on different occlusion patterns, making the training process more challenging. Zhang et al proposed an attention guided model [15] to reweigh convolutional channels that represent varying occlusion patterns. Other authors integrated additional features to improve pedestrian detection, for example Du et al [16] applied features from a pixel-wise semantic segmentation network, and Song et al [17] integrated temporal information from adjacent video frames. Different track of work focuses on improving crowded pedestrian detection by introducing new loss functions [6], [7], their goal is to minimize the distance between duplicate detection boxes of the same pedestrian and maximize the distance between adjacent pedestrians boxes, eventually preventing over elimination by Non-Maximum Suppression (NMS).

B. Anchor-Based-Detectors Vs Anchor-Free-Detectors

Anchor-based detectors overcome the increase in detectors complexity when using sliding window or region proposal models, by introducing pre-defined anchors at different scales, locations and aspect ratios. Despite the significant reduction in the complexity associated with applying anchor-based approaches, they limit the generality of detectors, since the design of anchor boxes varies to meet the requirements of different datasets. To overcome this problem, researchers introduced anchor-free-detectors [10], [12], [17], [18], in which the network predicts pedestrian locations directly without the need for defining and matching anchor boxes.

Center and Scale Detector (CSP) [12] outputs the center and the scale of pedestrians directly, without the need for predefined anchor boxes. During training, bounding boxes information provided with the dataset, is used to calculate pedestrian center (center point of the bounding box), scale (function of the bounding box height and width) and offset (to compensate for the drop in localization accuracy as the detection is performed in a lower resolution compared to the

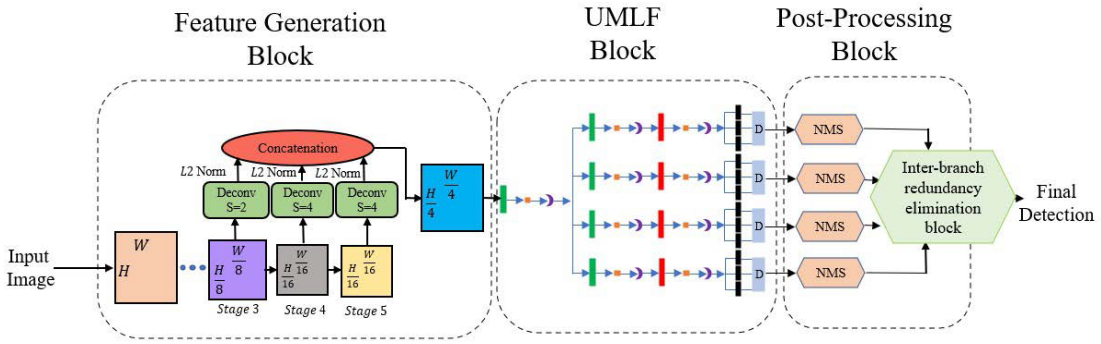

Fig. 2. MB-CSP architecture consisting of three blocks, namely feature generation block, UMLF block and post-processing block.

original image size). CSP model is trained by minimizing the summation of center, scale and offset losses.

III. Proposed Model

Detecting individual pedestrians in crowded areas is a challenging task, as people are often occluded. A pedestrian can be partially obstructed by objects of other classes such as vehicles and trees, which is referred to as inter-class occlusion. An intra-class occlusion occurs when a pedestrian is partially occluded by other pedestrians. In general, there are two hurdles when detecting occluded pedestrians.

- Real pedestrian features are mixed with features of the occluding barrier. This hurdle is present in both inter-class and intra-class occlusions, and can result in confusion when learning pedestrian characteristics, eventually leading to wrong detections. To overcome this hurdle, the proposed model utilizes part-based detectors, each of which is exclusively learned from visible pedestrian parts.
- Multiple-detection of a single pedestrian is a common problem in most detection models. The proposed multi-branches CSP model may exacerbate this problem by creating duplicates from its different branches. To address this issue, the proposed model utilizes non-maximum suppression to eliminate duplicates within the same branch, and proposes a novel post-processing algorithm for removing duplications across the different branches.

The block diagram of the proposed model, referred to as Multi-Branch Center and Scale Predictor (MB-CSP), is illustrated in Fig. 2. It is composed of the following three blocks.

1) **Feature Generation Block** to convert the input images into suitable feature maps for pedestrian detection at different scales.
2) **UMLF Block** to process features data in its four branches, each of which produces data maps indicating the location/scale/offset of potential pedestrian targets.
3) **Post-Processing Block** to fuse the data produced by the UMLF branches and determine the final detections.

Each block is discussed in detail in the following sections.

A. Feature Generation Block

Feature Generation Block is used to extract the basic features for pedestrian detection. The block is based on ResNet50 [19] and is pre-trained on ImageNet database in order to facilitate the learning process. The input images of the feature generation block are sized HxW. As shown in Fig. 2, the feature maps generated by stages 3, 4 and 5 are sized $(W/8)^2$, $(W/16)^2$ and $(W/16)^2$, respectively. These maps are processed by deconvolutional layers followed by L2-normalization in order to unify their dimensions to $(H/r) \times (W/r)$, where r = 4 is the downsampling factor. The upsampled feature maps, carrying context information at different scales, are then concatenated to form the input to the UMLF block.

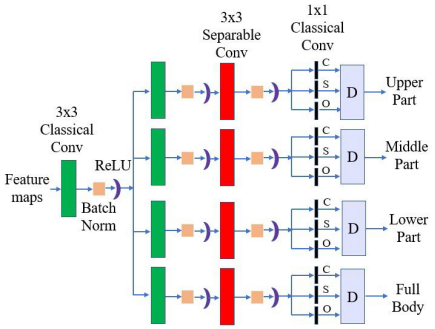

Fig. 3. UMLF network architecture, where C, S, O and D denote center map, scale map, offset maps and output decoder, respectively.

B. UMLF Block

If pedestrians appear partially occluded in an image, the pixel data in the occluded part carry the features of the occluding barriers, which can contribute adversely to the detection of the pedestrians. UMLF block is designed to mimic human perception of a partially occluded pedestrian by extracting the relevant features from visible pedestrian parts and ignoring data variations in the occluding barrier. To do so, one needs to partition the view of a pedestrian into parts so that visible areas containing actual features of real pedestrians are separated from the occluded areas, making it possible to exclude non-pedestrian features in the training process. As a result, the features of pedestrian parts will be processed in separate branches of the block, and each branch will learn features of its corresponding part.

A pedestrian appearance can be obstructed differently, and the patterns of occlusion are not unique. To partition a pedestrian view appropriately, the following elements are considered.

- The designed partitions must have recognizable and distinguishable patterns that discriminate pedestrians from irrelevant objects.
- Partitions must suit the different occlusion patterns so that each pedestrian has at least one visible part, without significant interference of occluding element, in most of the occlusion scenarios.
- The number of partitions must be reasonable, as more partitions implies more branches, therefore increasing the complexity of the overall system.

Taking the above-stated points into consideration, a pedestrian view is partitioned into overlapped upper, middle, lower and full body parts. Corresponding to these partitions, the proposed UMLF block has four branches to detect the four parts, respectively.

- **Upper Part Branch.** This branch is dedicated to detecting pedestrians face and shoulders using their distinguishable contours. The upper part detection is crucial in detecting highly occluded pedestrians, where face and shoulders might be the only visible part.
- **Middle Part Branch.** The features of the middle part, including the torso, of a pedestrian's view are very different from the upper or lower parts. This branch is trained to identify the patterns of the middle part, and its output data help to detect reasonable and partially occluded pedestrians.
- **Lower Part Branch.** This branch is specialized to detect the unique shape of the lower part, i.e. the trunk and legs of a pedestrian. If this part is visible, this branch will detect it and contribute to the correct final decision.
- **Full Body Branch.** In case of fully visible pedestrians, a full-body detection is evidently more advantageous than that of part-based, particularly when there are many fully visible pedestrians in the training samples. Hence, this full-body branch is placed to minimize the risk of missing fully visible pedestrian targets.

A good pedestrian detection needs a good identification of the patterns distinguishing the pedestrian targets from the rest of the image. The four-branch structure of the proposed UMLF Block permits each branch to be trained specifically to identify the distinguished patterns of the designated part. If the part is visible, the branch will generate a significant output, otherwise, no target patterns will be detected and the output will be weaker. The final detection decision is based on the outputs of all the four branches, dominated by the data generated from the visible parts.

The detailed structure of the UMLF block is illustrated in Fig. 3. The input data, i.e., the 2D maps carrying features extracted in different scales, are first fused by means of a convolutional layer of 256 kernels. The outputs of this layer are then applied to each of the four branches for the detections of the upper, middle, lower and full-body parts, respectively.

In each of the four branches, as shown in Fig. 3, the detection of the designated part is performed by two convolutional layers, each of which has 256 kernels. It should be noted that a standard 3×3 convolution is applied in the first layer, whereas the second layer is a 3×3 separable convolution (consisting of depth-wise filter of size 3×3 followed by 1×1 classical convolution filter). The separable convolution acts as a channel-wise attention mechanism to highlight the important features in each map. The output data containing information on targets centers, scales and offsets are then processed by 1×1 convolutions to generate the final center, scale and offset maps.

UMLF branches are configured identically. However, the convolution kernel parameters in each branch are designed to learn the associated features of each part. Fig. 4 illustrates two detection examples, each having an original input image and its associated upper, middle, lower and full-body center heat maps generated by the four UMLF branches. The first example involves two fully visible pedestrians, with their corresponding four center heat maps, indicating clearly and coherently the locations of their parts and full-bodies. The second example is a challenging heavy occlusion case, as one of the three pedestrians is severely occluded. Accordingly, the full-body branch can only detect two pedestrians, as shown in Fig. 4(j). So do the branches for the middle and lower parts. However, the center heat map in Fig. 4(g) produced by the upper part branch clearly indicates three pedestrian locations, which is crucial to detect the severely occluded third pedestrian. These two examples demonstrate the effectiveness of the UMLF branches in enhancing detection quality in the presence of significant heavy occlusion, without jeopardizing other cases.

As shown in Fig. 3, there is a decoder in each of the four UMLF branches. Each decoder converts the center, scale, and offset maps in each branch to a list of bounding boxes based on their predefined aspect ratios, illustrated in Fig. 5. It should be mentioned that a single pedestrian target can be detected multiple times in each of the four UMLF branches, which results in multiple overlapped full-length bounding boxes per branch, creating a type of redundancy referred to as intra-branch redundancy. Moreover, the same pedestrian may be detected by more than one branch, especially if a pedestrian is fully visible in the image, this type of redundancy is referred to as inter-branch redundancy. The post-processing block, presented in the following sub-section, is intended for bounding boxes refinement and redundancy elimination.

C. Post-Processing Block

The post-processing block is designed to eliminate duplicated pedestrian boxes, and to identify/preserve one bounding box per detected pedestrian. It is performed in two steps to eliminate intra-branch redundancy and inter-branch redundancy, respectively.

For intra-branch redundancy, the duplicated boxes generated in the same branch are removed by means of Non-Maximum Suppression (NMS). It is known that a single pedestrian can be indicated by highly overlapped boxes. The degree of overlapping reflects the likeness of the case, which is measured by

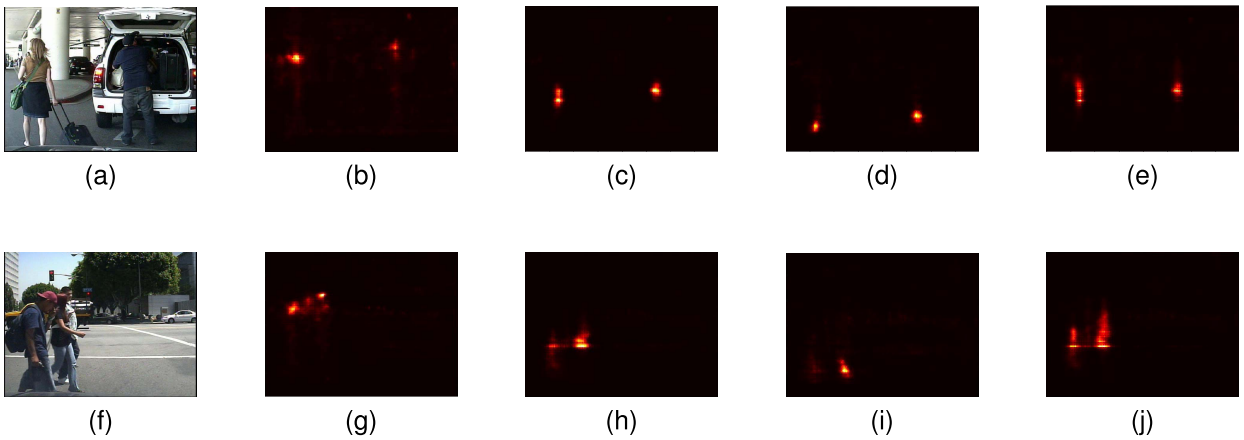

Fig. 4. Two detection examples of the proposed MB-CSP model. (a) and (f) Input images. (b) and (g) Center heat maps of the upper parts. (c) and (h) Center heat maps of the middle parts. (d) and (i) center heat maps of the lower parts. (e) and (j) Center heat maps of the full body.

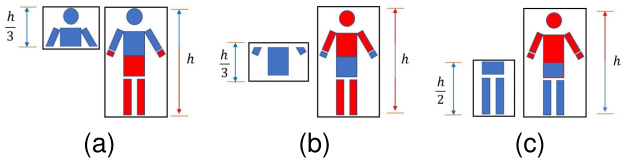

Fig. 5. Detection boxes of the three parts and their extension. (a) Upper part. (b) Middle part. (c) Lower part.

Intersection Over Union (IOU) index representing an overlap between 0% and 100%. If IOU value of two bounding boxes is higher than a threshold, they are considered to indicate the same pedestrian and the one having the lower confidence score will then be eliminated.

The above-mentioned threshold should be chosen very carefully. As NMS is performed in each of the four branches, the thresholds can be selected differently based on the detection criteria of different body parts. To decrease the risk of false eliminations, the IOU threshold of the upper part is set more cautiously to be 0.6, compared to 0.5 for the other branches. In case of detecting pedestrians that are heavily occluded by other pedestrians, only the upper parts of the occluded pedestrians can be differentiated, while their full-length boxes may highly overlap. In this case, setting the IOU threshold for the upper branch to 0.6 allows to preserve the two individual pedestrian upper parts.

The NMS performed in each of the four branches removes most of intra-branch redundant bounding boxes, and the remaining bounding boxes represent potential pedestrian candidates detected in each branch. The bounding-boxes lists generated by the four branches are then examined together, in the second step, to eliminate inter-branch redundancy.

The inter-branch redundancy can be caused by the detection of a single fully visible, or mostly visible, pedestrian in multiple branches, where the redundant bounding-boxes are usually highly overlapped. However, if two pedestrians are heavily occluded by each other, their boxes generated in the same branch or different branches, can also be overlapped. In order not to falsely eliminate the bounding boxes representing heavily occluded pedestrians, one needs to look into not only the overlap rate, but also other indications from the four bounding boxes lists. The operation, referred to as Boosted Identity Aware Non-Maximum Suppression (BIA-NMS), is to check if a group of overlapped boxes represent a single pedestrian or multiple heavily occluded ones.

BIA-NMS is proposed with a view to minimizing the risk of merging heavily overlapped boxes belonging to different pedestrians, while suppressing duplicated pedestrian boxes. The following two points are used to develop BIA-NMS algorithm.

1) BIA-NMS aims at eliminating duplicated detection boxes, generated by different branches, of the same pedestrian target. Hence, no boxes of the same branch can be merged in this procedure, to eliminate the risk of missing occluded targets. To be more specific, at a given location, the boxes to be checked must be from different branches and are eventually merged to be one.
2) At a given location, relatively high scores of multiple boxes from different branches indicate a detection of multiple parts of the same pedestrian, implying a high certainty of true detection. In this case, the final detection score will be boosted.

BIA-NMS is performed in the following steps.

1) Sort all the detected boxes in a descending order based on their confidence scores.
2) Identify the box with the highest score and refer to it as B_{max}.
3) Calculate the IOU between all the detected boxes and B_{max}.
4) Identify the boxes with IOU greater than 0.6 and add them to the new list $B_{duplicate}$, make sure that only one box per branch is added to $B_{duplicate}$ (the box with the highest IOU per branch).
5) Define N as the number of boxes in $B_{duplicate}$.
6) Modify the score of B_{max} as follows:

$$Score_B = (N-1) \times \beta \times Score_O + Score_O \quad (1)$$

where $Score_B$ denotes the boosted score of B_{max}, $Score_O$ is the original score of B_{max} and β is the boosting weight set to 0.08.

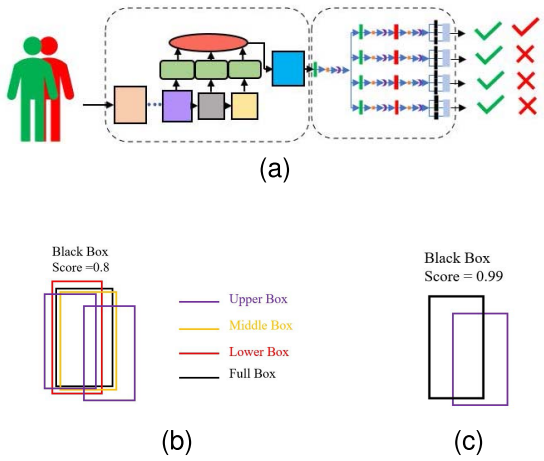

Fig. 6. (a) Example of input involving two pedestrians, of whom one is severely occluded. (b) Boxes generated by the four branches around the pedestrians locations. (c) Result produced by the proposed BIA-NMS. The two boxes from upper part branch should not be merged.

7) Add B_{max} and its boosted score to the final detection list.
8) Remove B_{max} and $B_{duplicate}$ from the initial list.
9) Repeat the process starting from step 1).

Fig. 6 presents an example of two pedestrians applied to the proposed MB-CSP model. The pedestrian in green is fully visible, hence the UMLF block can detect its upper, middle, lower and full body parts (indicated by the green check-marks). Meanwhile, the red pedestrian is highly occluded and only the upper part can be detected (red check-mark), and his middle, lower and full body parts are easily missed (red-crosses). The five detected boxes are depicted in (b), where the full-box for the green pedestrian is represented by a black colour and has the highest score of 0.8. The process of elimination starts by considering the IOU between all the detected boxes and the box with the highest score (the black box). In this example, the four boxes have IOU values greater than 0.5 with the black box. However in (c), BIA-NMS eliminates three of the four highly overlapped boxes and preserves one box (violet box). This is because violet boxes represent upper body boxes, and the black box is highly overlapped with two violet boxes, hence only one of them is eliminated (the one with highest IOU value). Finally, the score of the black box is boosted to become 0.99 using equation 3.

In Fig.7(a), an image including three pedestrians with different degrees of occlusion is illustrated. If NMS is applied in the second post-processing stage, one of the three pedestrians will be missed in the detection due to the heavy occlusion, as shown in Fig. 7(b). The proposed BIA-NMS helps to capture the missed one, so that all the three pedestrians are detected. Fig. 7(c) illustrates the detection result, indicated by the three boxes, before the boosting. The scores of the detected pedestrian boxes are boosted, by means of the calculation defined by Equation 3, as shown in Fig. 7(d).

D. Parts Annotation

Most pedestrians datasets provide annotation information that specify two bounding boxes for every pedestrian. Visible

Fig. 7. (a) Input image. (b) Detection result by MB-CSP and NMS, (c) by MB-CSP and BIA-NMS before boosting, and (d) by MB-CSP and BIA-NMS after boosting.

bounding box that indicates visible area of a pedestrian, and Full bounding box that describes the full pedestrian body including its extension if it is occluded. Annotation information is provided as follows:

$$Annotation = \left[x_f, y_f, w_f, h_f, x_v, y_v, w_v, h_v\right] \quad (2)$$

where x_f, y_f and x_v, y_v are the coordinates of the top left corner of the full box and the visible box, respectively. w_f, h_f and w_v, h_v are their corresponding width and height.

Since the proposed model has four detection ends, each pedestrian in the image is assigned four bounding boxes, namely BB_u, BB_m, BB_l and BB_f, to describe the upper, middle, lower and full pedestrian parts, respectively. Algorithm 1 presents detailed explanation of the annotation algorithm.

E. Model Loss

To calculate the total loss ($Loss_T$) of the proposed MB-CSP model, the four branches losses are combined as follows:

$$Loss_T = \alpha_1 Loss_U + \alpha_2 Loss_M + \alpha_3 Loss_L + \alpha_4 Loss_F \quad (3)$$

where $Loss_U$, $Loss_M$, $Loss_L$ and $Loss_F$ indicate the model loss of the upper, middle, lower and full branches, respectively. For simplicity, $\alpha_1, \alpha_2, \alpha_3$ and α_4 are set to 1s, however, adapting different weights can be investigated.

Furthermore, for branch P, the branch loss ($Loss_P$) can be expressed as:

$$Loss_P = Loss_{C\_P} + Loss_{S\_P} + Loss_{O\_P} \quad (4)$$

where $Loss_{C\_P}$, $Loss_{S\_P}$, and $Loss_{O\_P}$ are the center, scale and offset losses for branch P, respectively.

To calculate the center loss for every branch, the same procedure presented in [12] is followed. The main difference is, centers are calculated for every specific part instead of a

Algorithm 1 Parts Annotation

Input:
$BB_f = [x_f, y_f, w_f, h_f]$
$BB_v = [x_v, y_v, w_v, h_v]$

Output:
$BB_u = [x_u, y_u, w_u, h_u]$
$BB_m = [x_m, y_m, w_m, h_m]$
$BB_l = [x_l, y_l, w_l, h_l]$
$BB_f = [x_f, y_f, w_f, h_f]$

procedure PARTS ANNOTATION(BB_{full}, BB_{vis})
  for img in **Images do**
    for ped in **Pedestrians do**
      $BB_u = \left[x_f, y_f, w_f, \frac{h_f}{3}\right]$
      $BB_m = \left[x_f, y_f + \frac{h_f}{3}, w_f, \frac{h_f}{3}\right]$
      $BB_l = \left[x_f, y_f + \frac{h_f}{2}, w_f, \frac{h_f}{2}\right]$
      if $\frac{Area(BB_u \cap BB_v)}{Area(BB_u)} > 0.2$ **then**
        $BB_u \leftarrow BB_u$
      else
        $BB_u \leftarrow [0, 0, 0, 0]$
      if $\frac{Area(BB_m \cap BB_v)}{Area(BB_m)} > 0.2$ **then**
        $BB_m \leftarrow BB_m$
      else
        $BB_m \leftarrow [0, 0, 0, 0]$
      if $\frac{Area(BB_l \cap BB_v)}{Area(BB_l)} > 0.2$ **then**
        $BB_l \leftarrow BB_l$
      else
        $BB_l \leftarrow [0, 0, 0, 0]$
      Return BB_u, BB_m, BB_l, BB_f

single center for the entire pedestrian body. Following this procedure, the cross entropy center loss is defined as:

$$Loss_C = \begin{cases} \frac{-1}{N} \sum_{i=1}^{\frac{W}{r}} \sum_{j=1}^{\frac{H}{r}} (1 - p_{ij})^{\gamma} \, log(p_{ij}), & y_{ij} = 1 \\ \frac{-1}{N} \sum_{i=1}^{\frac{W}{r}} \sum_{j=1}^{\frac{H}{r}} (1 - M_{ij})^{\beta} \, p_{ij}^{\gamma} log(1 - p_{ij}), & y_{ij} = 0 \end{cases} \tag{5}$$

where N is the number of objects (specific body part) in the image, H, W and r are the height, width and downsampling factor of the image, respectively. M_{ij} is a 2d Gaussian map built around the center of every part, based on the height and width of the specific part, this is done to reduce the uncertainty created by the negatives surrounding center points, by reducing their effect on the total loss [12]. p_{ij} is the predicated probability for a center to be presented at location i, j, and y_{ij} is the ground truth value, equals to 1 if there is a center at location i, j and 0 otherwise. γ and β are hyper-parameters, γ is set to 2 as recommended by [20], and β is set to 4 [21]. Scale and offset losses of every branch are calculated using smooth L1 loss equation.

IV. Experiments

In this section, the performance evaluation of the proposed model is presented. To conduct the experiments for the evaluation, image samples from the three datasets, namely Caltech-USA [22], CityPersons [23] and EuroCity Persons [24], have been used. The evaluation process and the results obtained, in terms of detection miss-rate and processing time, are presented, in comparison with the existing pedestrian models.

TABLE I
EVALUATION SETTINGS FOR CALTECH-USA, CITYPERSONS AND EUROCITY PERSONS DATASETS

Setting	Visibility	Height
Bare (B)	$\geq$90%	$\geq$50 pixels
Reasonable (R)	$\geq$65%	$\geq$50 pixels
Partial Occlusion (P)	65% to 90%	$\geq$50 pixels
Heavy Occlusion (H)	20% to 65%	$\geq$50 pixels
All (A)	$\geq$0%	$\geq$20 pixels
Small (S)	$\geq$65%	50 to 75 pixels

A. Datasets

1) Caltech-USA: The image samples of this dataset are extracted from an approximately 10 hours video recorded by a car driving in the greater Los Angeles area. Images are of size (640 × 480). The dataset contains a total of 350,000 labelled bounding boxes in 250,000 frames. For the experiments, one image has been taken out of every 30 frames from the original sequence, and 4250 training images and 4024 testing images are obtained. Furthermore, The improved annotation, presented in [25], is adopted for the training and testing. The proposed model has been evaluated using a log-average miss rate for false positive per image in the range (10^{-2} to 1). The evaluation settings are presented in Table I.

2) Citypersons: The dataset consists of 2975 training images, 500 validation images and 1575 testing images captured in 27 different cities in Germany and neighbouring countries. All images are of size (2048 × 1024). The dataset has around 20K pedestrians, where only less than 30% of them are fully visible. The great variation in pedestrian scale, occlusion and background makes Citypersons a challenging dataset for pedestrian detection. In this paper, the validation images are used for testing.

3) EuroCity Persons (ECP): The image samples are taken under more diverse weather, illumination and background conditions than those in Caltech and Citypersions datasets. Images are of size (1920 × 1024), with over 200K annotated bounding-boxes. In this work, the validation set of ECP is used for testing.

B. Experiments Setup

Simulations have been performed using NVIDIA V100 Volta GPUs with 64G memory. Following the training implementation in [12], the backbone network is pre-trained on ImageNet, and the total model is fine-tuned using Adam optimizer. Furthermore, training images have been resized to reduce the training computational complexity. However, the full image size is used in the testing stage. Further, the implementation details are presented in Table II.

C. Ablation Study

The proposed MB-CSP model is designed to use the information of the upper, middle, lower and full-body parts in

TABLE II
TRAINING DETAILS

Dataset	GPUs	Images per GPU	Resized Image	Learning rate	Number of Iterations
Caltech-USA	2	8	336×448	10^{-4}	15K
EuroCity	2	6	512×960	2×10^{-4}	166K
CityPersons	4	2	640×1280	2×10^{-5}	37.5K

TABLE III
TRAINED ON CITYPERSONS TESTED ON CITYPERSONS

Method	R	H	P	B	Test-Time
UF	12.6%	46.62%	11.32%	8.87%	0.40 s/img
UMF	10.35%	46.82%	9.64%	6.74%	0.44 s/img
UML	10.71%	47.12%	10.35 %	6.95%	0.44 s/img
UMLF	10.08%	47.29%	10.22%	6.12%	0.48 s/img

an optimized manner, in order to minimize the interference of the features belonging to the occluding barriers. In this section, three alternatives of UMLF model are investigated, namely, *Upper and Full body parts* (UF) model, *Upper, Middle and Full body parts* (UMF) model and *Upper, Middle and Lower parts* (UML) model. Extensive simulations have been conducted in order to recognise and compare the pros and cons of each model.

UF model is the simplest block to design MB-CSP detector, in which, only upper body box and full pedestrian box are considered. UF model reported the best results compared to other models when tested on heavily occluded pedestrians with miss-rate of 46.62%, as it is clear in Table III. This is expected because lower and middle parts boxes carry no pedestrian information in this case. However, UF performs poorly for the remaining testing subsets.

On the other hand, UMF model, achieved better accuracy compared to UF model on *Reasonable*, *Partial* and *Bare* subsets with miss-rates of 10.35%, 9.64% and 6.74%, respectively. These results indicate the importance of middle body information for detecting visible and partially occluded pedestrians. Finally, UML utilises the information in different body parts and neglects full box information. Comparing UML to UMLF model shows a drop in the detection accuracy for all testing subsets when using UML. This observation suggests the importance of the full box information in detecting pedestrians at all occlusion patterns.

Table III presents the testing time required by the different models to process a single image and output pedestrians locations. UF model requires the minimum time of 0.4 seconds per image, while UMF and UML need 0.44 seconds per image compared to UMLF model with 0.48 seconds per image. In general, the increment in UMLF processing time is minor, as it only adds a few convolutional layers to predict the different body parts.

D. Comparison With the State of the Art Methods

1) Testing on Caltech-USA Dataset: The proposed MB-CSP+BIA-NMS model has been compared to the-state-of-arts detectors on Caltech-USA testing sets. MB-CSP+BIA-NMS refers to the proposed model trained on Caltech-USA training sets, and MB-CSP+BIA-NMS (City) indicates the model pre-trained on Citypersons training sets and fine-tuned on Caltech-USA training sets. Fig. 8 compares the proposed model to the-state-of-arts detectors reported in Caltech-USA dataset website.[1] All the algorithms are evaluated on the improved annotated testing subsets, hence there is a variation in their results compared to the ones reported on Caltech-USA website.

In Fig. 8 (a), MB-CSP+BIA-NMS (City) achieved the lowest miss-rate of 4.38% on *Reasonable* subset, Compared to 5.11% for AdaptFasterRCNN [23] and 5.13% for AR-Ped [26]. These results reflects the advantage of using the proposed model for detecting fully visible and partially occluded pedestrians, particularly by boosting pedestrians scores using BIA-NMS method in post-processing.

For *Heavy* occlusion subset depicted in Fig. 8 (b). MB-CSP+BIA-NMS (City) and MB-CSP+BIA-NMS reported superior miss-rates of 27.83% and 30.55%, respectively. Lower by 4.4% compared to the best reported method F-DNN2+SS [27] with a miss-rate of 32.28%. This gain in performance is attributed to the proper design of the multi-branch model. Finally, the proposed methods showed decent performance on Caltech-USA *All* subset in Fig. 8(c), with miss-rates of 50.18% and 51.14%, respectively.

To further investigate the performance of the proposed model. Table IV presents the results of recent state-of-arts detectors that have not been included in Caltech-USA website. The proposed model shows improvement over the Original CSP [12] in all testing subsets. Furthermore, MB-CSP+BIA-NMS surpassed all detectors in *Reasonable* and *Heavy* occlusion subsets.

2) Testing on CityPersons Dataset: The performance of the proposed model is compared to the state-of-the-arts methods on Cityperons validation set in Table IV. The proposed model in this case, has been trained on CityPersons Dataset. MB-CSP+BIA-NMS outperformed all the reported methods at all testing subsets. For *Reasonable* and *Bare* subsets, MB-CSP+BIA-NMS reported miss-rates of 10.08% and 6.12%, respectively. Surpassing the best reported miss-rate by almost 1%. This improvement emphasizes the benefits of using the proposed model in detecting highly visible pedestrians. Furthermore, when detecting occluded pedestrians, MB-CSP+BIA-NMS scored 47.29% and 10.22% for *Heavy* and *Partial* occlusions, compared to 49.3% and 10.4% for CSP [12]. Proving the superiority of the proposed model in detecting heavily occluded pedestrians with more than 2% gain on Caltech-USA and Citypersons dataset.

3) EuroCity Persons (ECP) Dataset: Finally, Table V evaluates the proposed MB-CSP+BIA-NMS algorithm, trained and tested on EuroCity dataset. In the *heavy* occlusion setting, CSP+BIA-NMS obtained state-of-the-arts results, similar to Cascade R-CNN [28] with miss-rate of 33.3%, demonstrating the effectiveness of the proposed model in detecting occluded pedestrians. Moreover, On *small* setting, CSP+BIA-NMS outperformed all other models with a miss-rate of 10.5%. Finally, MB-CSP+BIA-NMS produced a miss-rate of 10.4% on *Reasonable* subset, which is comparable to the best reported result using SSD model [24].

[1] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

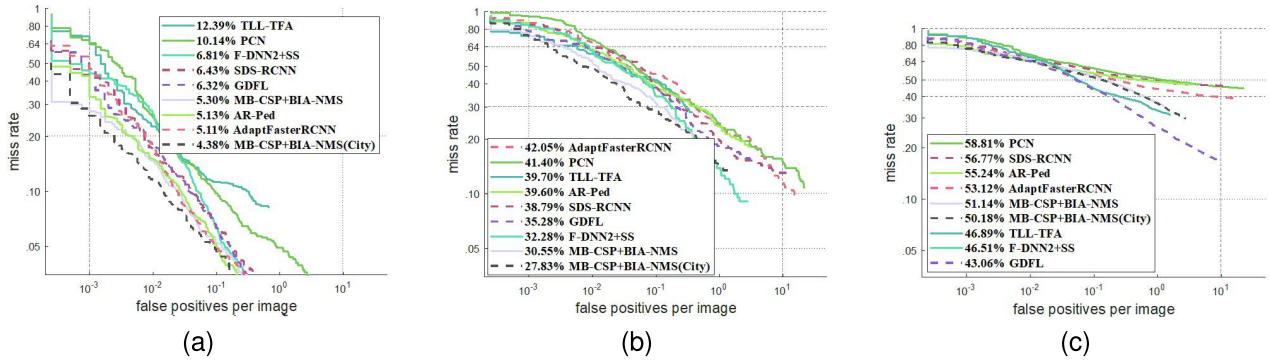

Fig. 8. Comparison of the proposed model and the state-of-the-art methods on Caltech-USA, using average miss rate (MR%) on (a) *reasonable*, (b) *heavy*, and (c) *all* subsets.

TABLE IV

COMPARISON WITH THE STATE-OF-THE-ART ON CALTECH AND CITYPERSONS DATASETS

Method	Dataset	R	H	P	B	A
PAMS-FCN [29]	Caltech	N.A.	47.4%	N.A.	N.A.	53.7%
CSP [12]	Caltech	4.5%	45.8%	N.A.	N.A.	56.9%
CircleNet [30]	Caltech	10.2%	44.5%	N.A.	N.A.	46.4%
CSP (City) [12]	Caltech	3.8%	38.5 %	N.A.	N.A.	54.4 %
FRCN+A+DT [31]	Caltech	8.0%	37.9%	N.A.	N.A.	N.A.
Couple [32]	Caltech	4.7%	34.6%	N.A.	N.A.	N.A.
MB-CSP+BIA-NMS	Caltech	5.30%	30.55%	N.A.	N.A.	51.14%
MB-CSP+BIA-NMS (City)	Caltech	4.38%	27.83%	N.A.	N.A.	50.18%
TLL [17]	City	14.4%	52.0%	15.9%	9.2%	N.A.
RepLoss [7]	City	13.2%	56.9%	16.8%	7.6%	N.A.
OR-CNN [33]	City	12.8%	55.7%	15.3%	6.7%	N.A.
Couple [32]	City	12.2%	49.8%	N.A.	N.A.	N.A.
ALFNet [34]	City	12.0%	51.9%	11.4%	8.4%	N.A.
CircleNet [30]	City	11.7%	50.2%	12.2%	7.1%	N.A.
CSP [12]	City	11.0%	49.3%	10.4%	7.3%	N.A.
MB-CSP BIA-NMS	City	10.08%	47.29%	10.22%	6.12%	N.A.

TABLE V

COMPARISON WITH THE STATE-OF-THE-ART ON EUROCITY PERSONS (ECP) DATASET

Method	Dataset	R	H	S
Faster R-CNN [24]	ECP	7.3%	52%	16.6%
YOLOv3 [24]	ECP	8.5%	37%	17.8%
SSD [24]	ECP	10.5%	42.0%	20.5%
Cascade R-CNN [28]	ECP	6.6%	33.3%	13.6%
MB-CSP BIA-NMS	ECP	10.4%	33.3%	10.5%

TABLE VI

PROCESSING TIME COMPARISON ON CALTECH-USA DATASET

Method	Test Time
ALFNet [34]	0.27 s/img
CSP [12]	0.33 s/img
MB-CSP BIA-NMS	0.48 s/img

Moreover, the time required to process one image by the proposed MB-CSP+BIA-NMS model is investigated and compared to the processing time of CSP [12] and ALFNet [34] models as shown in Table VI. On average, MB-CSP+BIA-NMS requires 0.48 seconds to compute pedestrians locations in one image compared to 0.33 seconds and 0.27 seconds for CSP [12] and ALFNet [34] models, respectively. The increment in the processing time is expected as the proposed model detects four body parts and has more convolutional layers. However, the reported processing time, is still sufficient for accurate pedestrian detection in real time.

V. CONCLUSION

In this paper, a multi-branch deep learning model to improve the accuracy of occluded pedestrian detection has been proposed. The proposed model, referred to as MB-CSP, is based on Center and Scale Prediction (CSP) framework. MB-CSP model involves four detection branches to detect upper, middle, lower and full-body pedestrian parts, respectively. A new post-processing algorithm called Boosted Identity Aware Non-Maximum Suppression (BIAS-NMS) is utilized to merge the four branch outputs and produce final detection results. Furthermore, a new part annotation has been introduced based on parts visibility for every pedestrian sample in order to insure accurate part training. Finally, it is important to note that the proposed model is able to function effectively under the condition of the minimum pedestrian height of 50 pixels. Beyond this limit, the information in partitioned pedestrian parts is insufficient for the proposed model to operate well.

REFERENCES

[1] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1505–1512.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[3] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1904–1912.

[4] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2056–2063.

[5] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3506–3515.

[6] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 637–653.

[7] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[11] W. Liu *et al.*, "SSD: Single shot multibox detector," 2015, *arXiv:1512.02325*.
[12] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5182–5191.
[13] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 90–97.
[14] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 135–151.
[15] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.
[16] 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 *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.
[17] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 536–551.
[18] L. Tychsen-Smith and L. Petersson, "DeNet: Scalable real-time object detection with directed sparse sampling," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 428–436.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[21] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[22] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[23] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4457–4465.
[24] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "EuroCity persons: A novel benchmark for person detection in traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1844–1861, Aug. 2019.
[25] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1259–1267.
[26] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7231–7240.
[27] X. Du, M. El-Khamy, V. I. Morariu, J. Lee, and L. Davis, "Fused deep neural networks for efficient pedestrian detection," 2018, *arXiv:1805.08688*.
[28] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2019.
[29] P. Yang, G. Zhang, L. Wang, L. Xu, Q. Deng, and M.-H. Yang, "A part-aware multi-scale fully convolutional network for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1125–1137, Feb. 2021.
[30] T. Zhang, Z. Han, H. Xu, B. Zhang, and Q. Ye, "CircleNet: Reciprocating feature adaptation for robust pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4593–4604, Nov. 2020.
[31] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9556–9565.
[32] T. Liu, W. Luo, L. Ma, J.-J. Huang, T. Stathaki, and T. Dai, "Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling," *IEEE Trans. Image Process.*, vol. 30, pp. 754–766, 2021.
[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 657–674, Oct. 2018.
[34] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 618–634.

Ameen Abdelmutalab received the B.Sc. degree from the University of Khartoum, Sudan, in 2011, and the M.Sc. degree from the American University of Sharjah, United Arab Emirates, in 2015. He is currently pursuing the Ph.D. degree with Concordia University, Montreal, QC, Canada. His research interests include machine-learning applications, image and video processing wireless communications, and cognitive radio.

Chunyan Wang received the B.Eng. degree in electronics from Shanghai Jiao Tong University, Shanghai, China, and the M.Eng. and Ph.D. degrees from Université Paris-Sud, Paris, France. She is currently a Professor of electrical and computer engineering with Concordia University, Montreal, QC, Canada. Her current research interests include digital image processing and VLSI circuits for signal processing.