Mask-Guided Attention Network and Occlusion-Sensitive Hard Example Mining for Occluded Pedestrian Detection

Jin Xie, *Graduate Student Member, IEEE*, Yanwei Pang, *Senior Member, IEEE*, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, *Member, IEEE*, and Ling Shao, *Senior Member, IEEE*

***Abstract*— Pedestrian detection relying on deep convolution neural networks has made significant progress. Though promising results have been achieved on standard pedestrians, the performance on heavily occluded pedestrians remains far from satisfactory. The main culprits are intra-class occlusions involving other pedestrians and inter-class occlusions caused by other objects, such as cars and bicycles. These result in a multitude of occlusion patterns. We propose an approach for occluded pedestrian detection with the following contributions. First, we introduce a novel mask-guided attention network that fits naturally into popular pedestrian detection pipelines. Our attention network emphasizes on visible pedestrian regions while suppressing the occluded ones by modulating full body features. Second, we propose the occlusion-sensitive hard example mining method and occlusion-sensitive loss that mines hard samples according to the occlusion level and assigns higher weights to the detection errors occurring at highly occluded pedestrians. Third, we empirically demonstrate that weak box-based segmentation annotations provide reasonable approximation to their dense pixel-wise counterparts. Experiments are performed on CityPersons, Caltech and ETH datasets. Our approach sets a new state-of-the-art on all three datasets. Our approach obtains an absolute gain of 10.3% in log-average miss rate, compared with the best reported results on the heavily occluded HO pedestrian set of the CityPersons test set. Code and models are available at: https://github.com/Leotju/MGAN.**

***Index Terms*— Pedestrian detection, attention, convolutional neural networks, hard example mining.**

I. Introduction

PEDESTRIAN detection [1]–[3] is a challenging computer vision problem with numerous real-world applications. Recently, deep convolutional neural networks (CNNs) have pervaded many areas of computer vision ranging from object recognition [4], [5], to generic object detection [6]–[8], to pedestrian detection [9]–[18]. Despite the recent progress on standard benchmarks with non-occluded or rarely occluded pedestrians, the state-of-the-art approaches still struggle under severe occlusions. For example, when walking in proximity, a pedestrian is likely to be obstructed by other pedestrians and/or other objects like cars and bicycles. For illustration, Fig. 1 displays the performance of the baseline Faster R-CNN pedestrian detector [6] under heavy occlusions. Handling occlusions is a key challenge; they frequently occur in real-world applications of pedestrian detection. Therefore, recent benchmarks specifically focus on heavily occluded pedestrian detection. For instance, the CityPersons [19] dataset has around 70% of pedestrians depicting various degrees of occlusions.

Most existing approaches employ a holistic detection strategy [9]–[12] that assumes entirely visible pedestrians when trained using full body annotations. However, such a strategy is sub-optimal under partial or heavy occlusions since most of the pedestrian's body is invisible. This deteriorates the performance by degrading the discriminative ability of the pedestrian model due to the inclusion of background regions inside the full body detection window.

Lately, several pedestrian detection methods [20]–[23] tackle occlusions by learning a series of part detectors that are integrated to detect partially occluded pedestrians. They either learn an ensemble model and integrate their outputs or jointly train different occlusion patterns to handle occlusions. Ensemble-based approaches are computationally expensive which prohibits real-time detection. The parts used in these methods are usually designed manually, which is not necessary to be optimal. Instead, the visible regions in our method are predicted without employing specific hand-crafted parts.

In contrast to part-based approaches for handling occlusions, a few methods [24], [25] exploit visible-region information, available with standard pedestrian detection benchmarks [19], [26], to either output visible part regions for proposal generation [25] or employ as extraneous supervision to learn occlusion patterns [24]. In this work, we follow the footsteps of these recent methods to tackle the problem of occluded detection. Different to [24], [25], we make use of visible body information to produce a spatial attention to modulate the multichannel convolutional features in the standard full

Manuscript received May 20, 2020; revised October 5, 2020; accepted November 10, 2020. Date of publication December 4, 2020; date of current version March 29, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102800; in part by the National Natural Science Foundation of China under Grant 61632018; in part by the MBZUAI Starting Grant GR010, Grant GR008, and Grant GR007; and in part by the VR Starting Grant 2016-05543. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Husrev T. Sencar. *(Corresponding author: Yanwei Pang.)*

Jin Xie and Yanwei Pang are with the Tianjin Key Laboratory of Brain-Inspired Intelligence Technology, School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: jinxie@tju.edu.cn; pyw@tju.edu.cn).

Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao are with the Computer Vision Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: muhammad.haris@mbzuai.ac.ae; rao.anwer@mbzuai.ac.ae; fahad.khan@mbzuai.ac.ae; ling.shao@mbzuai.ac.ae).

Digital Object Identifier 10.1109/TIP.2020.3040854

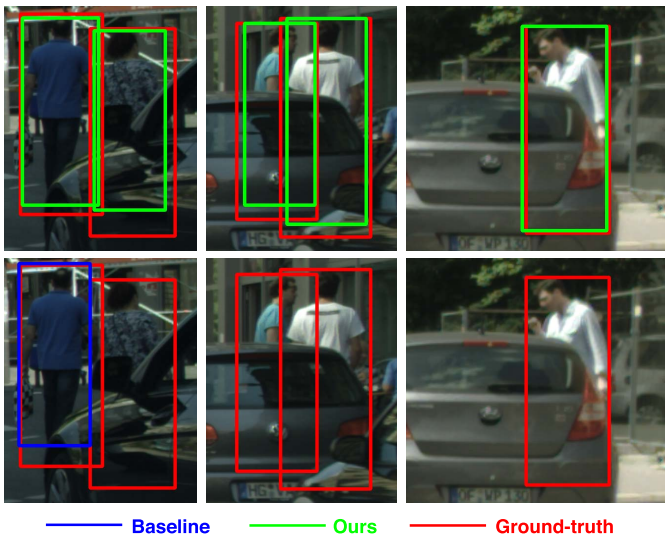


Fig. 1. Detection examples using our approach (top row) and the baseline Faster R-CNN [6] (bottom row). For improved visualization, detection regions are cropped from images of the CityPersons dataset [19]. All results are obtained using the same false positive per image (FPPI) criterion. Our approach robustly handles occlusions.

body estimation branch. The proposed mask-guided spatial attention network can be easily integrated into mainstream pedestrian detectors and is not limited to specific occlusion patterns. Fig. 1 shows that the proposed approach is able to detect occluded pedestrians over a wide spectrum ranging from partial to heavy occlusions.

When training the pedestrian detector, hard examples can help improve the pedestrian detection performance. Most of these existing pedestrian detection methods select examples randomly. However, the most of selected samples by random sampling scheme are easy samples. The random sampling would limit the detection performance. In object detection, some hard example mining methods [8], [27], [28] are proposed to solve the problem of random sampling scheme. OHEM [27] automatically selects hard examples according to the losses. Focal Loss [8] reshapes standard cross entropy loss to focus training on hard examples. Libra RCNN [28] selects hard examples according to their Intersection over Union (IoU). Different to these methods, we propose the occlusion-sensitive hard example mining method and the occlusion-sensitive loss by introducing occlusion levels into sampling procedure and loss function. In addition, it can be easily applied in any two-stage pedestrian detectors to improve the occluded pedestrian detection performance.

In summary, the main contributions of this article lie in:

- We propose a deep architecture termed as mask-guided attention network (MGAN), which comprises two branches: the standard pedestrian detection branch and a novel mask-guided attention branch. The standard pedestrian detection branch generates features using full body annotations for supervision. The proposed mask-guided attention branch produces a spatial attention map using visible-region information, thereby highlighting the visible body region while suppressing the occluded part of the pedestrian. The spatial attention map is then deployed to modulate the standard full body features by emphasizing regions likely belonging to visible part of the pedestrian.
- We propose the occlusion-sensitive hard example mining method which introduces occlusion levels into sampling procedure to improve the detection performance of occluded pedestrians. And the occlusion-sensitives loss is proposed to alleviate classification and regression problems in occluded pedestrian detection.
- We empirically demonstrate that for occluded pedestrian detection, the weak approximation of dense pixel-wise annotations yields similar results.

We perform experiments on three pedestrian detection benchmarks: CityPersons [19], Caltech [26] and ETH [29]. On all datasets, our approach displays superior results compared with the existing pedestrian detection methods. Further, our approach improves the state-of-the-art [25] from 44.2% to 37.2% in log-average miss rate on the **HO** set of the CityPersons validation set, which has 35-80% occluded pedestrians, using the *same* level of supervision, input scale and backbone network.

Preliminary results of this work have been published in [30]. This article has been improved and extended to the conference version as follows.

- The occlusion-sensitive hard example mining is proposed to mine hard examples during training.
- The occlusion-sensitive regression loss function is proposed to alleviate the inaccurate localization problem of occluded pedestrians.
- More detailed ablation studies are conducted to analyze the effectiveness of our approach.
- Experiment results on the ETH dataset [29] are shown to demonstrate the generalization capability of our method.

II. Related Work

Pedestrian detection is an important but challenging computer vision task and the prerequisite as the basis for some popular computer vision tasks [31]–[34]. In this section, we first review recent deep neural networks based pedestrian detectors and then discuss relevant occlusion handling methods for pedestrian detection.

A. Deep Pedestrian Detection

Recently, pedestrian detection approaches based on deep learning techniques exhibited the state-of-the-art performance [9]–[12], [15], [16], [35], [36]. CNN-based detectors can be roughly divided into two categories: the two-stage approach comprising separate proposal generation followed by confidence computation and bounding-box regression of proposals and the single-stage approach regressing and classifying default anchors directly. Most existing pedestrian detection methods either employ the single-stage [9], [10], [37] or two-stage strategy [11], [12], [15], [16], [38] as their backbone architecture. The work of [9] proposed a recurrent rolling convolution architecture that aggregates useful contextual information among the feature maps to improve single-stage

detectors. Liu *et al.* [10] extended the single-stage architecture with an asymptotic localization fitting module storing multiple predictors to evolve default anchor boxes. This improves the quality of positive samples while enables hard negative mining with increased thresholds.

In the two-stage detection strategy, the work of [11] proposed a deep multi-scale detection approach where intermediate network layers, with receptive fields similar to different object scales, are employed to perform the detection task. Mao *et al.* [12] proposed to integrate channel features (*i.e.*, edge, heatmap, optical flow and disparity) into a two-stage deep pedestrian detector. The work of [15] introduced a multi-task approach for joint supervision of pedestrian detection and semantic segmentation. The work of [39] employed a two-stage pretrained person detector (Faster R-CNN) and an instance segmentation model for person re-identification. Each detected person is cropped out from the original image and fed to another network. Wang *et al.* [16] introduced repulsion losses that prevent a predicted bounding-box from shifting to neighboring overlapped objects to counter occlusions. Due to their superior performance on the pedestrian benchmarks [19], we deploy the two-stage detection strategy as backbone pipeline in our work.

B. Occlusion Handling in Pedestrian Detection

Several works [40]–[42] investigated the problem of handling occlusions in pedestrian detection. A common strategy [20]–[23] is the part-based approach where a set of part detectors are learned with each part designed to handle a specific occlusion pattern. Some of these part-based approaches [21], [22] train an ensemble model for most occurring occlusion patterns and are computationally expensive due to the deployment of large number of part detectors. Alternatively, some part-based approaches [20], [23] rely on joint learning of collection of parts to capture occlusion patterns.

Contrary to the aforementioned methods, recent approaches exploited visible body information either as an explicit branch to regress visible part regions for proposal generation [25] or as external guidance to learn specific occlusion modes (full, upper-body, left-body and right-body visible) in a supervised fashion [24]. Different to [25], we utilize the visible branch to generate a spatial attention map that is used to modulate multichannel convolutional features in the standard full body estimation branch. Unlike ATT-vbb [24], we propose a spatial attention network that is not restricted to only certain type of occlusion patterns. Further, when using the *same* level of supervision, input scale, backbone and training data, our proposed approach significantly reduces the log-average miss rate by 7.0% and 7.8% compared with [25] and [24], respectively on the **HO** set of the CityPersons validation set.

III. Proposed Approach

We propose a mask-guided attention network (MGAN) that features a novel mask-guided attention branch. It produces a spatial attention map, highlighting the visible body part while suppressing the occluded part in the full body features. This branch is a lightweight, easy to implement module and can be easily integrated into the standard pedestrian pipeline, thereby making a single, coherent architecture capable of end-to-end training. In addition, we propose the occlusion-sensitive hard example mining method and the occlusion-sensitive loss which introduce occlusion levels into sampling procedure and detection loss computing. It is integrated seamlessly into the two-stage pedestrian detectors to improve the detection performance of occluded pedestrians.

The overall proposed architecture comprises two main branches: a **standard pedestrian detector** (SPD) branch that detects pedestrian [6] using full body information whom components are shown in blue in Fig. 2, and a novel mask-guided attention (MGA) branch that produces a spatial attention map employing visible bounding-box information. This branch modulates the full body features and shown with a red dashed box in Fig. 2.

Next, we review the SPD branch in Sec. III-A, detail the design of our MGA branch in Sec. III-B, explain the design of our proposed occlusion-sensitive hard example mining method in Sec. III-C, and detail the overall loss function in Sec. III-D.

A. Standard Pedestrian Detector Branch

We choose Faster R-CNN [6] as the standard pedestrian detection branch mainly for its state-of-the-art performance. It takes a raw image as input, first deploys a ImageNet [43] pretrained model such as VGG-16 [4] to extract features, and then utilizes a region proposal network (RPN) to generate region proposals. Extracts proposal features by cropping the corresponding region-of-interest (RoI) in the extracted feature maps and further resizes them to fixed dimensions with a RoI pooling layer. Note, we replace RoI pooling layer with RoI Align layer [44] in our experiments. This makes every proposal to have the same feature length. These features go through a classification net that generates the classification score (*i.e.* the probability that this proposal contains a pedestrian) and the regressed bounding box coordinates for every proposal. Fig. 2 visually illustrates the aforementioned steps. Since every layer in Faster R-CNN is differentiable, it is trainable end-to-end with the following loss function:

$$L_0 = L_{rpn} + L_{rcnn}. \tag{1}$$

Each term has a classification loss and a bounding box regression loss. Thus, Eq. 1 can be written as:

$$L_0 = L_{rpn\_cls} + L_{rpn\_reg} + L_{rcnn\_cls} + L_{rcnn\_reg}, \tag{2}$$

where L_{rpn_cls} and L_{rcnn_cls} refer to the classification loss of RPN and R-CNN, respectively, and L_{rpn_reg} and L_{rcnn_reg} are the bounding box regression loss of RPN and R-CNN, respectively. Here, the classification loss is Cross-Entropy loss and the bounding box regression loss is Smooth-L1 loss.

Discussion: Despite achieving impressive results for non-occluded pedestrians, the standard pedestrian detector struggles - showing high miss rates - in the presence of partial and heavy occlusions. Fig. 3 depicts pedestrian detector trained using full body bounding-box annotations produces less false positives but miss several pedestrians. This is likely due to the contribution of features towards the scoring of a proposal

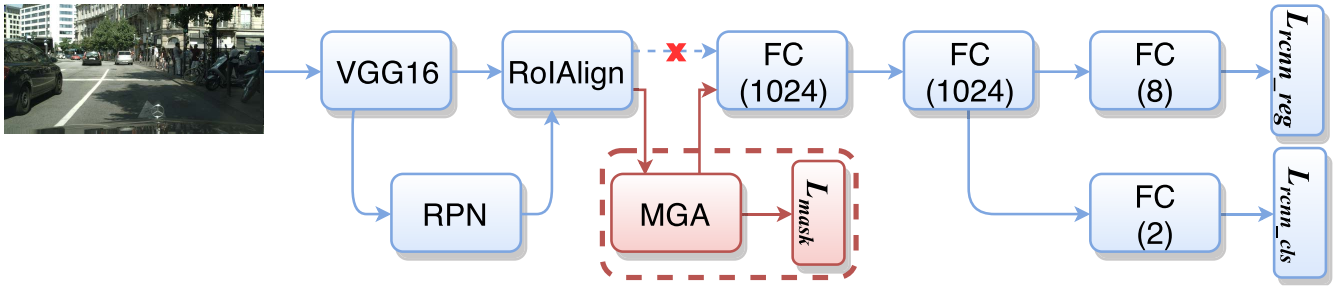


Fig. 2. The overall network architecture of our mask-guided attention network (MGAN). It comprises a standard pedestrian detection (SPD) branch, whose components are shown in blue. It introduces a novel mask-guided attention (MGA) module enclosed in the red dashed box. Note, after RoI Align there is a classification stage in the SPD branch whose first layer is shown by FC (1024). In our architecture, the standard full body features in SPD branch after RoI Align layer are modulated by the MGA branch before getting scored by the classification stage. This is in contrast to the baseline SPD where these features directly become the input to the classification stage without any modulation.

Fig. 3. Results of a pedestrian detector trained by full body bounding-box annotations. We show three different occluded scenarios. Solid green boxes denote predictions by the detector and dashed green boxes represent the missed detection results. The detector cannot capture heavily occluded pedestrians and might result in high miss rates under similar circumstances.

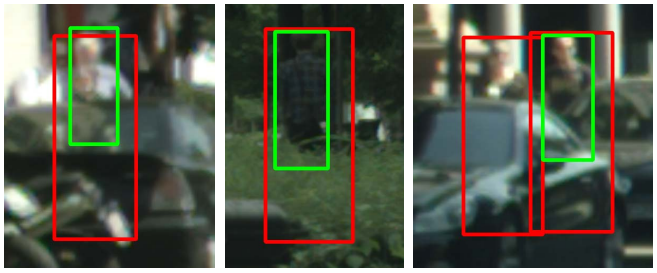


Fig. 4. Detection examples of the standard pedestrian detector (SPD). Red boxes denote the ground-truth bounding-box, and green boxes represent the SPD predictions. The SPD might result in inaccurate locations and high miss rates for occluded pedestrians.

corresponding to the occluded parts of the pedestrian. As the occlusion modifies the pedestrian appearance, the features for the occluded part are vastly different to the visible part. We show how to suppress these (occluded) features and enhance the visible ones to obtain more robust features .[1] We present a **mask-guided spatial attention** approach that greatly alleviates the impact of occluded features while stresses the visible-region features, and is not restricted to certain occlusion types. This mask-guided attention network is a lightweight branch integrated into the standard pedestrian detection network.

Secondly, the standard pedestrian detector treats each sample as equal during training and ignores occlusion level of different samples. It would limit the detection performance of occluded pedestrians. To counter this issue, we propose occlusion-sensitive hard example mining by introducing occlusion level into sampling procedure.

At last, the standard pedestrian detector struggles on classification and localization problems (shown in Fig. 4). We propose the occlusion-sensitive loss by introducing occlusion level into loss computing to improve classification problem and alleviate the inaccurate localization problem.

B. Mask-Guided Attention Branch

The proposed mask-guided attention branch is highlighted with the red annotated box in Fig. 2. It produces a spatial attention mask supervised by visible-region bounding box information and using this modulates the multichannel features generated by the RoI Align layer. Fig. 5 shows three different occluded pedestrians and their corresponding spatial attention masks. These masks accurately reveal the visible part and hide the occluded part for three variable occlusion patterns. The modulated features with these masks help classification network detect partially and heavily occluded pedestrians with higher confidence, which otherwise might not get detected due to being scored poorly. The following subsections detail our mask-guided attention branch.

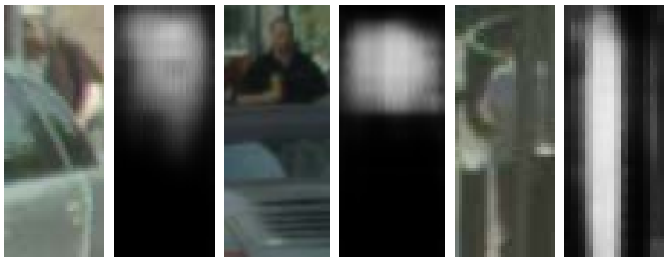


Fig. 5. Spatial attention masks generated by our MGA branch. Three spatial attention masks correspond to differently occluded pedestrians *i.e.* partial and heavy. Note the enhancing of visible part and the hiding of occluded part in each mask.

1) MGA Architecture: The proposed MGA branch architecture is depicted in Fig. 6. The input to MGA branch are the multichannel features from RoI Align layer and the output are the modulated multichannel features. The modulated features

[1]One might argue that a simple solution can be to train a pedestrian detector supervised only by visible-region annotations. Though the resulting detector will capture occluded pedestrians and will decrease the miss rates, it would result in high false positive detection results.

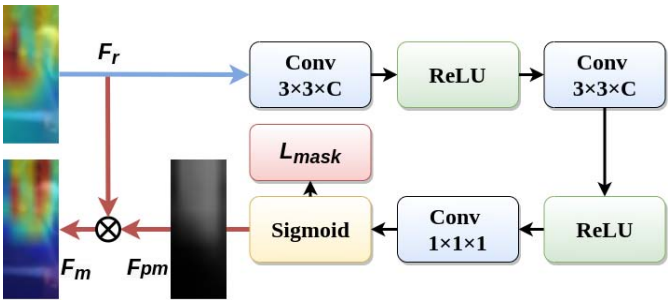


Fig. 6. The network architecture of our mask-guided attention (MGA) Branch. It takes the RoI features and generates the modulated features using a small stack of conv. operations.

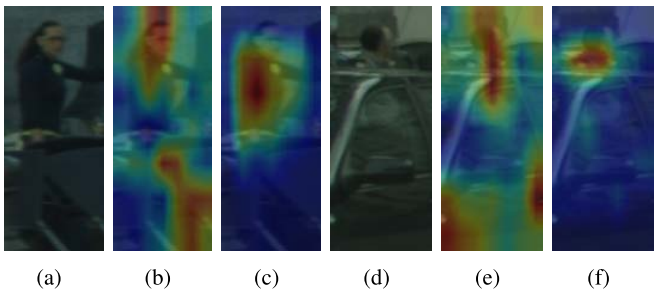


Fig. 7. Visual comparison of the RoI features and the corresponding modulated features. (a) and (d) are two different proposals. (b) and (e) depict their corresponding RoI features. (c) and (f) show their corresponding modulated features. In contrast to the RoI features, the modulated features from our MGA branch have visible region signified and occluded part concealed.

are generated using pedestrian probability map, termed as the spatial attention mask. We denote the input features as $F_r \in [H \times W \times C]$, where the first two dimensions are the resolution and the last one is the depth. Firstly, two 3×3 filter size convolution layer followed by Rectified Linear Unit (ReLU) extracts features. Then, a 1×1 filter size conv. layer followed by a sigmoid layer generates the probability map $F_{pm} \in [H \times W \times 1]$. In our experiments, H and W are set to 7, and C is set to 512.

These probability maps F_{pm} modulate the multichannel features F_r of a proposal to obtain the re-weighted features F_m. We achieve this by taking the element-wise product of every feature channel in F_r with F_{pm} as:

$$F_{m_i} = F_{r_i} \odot F_{pm}, \quad i = 1, 2, \ldots, C, \tag{3}$$

where i is the channel index and $\odot$ is the element-wise product. Instead of the RoI features F_r, we feed the modulated features F_m to the classification net for scoring proposals. Fig. 7 illustrates that in contrast to the RoI features, the modulated features from MGA branch have visible region signified and occluded part concealed thereby leading to a relatively high confidence for occluded proposals.

2) Weak Box-Based Segmentation Annotation: The spatial attention masks for a proposal and image-level segmentation requires supervision in the form of dense pixel-wise segmentation annotation. This, however, is tedious to acquire in many computer vision tasks including pedestrian detection. We therefore adapt visible-region bounding box annotation as an approximate alternative. Such annotations are readily available for the popular pedestrian detection benchmarks [19], [26].

The adaption is as follows. If a pixel lies in the visible-region bounding-box annotation; it is a foreground pixel with a label one. Similarly, a pixel outside this region is a background pixel and its label is zero. This labelling process creates a weak box-based segmentation annotation. Importantly, such weakly labelled annotations have generated accurate masks in our experiments (see Fig. 5). Description of MGA branch finishes here and the following subsection discusses the occlusion-sensitive hard example mining and the loss function.

C. Occlusion-Sensitive Hard Example Mining

The standard pedestrian detectors (SPD) usually utilized random sampling procedure. Suppose we need to sample N_{ps} positive proposals from N_p corresponding candidate positive proposals. The selected probability pb_i for i-th positive sample under random sampling procedure can be computed as $pb_i = \frac{N_{ps}}{N_p}$. To raise the selected probability of occluded samples, the proposed occlusion-sensitive hard example mining (OSEM) introduces occlusion ratios into sampling procedure. The selected probability under our OSEM can be computed as $pb_i = (1 + \lambda \cdot ocr_i)\frac{N_{ps}}{N_p}$, where ocr_i is the occlusion ratio of i-th positive sample, and λ is a balancing factor, empirically set to 1 and fixed for all datasets. It can be observed that, if the occlusion level of a sample is higher, the selected probability will be larger. This implies that our OSEM would sample more positive proposals from heavily occluded pedestrians.

D. Loss Function

Here, we present our loss function of the proposed approach. The overall loss formulation L is:

$$L = L_0 + \alpha L_{mask} + \beta L_{occ}, \tag{4}$$

where L_0 is the loss term for Faster R-CNN as in Eq.(1), L_{mask} is the loss term for the proposed MGA branch, and L_{occ} is the occlusion-sensitive loss term. Note that we tend to jointly optimize all the losses in the spirit of end-to-end training. In our experiments, we set $\alpha = 0.5, \beta = 1$ by default. L_{mask} on weak box-based supervision is formulated as a binary cross-entropy loss (BCE loss):

$$L_{mask} = \frac{1}{NHW} \sum_{i}^{N} \sum_{y}^{H} \sum_{x}^{W} BCELoss(p_i^M(x, y), \hat{p}_i^M(x, y)), \tag{5}$$

where $\hat{p}_i^M(x, y)$ is the predictions produced by MGA branch and $p_i^M(x, y)$ represents the ground truth. N represents the number of proposals. H and W are the height and width of the RoI features.

Further, to make the classification and regression loss aware of variable occlusion levels, we introduce an occlusion sensitive loss term L_{occ}. It simply weights pedestrian proposals

based on their occlusion ratios. L_{occ} can be computed as:

$$\begin{aligned} L_{occ} &= L_{occ\_cls} + L_{occ\_reg}. \\ &= \frac{\mu_{cls}}{N_{ps}} \sum_{i}^{N_{ps}} ocr_i \cdot CELoss(p_i^{rcnn\_cls}, \hat{p}_i^{rcnn\_cls}) \\ &\quad + \frac{\mu_{reg}}{N_{ps}} \sum_{i}^{N_{ps}} ocr_i \cdot SmoothL1(p_i^{rcnn\_reg}, \hat{p}_i^{rcnn\_reg}), \end{aligned} \tag{6}$$

where L_{occ_cls} and L_{occ_reg} are the occlusion-sensitive classification loss and the occlusion-sensitive regression loss respectively, μ_{cls} and μ_{reg} are balancing factors and empirically set to 1, ocr_i is the occlusion ratio, N_{ps} is the number of positive samples, $\hat{p}_i^{rcnn\_cls}$ are the predictions produced by the classification branch of RCNN, $p_i^{rcnn\_cls}$ represents the ground-truth, and $CELoss$ represents Cross-Entropy loss. $\hat{p}_i^{rcnn\_reg}$ are the predictions produced by the regression branch of RCNN, $p_i^{rcnn\_reg}$ represents the ground-truth, and $SmoothL1$ represents Smooth-L1 loss proposed in [45]. L_{occ} are defined on positive proposals. It can be observed that, the occlusion level of a sample is higher, the classification and regression loss values are larger. It would help improve the classification problem of occluded pedestrians and alleviate the inaccurate localization problems of occluded pedestrians.

IV. Experiments

In this section, we denote the proposed method in the preliminary version [30] as MGAN. And we denote MGAN with the proposed occlusion-sensitive regression loss and occlusion-sensitive hard example mining as MGAN+.

A. Datasets and Evaluation Metrics

1) Datasets: We perform experiments on three pedestrian detection benchmarks: CityPersons [19], Caltech [26], and ETH [29]. CityPersons [19] is a challenging dataset for pedestrian detection and exhibits large diversity. It consists of 2975 training images, 500 validation images, and 1525 test images. Caltech is a popular pedestrian dataset [26] featuring 11 sets of videos. First 6 sets (0-5) correspond to training and the last 5 sets (6-10) are for testing. To increase training set size, the frames are sampled at 10Hz. The test images are captured at 1 Hz. Finally, the training and test sets have 42782 and 4024 images, respectively. Both the CityPersons and Caltech datasets provide bounding box annotations for full body and visible region. ETH [29] is also a popular pedestrian dataset which consists of three different sequences. And there are 1804 images in total.

2) Evaluation Metrics: We report the detection performance using the standard log-average miss rate in experiments; it is computed over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ [26]. We select MR^{-2} as evaluation metrics. And its lower value reflects better detection performance. On the Caltech dataset, we report results across three different occlusion degrees: Reasonable (**R**), Heavy (**HO**) and the combined Reasonable+Heavy (**R+HO**). For the CityPersons dataset, we follow [19] and report results on the Reasonable (**R**) and the Heavy (**HO**) sets. The visibility ratio in the **R** set is larger than 65%, and the visibility ratio in the **HO** set ranges from 20% to 65%. Similarly, the visibility ratio in the **R+HO** set is larger than 20%. In all subsets, the height of pedestrians over 50 pixels is taken for evaluation, as in [24]. Note that the **HO** set is designed to evaluate performance in the case of severe occlusions.

B. Implementation and Training Details

We first explain the general settings common to both the CityPersons and Caltech datasets. And then detail implementation the settings specific to the CityPersons and Caltech datasets.

For both datasets, the networks are trained on NVIDIA GPUs and a mini-batch comprises 2 images per GPU. We select the Adam [46] solver as optimizer. For a fair comparison with other methods, we perform image horizontal flipping only for data augmentation.

We then detail settings specific to the two datasets.

CityPersons: We fine-tune the ImageNet pretrained VGG-16 [4] models on the CityPersons training set. Except we use two fully-connected layers with 1024 output dimensions instead of 4096 output dimensions, we follow the same experimental protocol as in [19]. We start with the initial learning rate of 1×10^{-4} for the first 8 epochs and further decay it to 1×10^{-5} and perform 3 epochs.

Caltech: We start with the model pretrained on the CityPersons dataset. To fine-tune the model, an initial learning rate of 10^{-4} is used for first 3 training epochs. The training is further performed for another 1 epoch after decaying the initial learning rate by a factor of 10.

C. Ablation Study

We evaluate our approach by performing an ablation study on the CityPersons dataset.

1) Baseline Comparison: Tab. I shows the baseline comparison. For a fair comparison, we use the same set of ground-truth pedestrian examples during training for all methods. We select ground-truth pedestrian examples which are at least 50 pixels tall with visibility $\geq$ 65% for the training purpose. The baseline SPD (first row) detector obtains a log-average miss rate of 13.8% and 57.0% on the **R** and **HO** sets of the CityPersons dataset, respectively. Our MGAN (third row) based on the SPD with the MGA branch and occlusion-sensitive classification loss (OSL(Cls)) term significantly reduces the error on both the **R** and **HO** sets. On the **HO** set, our MGAN achieves an absolute reduction of 5.3% in a log-average miss rate, compared with the baseline. Our MGAN+ (last row) based on our MGAN with the occlusion-sensitive regression loss (OSL(Reg)) and the occlusion-sensitive hard example mining (OSEM) reduces the miss rates on both the **R** and **HO** sets further. On the **HO** set, our MGAN+ achieves an absolute reduction of 7.5% and 2.2% in log-average miss rates, compared with the baseline SPD and our MGAN, respectively. The significant reduction

TABLE I

COMPARISON (IN LOG-AVERAGE MISS RATES (%)) OF OUR METHOD WITH THE BASELINE ON THE CITYPERSONS VALIDATION SETS. WE SHOW THE PERFORMANCE OF OUR MGAN (THIRD ROW) AND OUR MGAN+ (LAST ROW). FOR A FAIR COMPARISON, WE USE THE SAME TRAINING DATA, INPUT SCALE (×1) AND NETWORK BACKBONE (VGG-16). SPD: STANDARD PEDESTRIAN DETECTOR, MGA: MASK-GUIDED ATTENTION BRANCH, OSL(CLS): OCCLUSION-SENSITIVE CLASSIFICATION LOSS, OSL(REG): OCCLUSION-SENSITIVE REGRESSION LOSS. OSEM: OCCLUSION-SENSITIVE HARD EXAMPLE MINING. ON THE HEAVY OCCLUSION SET (**HO**), OUR DETECTOR SIGNIFICANTLY REDUCES THE ERROR FROM 57.0% TO 49.5%, COMPARED WITH THE BASELINE. THE BEST RESULTS ARE IN BOLD

SPD	MGA	OSL(Cls)	OSL(Reg)	OSEM	**R**	**HO**
✓					13.8	57.0
✓	✓				11.9	52.7
✓	✓	✓			11.5	51.7
✓	✓		✓		11.4	51.2
✓	✓	✓	✓		11.2	50.5
✓	✓	✓	✓	✓	**11.0**	**49.5**

TABLE II

COMPARISON (IN LOG-AVERAGE MISS RATES (%)) OF OUR MGAN DETECTOR WHEN USING DENSE PIXEL-WISE LABELING WITH WEAK BOX-BASED SEGMENTATION OBTAINED THROUGH VISIBLE BOUNDING BOX INFORMATION IN OUR MGA BRANCH. REPLACING FORMER WITH LATTER IN OUR MGA BRANCH RESULTS IN NO SIGNIFICANT DETERIORATION IN DETECTION PERFORMANCE. ON BOTH SETS, OUR APPROACH BASED ON WEAK BOX-BASED SEGMENTATION PROVIDES A TRADE-OFF BETWEEN ANNOTATION COST AND ACCURACY

Set	Dense Pixel-wise Annotations	Weak Box-based Annotations
R	11.2	11.9
HO	51.7	52.7

TABLE III

COMPARISON (IN LOG-AVERAGE MISS RATES (%)) BY DIVIDING PEDESTRIANS W.R.T. THEIR HEIGHT (PIXELS): SMALL [50-75], MEDIUM [75-125] AND LARGE (>125) REPRESENTING 28%, 37% AND 35%, RESPECTIVELY OF THE CITYPERSONS **HO** SET. THE BEST RESULTS ARE BOLDFACED IN EACH CASE

Method	[50, 75]	[75, 125]	>125
Baseline SPD	66.3	59.7	43.1
MGAN	**61.7**	**52.3**	**37.6**

in error on the **HO** set demonstrates the effectiveness of our MGAN and MGAN+ against the baseline SPD.

2) Comparison With Other Attention Strategies: We compare our approach with other attention strategies proposed by [24]. The work of [24] investigates channel attention (CA), visible box attention (CA-VBB) and part attention (CA-Part). Both CA and CA-VBB exploit channel-wise attention, with the latter also using VBB information. In addition, CA-Part utilizes a part detection network pretrained on MPII Pose dataset. In contrast to CA-Part, our method does not require extra annotations for part detection.

We perform an experiment integrating the CA and CA-VBB attention strategies [24] in our framework. On the **R** and **HO** sets of the CityPersons validation set, the CA attention strategy achieves a log-average miss rate of 17.3% and 54.5%, respectively. The CA-VBB attention scheme obtains a log-average miss rate of 14.0% and 54.1% on the **R** and **HO** sets, respectively. Our approach with our MGA branch outperforms both the CA and CA-VBB strategies on both the **R** and **HO** sets by achieving a log-average miss rate of 11.9% and 52.7%, respectively.

3) Impact of Weak Box-Based Segmentation: As discussed in Sec. III-B.2, dense pixel-wise labelling is expensive to acquire. Further, such dense annotations are only available for the CityPersons dataset and not for the Caltech dataset. We validate our approach using weak box-based segmentation and compare it with using dense pixel-wise labelling in Tab. II. On both sets, similar results are obtained with the coarse level information and dense pixel-wise labelling in our MGA branch. Our results in Tab. II are also aligned to the prior work in instance segmentation [47]. Further, our final output is a detection box which does not require a precise segmentation mask prediction as in [47]. In addition, the difference between the two set of annotations is likely to reduce further for small pedestrians due to high-level of pooling operations undertaken by the network (*i.e.*, we use RoI features from conv5_3 of VGG). Our approach therefore provides a trade-off between annotation cost and accuracy.

4) Heavy Occlusion and Size Variation: We also evaluate the effectiveness of our MGAN on heavily occluded pedestrians with varying sizes, especially small pedestrians. Tab. III shows that our approach provides improvement for all cases with a notable gain of 4.6% for the small sized (50-75 pixels tall) heavily occluded pedestrians, compared with the baseline.

5) Comparison With Other Hard Example Mining Methods: We perform an experiment comparing our occlusion-sensitive hard example mining (OSEM) and occlusion-sensitive loss (OSL) with the existing hard example mining methods (*i. e.* OHEM [27] and Focal Loss [8]). For a fair comparison, the experiment is performed by integrating all the methods into the same baseline (SPD+MGA). Tab. IV reports the results. Compared with the OHEM and Focal Loss, our method (OSEM+OSL) significantly reduces the log-average miss rates by 2.3% and 3.0% on the **HO** set of the CityPersons validation set, respectively.

6) Robust to Various Backbone Networks: We also evaluate the effect of backbone networks to our MGA branch in terms of log-average miss rates. Two different backbone networks including VGG16 [4] and ResNet50 [5] are chosen to conduct experiments. For each backbone network, the same training data (visibility is larger than 65%) and the same input scale (1.0×) are used in training. For ResNet50, we remove down-sampling operation of block 4 and replace the general convolution layer of block 4 with the dilated convolution layer and dilated factor of 2. For VGG16, our MGA branch achieves an absolute gain of 1.9% and 4.3% over the baseline SPD on the **R** and **HO** set, respectively. For ResNet50, our MGA branch

Fig. 8. Detection examples on the CityPersons validation dataset using our proposed pedestrian detector. The ground-truth and our detector predictions are shown in red and green respectively. Our detector accurately detects pedestrians under partial and heavy occlusions.

TABLE IV

COMPARISON (IN LOG-AVERAGE MISS RATES (%)) WITH OTHER HARD EXAMPLE MINING METHODS. FOR A FAIR COMPARISON, WE USE THE SAME PEDESTRIAN DETECTOR (STANDARD PEDESTRIAN DETECTOR WITH MASK-GUIDED ATTENTION BRANCH) FOR ALL EXPERIMENTS. THE BEST RESULTS ARE BOLDFACE. OUR METHOD OUTPERFORMS THESE METHODS ON BOTH THE **R** AND **HO** SETS. SPD: STANDARD PEDESTRIAN DETECTOR, MGA: MASK-GUIDED ATTENTION BRANCH, OHEM [27]: ONLINE HARD EXAMPLE MINING, FL [8]: FOCAL LOSS, OSEM: OUR PROPOSED OCCLUSION-SENSITIVE HARD EXAMPLE MINING, OSL: OUR OCCLUSION-SENSITIVE LOSS

Hard Example Mining Methods	**R**	**HO**
Baseline (SPD+MGA)	11.9	52.7
FL [8]	11.8	52.5
OHEM [27]	11.6	51.8
Ours (OSEM+OSL)	11.0	49.5

TABLE V

COMPARISON OF THE PROPOSED MGAN+ WITH THE BASELINE SPD IN TERMS OF NUMBER OF NETWORK PARAMETERS AND TEST TIME

Method	# Parameters(M)	Test time(Second)	MR^{-2}	
			R	**HO**
Bseline SPD	43.9	0.13	13.8	57.0
MGAN+	48.6	0.15	11.0	49.5

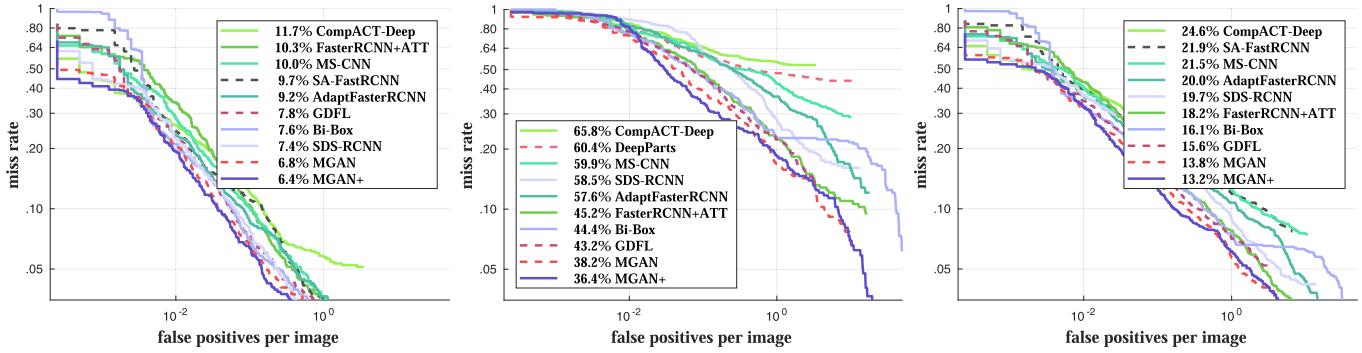
achieves an absolute gain of 1.5% and 4.2% over the baseline SPD on the **R** and **HO** set, respectively. It can be observed that our proposed MGA branch can be successfully used with various backbone networks.

7) Model Parameters and Test Time: We also evaluate the model size and test time of our proposed MGAN+. The results are reported in Tab. V. It can be observed that compared with the standard pedestrian detector (SPD), the proposed MGAN+ has merely additional 4.7M parameters overhead. For a fair comparison, the test time of both the SPD and our MGAN+ is measured on a single NVIDIA V100 GPU. For a 1024×2048 input, compared with the baseline SPD, our MGAN+ only has addition 0.02s for one image on the CityPersons validation set. Although the test time of our MGAN+ is slightly slower than the baseline SPD, our MGAN+ significantly reduces the log-average miss rate from 57.0% to 49.5% on the **HO** set.

D. State-of-the-Art Comparison on CityPersons

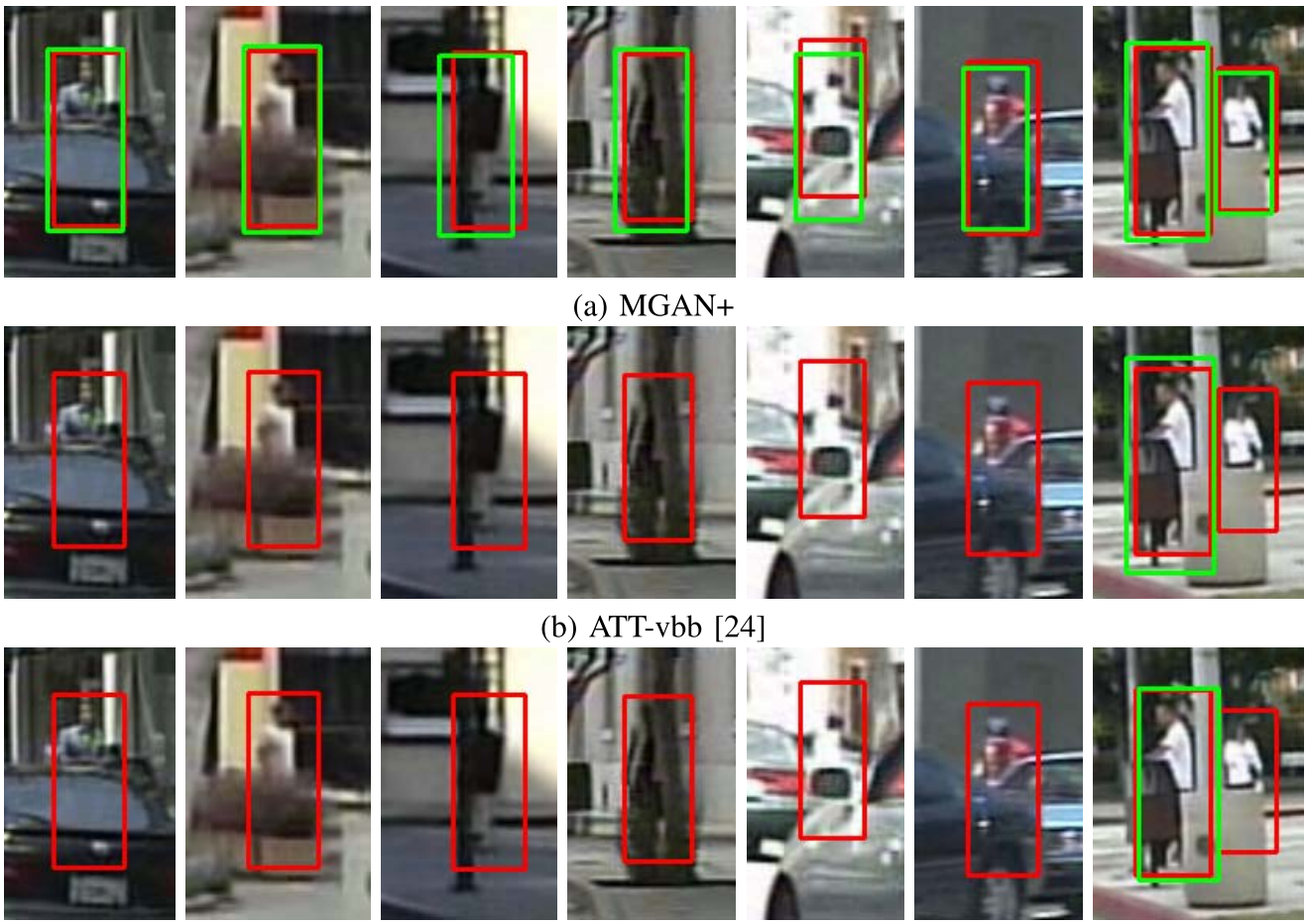
Our proposed MGAN and MGAN+ are compared with the recent state-of-the-art methods, namely Repulsion Loss [16], ATT-part [24], ALFNet [10], OR-CNN [48], TLL [50], Bi-Box [25], FRCN+A+DT [49] on the CityPersons validation set. It is worth mentioning that existing pedestrian detection methods employ different set of ground-truth pedestrian examples for training. We therefore select the same set of ground-truth pedestrian examples and input scale when comparing with each state-of-the-art method.

Among existing methods, ATT-vbb [24], OR-CNN [48], Bi-Box [25] and FRCN+A+DT [49] employ both the visible bounding box (VBB) and full body information similar to our method. We therefore first compare our approach with these four methods. Tab. VI shows the comparison in terms of the log-average miss rates on the **R** and **HO** sets of the CityPersons validation dataset. Our MGAN and MGAN+ outperform all four methods on both the **R** and **HO** sets.



(a) **R** (b) **HO** (c) **R+HO**

Fig. 9. The state-of-the-art comparison on the **R**, **HO** and **R+HO** subsets of the Caltech dataset. The legend in each plot represents the log-average miss rates over FPPI= $[10^{-2}, 10^{0}]$. Our approach provides superior results compared with the existing approaches on all three subsets.



(a) MGAN+

(b) ATT-vbb [24]

(c) GDFL [53]

Fig. 10. Qualitative comparison of (a) MGAN+ with (b) ATT-vbb [24] and (c) GDFL [53] on the Caltech test set. Red boxes denote the ground-truth and green boxes indicate detector predictions. The example images depict varying level of occlusions.

When using an input scale of 1×, the OR-CNN method [48] employs both full body and visible region information and enforces the pedestrian proposals to be close and compactly located to corresponding objects, achieves a log-average miss rate of 12.8% and 55.7% on the **R** and **HO** sets, respectively. The proposed MGAN and MGAN+ achieve an absolute gain of 8.5% and 11.0% on the **HO** set, compared with OR-CNN. The detection results of OR-CNN [48] are improved when using an input scale of 1.3×. Our proposed MGAN and MGAN+ outperform OR-CNN with a significant margin on both input scales. For an input scale of 1×, the ATT-vbb approach [24] employing Faster RCNN detector with a visible bounding box channel attention network obtains a log-average miss rate 16.4% and 57.3% on the **R** and **HO** sets, respectively. Our MGAN provides superior detection results with a log-average miss rate of 11.5% and 51.7% on the **R** and **HO** sets, respectively. Further, our MGAN+ obtains the best detection results with a log-average miss rate of 11.0%

TABLE VI

COMPARISON (IN TERMS OF LOG-AVERAGE MISS RATES(%)) WITH THE STATE-OF-THE-ART METHODS THAT USE BOTH THE VISIBLE BOUNDING BOX (VBB) *and* FULL BODY INFORMATION ON THE CITYPERSONS VALIDATION SET. FOR A FAIR COMPARISON, WE USE THE SAME SET OF GROUND-TRUTH PEDESTRIAN EXAMPLES (VISIBILITY) AND INPUT SCALE FOR TRAINING WHEN COMPARING WITH EACH METHOD. OUR MGAN AND MGAN+ OUTPERFORMS ALL FOUR METHODS ON BOTH SETS. UNDER HEAVY OCCLUSIONS (**HO**), OUR MGAN+ SIGNIFICANTLY REDUCES THE ERROR FROM 44.2% TO 37.2%, COMPARED WITH THE RECENTLY INTRODUCED BI-BOX [25]. THE BEST RESULTS ARE BOLDFACED IN EACH CASE

Method	VBB	Backbone	Data (visibility)	Scale	**R**	**HO**
OR-CNN [48]	✓	VGG	$\geq 50\%$	$\times 1$	12.8	55.7
MGAN	✓	VGG	$\geq 50\%$	$\times 1$	10.5	47.2
MGAN+	✓	VGG	$\geq 50\%$	$\times 1$	**10.2**	**44.7**
OR-CNN [48]	✓	VGG	$\geq 50\%$	$\times 1.3$	11.0	51.3
MGAN	✓	VGG	$\geq 50\%$	$\times 1.3$	9.9	45.4
MGAN+	✓	VGG	$\geq 50\%$	$\times 1.3$	**9.7**	**43.1**
ATT-vbb [24]	✓	VGG	$\geq 65\%$	$\times 1$	16.4	57.3
MGAN	✓	VGG	$\geq 65\%$	$\times 1$	11.5	51.7
MGAN+	✓	VGG	$\geq 65\%$	$\times 1$	**11.0**	**49.5**
Bi-Box [25]	✓	VGG	$\geq 30\%$	$\times 1.3$	11.2	44.2
FRCN+A+DT [49]	✓	VGG	$\geq 30\%$	$\times 1.3$	11.1	44.3
MGAN	✓	VGG	$\geq 30\%$	$\times 1.3$	10.5	39.4
MGAN+	✓	VGG	$\geq 30\%$	$\times 1.3$	**10.3**	**37.2**

TABLE VII

COMPARISON (IN TERMS OF LOG-AVERAGE MISS RATES(%)) OF OUR MGAN AND MGAN+ WITH THE STATE-OF-THE-ART METHODS IN LITERATURE ON THE CITYPERSONS VALIDATION SET. OUR MGAN+ SETS A NEW STATE-OF-THE-ART BY OUTPERFORMING ALL EXISTING METHODS. THE BEST RESULTS ARE BOLDFACED IN EACH CASE

Method	Data (visibility)	Scale	**R**	**HO**
TLL [50]	-	$\times 1$	14.4	52.0
ATT-part [24]	$\geq 65\%$	$\times 1$	16.0	56.7
Rep. Loss [16]		$\times 1$	13.2	56.9
Adaptive-NMS [51]		$\times 1$	11.9	55.2
MGAN		$\times 1$	11.5	51.7
MGAN+		$\times 1$	**11.0**	**49.5**
OR-CNN [48]	$\geq 50\%$	$\times 1$	12.8	55.7
MGAN		$\times 1$	10.5	47.2
MGAN+		$\times 1$	**10.2**	**44.7**
ALFNet [10]	$\geq 0\%$	$\times 1$	12.0	51.9
CSP [52]		$\times 1$	11.0	49.3
MGAN		$\times 1$	11.3	42.0
MGAN+		$\times 1$	**11.0**	**39.7**
Rep. Loss [16]	$\geq 65\%$	$\times 1.3$	11.6	55.3
Adaptive-NMS [51]		$\times 1.3$	10.8	54.0
MGAN		$\times 1.3$	10.3	49.6
MGAN+		$\times 1.3$	**10.2**	**47.2**
OR-CNN [48]	$\geq 50\%$	$\times 1.3$	11.0	51.3
MGAN		$\times 1.3$	9.9	45.4
MGAN+		$\times 1.3$	**9.7**	**43.1**
Bi-Box [25]	$\geq 30\%$	$\times 1.3$	11.2	44.2
FRCN+A+DT [49]		$\times 1.3$	11.1	44.3
MGAN		$\times 1.3$	10.5	39.4
MGAN+		$\times 1.3$	**10.3**	**37.2**

and 49.5% on the **R** and **HO** sets. The Bi-Box method [25] utilizes visible bounding box (VBB) information to generate visible part regions for pedestrian proposal generation. On the **R** and **HO** sets, when using an input scale of $1.3\times$, the Bi-Box approach [25] yields a log-average miss rate of 11.2% and 44.2%, respectively. Our MGAN+ outperforms Bi-Box on both sets by achieving a log-average miss rate of 10.3% and 37.2%, respectively. Moreover, the recently introduced FRCN+A+DT approach [49] employing Fast RCNN detector with a discriminative feature transformation obtains a log-average miss rate of 11.1% and 44.3%. The proposed MGAN and MGAN+ achieve an absolute gain of 4.9% and 7.1% on the **HO** set, compared with FRCN+A+DT. To summarize, the results in Tab. VI clearly signify the effectiveness of our MGAN and MGAN+ towards handling heavy occlusions (**HO**) compared with these methods [24], [25], [48], [49] using the *same* level of supervision, ground-truth pedestrian examples during training, input scale and backbone.

TABLE VIII

COMPARISON (IN TERMS OF LOG-AVERAGE MISS RATES(%)) OF MGAN WITH STATE-OF-THE-ART METHODS ON THE CITYPERSONS TEST SET. THE TEST SET IS WITHHELD AND RESULTS ARE OBTAINED BY SENDING OUR DETECTION PREDICTIONS TO THE AUTHORS OF THE CITYPERSONS DATASET [19] FOR EVALUATION

Method	**R**	**HO**
MS-CNN [54]	13.2	51.9
Adaptive Faster RCNN [19]	13.0	50.5
Rep. Loss [16]	11.5	52.6
OR-CNN [48]	11.3	51.4
Adaptive NMS [51]	11.4	47.0
Cascade MS-CNN [54]	11.6	47.1
MGAN	9.3	41.0
MGAN+	**9.3**	**36.7**

Tab. VII further shows the comparison with the published state-of-the-art methods on the CityPersons validation set. The best reported result for the **HO** set is 44.2%, in terms of a log-average miss rate, obtained by the recently introduced Bi-Box [25] with an input scale of $1.3\times$. Our MGAN+ sets a new state-of-the-art on the **HO** set with a log-average miss rate of 37.2%. Our detector also outperforms existing methods on the **R** set.

Fig. 8 displays detection examples from our MGAN+ on CityPersons. Examples show a range of occlusion degrees *i.e.* from partial to heavy.

Finally, Tab. VIII shows the state-of-the-art comparison on the CityPersons test set. Note that the test set is withheld and the results are obtained by sending our detector predictions to the authors of CityPersons [19]. Our proposed MGAN and MGAN+ outperform all reported methods on both the **R** and **HO** sets of the test set. On the heavy occlusion set, our proposed MGAN+ outperforms the existing best performing method Adaptive NMS [51] that applies a dynamic suppression threshold and learns density scores by significant margin after reducing the log-average miss rate by 10.3%.

E. Caltech Dataset

Here, our MGAN and MGAN+ are compared with the following recent state-of-art methods: CompACT-Deep [55], DeepParts [22], MS-CNN [11], RPN+BF [35], SA-FastRCNN [57], MCF [56], SDS-RCNN [15], Adapt-FasterRCNN [19], F-DNN+SS [36], FasterRCNN+ATT-vbb [24], GDFL [53], Bi-Box [25], AR-Ped [58], and FRCN+A+DT [49]. Tab. IX compares our MGAN and

TABLE IX

COMPARISON (IN TERMS OF LOG-AVERAGE MISS RATES(%)) OF OUR MGAN AND MGAN+ WITH THE STATE-OF-ART METHODS ON THE CALTECH DATASET. THE SECOND COLUMN INDICATES WHETHER THE METHOD IS SPECIFICALLY TARGETED TO HANDLING OCCLUSION. THE BEST RESULTS ARE IN BOLD. FURTHER, OUR DETECTOR PROVIDES SUPERIOR RESULTS COMPARED WITH ALL PUBLISHED METHODS ON THE REASONABLE SET (**R**), THE HEAVY OCCLUSIONS SET (**HO**) AND THE COMBINED SET OF REASONABLE AND HEAVY OCCLUSIONS (**R+HO**)

Detector	Occl.	**R**	**HO**	**R+HO**
DeepParts [22]	✓	11.9	60.4	22.8
CompACT-Deep [55]	×	11.8	65.8	24.6
MCF [56]	×	10.4	66.7	22.9
FasterRCNN+ATT-vbb [24]	✓	10.3	45.2	18.2
MS-CNN [11]	×	10.0	59.9	21.5
SA-FastRCNN [57]	×	9.7	64.4	21.9
RPN+BF [35]	×	9.6	74.4	24.0
AdaptFasterRCNN [19]	×	9.2	57.6	20.0
F-DNN+SS [36]	×	8.2	53.8	18.8
FRCN+A+DT [49]	✓	8.0	37.9	-
GDFL [53]	×	7.9	43.2	15.6
Bi-Box [25]	✓	7.6	44.4	16.1
SDS-RCNN [15]	×	7.4	58.6	19.7
AR-Ped [58]	×	6.5	48.8	16.1
MGAN	✓	6.8	38.2	13.8
MGAN+	✓	**6.4**	**36.4**	**13.2**

MGAN+ with the state-of-the-art methods under all three occlusion subsets: **R**, **HO** and **R+HO**. The AR-Ped detector [58] reports the best results among existing methods with a log-average miss rate of 6.5% on the **R** set. Our MGAN+ achieves superior results with a log-average miss rate of 6.4%. Note that our MGAN+ approach outperforms the AR-Ped detector with an absolute gain of 12.4% on the **HO** set. Among existing methods, the FRCN+A+DT approach [49] reports a log-average miss rate of 37.9% on the **HO** set. Our MGAN+ achieves superior results with a log-average miss rate of 36.4% on this set. On the **R+HO** set, the GDFL detector [53] provides the best results among the existing methods with a log-average miss rate of 15.6%. Our MGAN+ detector outperforms GDFL with an absolute gain of 2.4% on **R+HO** set. Fig. 9 shows the comparison of our MGAN and MGAN+ with existing methods over the whole spectrum of false positives per image metric.

We further signify the effectiveness of MGAN+ towards handling occlusions by drawing visual comparison with ATT-vbb [24], and GDFL [53] in Fig. 10. All results are obtained using the same FPPI. ATT-vbb captures pedestrians using channel-wise attention and occlusion patterns. GDFL introduces graininess-aware deep features designed to enhance local details and context information. Our MGAN+ accurately detects pedestrians in all scenarios.

F. ETH Dataset

To investigate the generalization ability of our proposed method, we apply the model trained on the CityPersons dataset on the ETH dataset. Because there are no visible bonding-box annotations in the ETH dataset, we report the log-average miss rate on all occlusion level pedestrians. Our MGAN+ is compared with the following state-of-the-art methods: LDCF [59], SpatialPooling [60], FasterFCNN [24], TA-CNN [61], RPN+BF [35], FasterRCNN+ATT-part [24], F-DNN2+SS [62] and OR-CNN [48]. Tab. X shows comparison of our MGAN+ with the state-of-the-art methods on the ETH dataset. Our MGAN+ sets a new state-of-the-art by outperforming the existing best method (OR-CNN [48]) after significantly reducing the log-average miss rate by 1.5%. The results on the ETH datasets verify the generalization ability of our MGAN+.

TABLE X

COMPARISON (IN TERMS OF LOG-AVERAGE MISS RATES(%)) OF MGAN+ WITH THE STATE-OF-THE-ART METHODS ON THE ETH DATASET. THE BEST RESULTS ARE IN BOLD. OUR DETECTOR PROVIDES SUPERIOR RESULTS COMPARED WITH ALL PUBLISHED METHODS)

Method	
LDCF [59]	45.0
SpatialPooling [60]	37.4
FasterRCNN [24]	35.6
TA-CNN [61]	35.0
FasterRCNN+ATT-part [24]	33.8
RPN+BF [35]	30.2
F-DNN2+SS [62]	30.0
OR-CNN [48]	24.5
MGAN+	**23.0**

V. CONCLUSION

We proposed a mask-guided attention network (MGAN) for occluded pedestrian detection. The mask-guided attention (MGA) module generates spatial attention mask using visible body region information. The resulting spatial attention mask modulates the full body features (*i.e.*, highlighting the features of pedestrian visible region, and suppressing the background). Instead of dense pixel labelling, we employ weak box-based segmentation information for visible regions. In addition to MGA, we introduced the occlusion-sensitive hard example mining (OSEM) and the occlusion-sensitive loss (OSL). Experiments on three datasets clearly show the effectiveness of our approach, especially for heavily occluded pedestrians.

REFERENCES

[1] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3820–3834, 2020.

[2] W. Liu, S. Liao, and W. Hu, "Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding," *IEEE Trans. Image Process.*, vol. 29, pp. 1413–1425, 2020.

[3] J. Cao, Y. Pang, J. Han, B. Gao, and X. Li, "Taking a look at small-scale pedestrians and occluded pedestrians," *IEEE Trans. Image Process.*, vol. 29, pp. 3143–3152, 2020.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[5] 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.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2015, pp. 91–99.

[7] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[9] J. Ren *et al.*, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5420–5428.

[10] 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.*, 2018, pp. 618–634.

[11] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.

[12] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3127–3136.

[13] J. Xie, Y. Pang, H. Cholakkal, R. Anwer, F. Khan, and L. Shao, "PSC-net: Learning part spatial co-occurrence for occluded pedestrian detection," *Sci. China Inf. Sci.*, vol. 64, no. 2, Feb. 2021, Art. no. 120103.

[14] J. Wu, C. Zhou, Q. Zhang, M. Yang, and J. Yuan, "Self-mimic learning for small-scale pedestrian detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2012–2020.

[15] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4950–4959.

[16] 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.

[17] J. Xie *et al.*, "Count-and similarity-aware R-CNN for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–16.

[18] J. Wu, C. Zhou, M. Yang, Q. Zhang, Y. Li, and J. Yuan, "Temporal-context enhanced detection of heavily occluded pedestrians," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13430–13439.

[19] 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. 3213–3221.

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

[21] 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.

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

[23] 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. 3486–3495.

[24] 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.

[25] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–151.

[26] P. Dollar, 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.

[27] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[28] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830.

[29] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[30] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4967–4975.

[31] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3633–3645, Aug. 2014.

[32] F. S. Khan, J. Xu, J. van de Weijer, A. D. Bagdanov, R. M. Anwer, and A. M. Lopez, "Recognizing actions through action-specific person detection," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4422–4432, Nov. 2015.

[33] T. Wang *et al.*, "Deep contextual attention for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5694–5702.

[34] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8679–8687.

[35] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–457.

[36] 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.*, Mar. 2017, pp. 953–961.

[37] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 966–974.

[38] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "PedHunter: Occlusion robust pedestrian detector in crowded scenes," in *Proc. AAAI*, 2020, pp. 10639–10646.

[39] D. Chen, S. Zhang, W. Ouyang, J. Yang1, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[40] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3124–3129.

[41] W. Ouyang, X. Zeng, and X. Wang, "Single-pedestrian detection aided by two-pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1875–1889, Sep. 2015.

[42] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1874–1887, Aug. 2018.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[44] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[45] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[47] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1635–1643.

[48] 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.*, 2018, pp. 637–653.

[49] 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. 9557–9566.

[50] 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.*, 2018, pp. 536–551.

[51] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6459–6468.

[52] 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. 5187–5196.

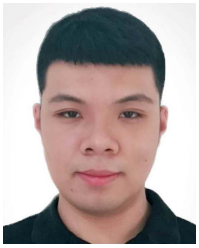
[53] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 732–747.

[54] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," 2019, *arXiv:1906.09756*. [Online]. Available: http://arxiv.org/abs/1906.09756

[55] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3361–3369.

[56] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3210–3220, Jul. 2017.

[57] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018, doi: 10.1109/TMM.2017.2759508.
[58] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7231–7240.
[59] J. H. H. Woonhyun Nam and P. Dollár, "Local decorrelation for improved pedestrian detection," in *Proc. Adv. Neural Inform. Process. Syst.*, 2014, pp. 424–432.
[60] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 546–561.
[61] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5079–5087.
[62] 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*. [Online]. Available: http://arxiv.org/abs/1805.08688



Jin Xie (Graduate Student Member, IEEE) received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2016, where he is currently pursuing the Ph.D. degree under the supervision of Prof. Y. Pang. His research interests include machine learning and computer vision.

Yanwei Pang (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China in 2004. He is currently a Professor with Tianjin University, China, and also the Founding Director of the Tianjin Key Laboratory of Brain Inspired Intelligence Technology (BIIT), China. His research interests include object detection and image recognition, in which he has published 150 scientific articles, including 40 IEEE TRANSACTIONS articles and 30 top conferences (e.g., CVPR, ICCV, and ECCV) papers. He is an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (T-NNLS) and *Neural Networks* (Elsevier) and a Guest Editor of *Pattern Recognition Letters*.

Muhammad Haris Khan received the Ph.D. degree in computer vision from the University of Nottingham, U.K. He is currently a Faculty Member with the Mohamed Bin Zayed University of Artificial Intelligence, United Arab Emirates. Prior to MBZUAI, he was a Research Scientist with the Inception Institute of Artificial Intelligence, United Arab Emirates. He also stayed as a Postdoctoral Fellow with the University of Nottingham. He has published several papers in top-ranked computer vision conferences. His research interests include computer vision and machine/deep learning.

Rao Muhammad Anwer is currently an Assistant Professor with Computer Vision Department, Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI). Prior to joining MBZUAI, he was Research Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. Before joining IIAI, he was a Postdoctoral Research Fellow with Aalto University, Finland. His research interests include computer vision, including object detection and segmentation.

Fahad Shahbaz Khan (Member, IEEE) received the M.Sc. degree in intelligent systems design from the Chalmers University of Technology, Sweden, and the Ph.D. degree in computer vision from the Autonomous University of Barcelona, Spain. He is currently a Faculty Member with the MBZ University of AI (MBZUAI), United Arab Emirates, and Linkoping University, Sweden. Prior to joining MBZUAI, he has worked as a Lead Scientist with the Inception Institute of Artificial Intelligence (IIAI), United Arab Emirates. His research interests include computer vision and machine learning, such as object recognition, object detection, action recognition, and visual tracking. He has published articles in high-impact computer vision journals and conferences in these areas. He has achieved top ranks on various international challenges (Visual Object Tracking VOT: 1st 2014 and 2018, 2nd 2015, 1st 2016; VOT-TIR: 1st 2015 and 2016; OpenCV Tracking: 1st 2015; and 1st PASCAL VOC 2010) and the Best Paper Award at ICPR 2016.

Ling Shao (Senior Member, IEEE) received the B.Eng. degree in electronic and information engineering from the University of Science and Technology of China (USTC), and the M.Sc. degree in medical image analysis and the Ph.D. degree in computer vision from the Robotics Research Group, University of Oxford. He was the Chair Professor and the Director of the Artificial Intelligence Laboratory, University of East Anglia, Norwich, U.K. He is currently the CEO of the Inception Institute of Artificial Intelligence and the Executive Vice President and a Provost of the Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, deep learning/machine learning, multimedia, and image/video processing. He has published over 300 articles at top venues such as TPAMI, TIP, IJCV, ICCV, CVPR, and ECCV.