Mask-Guided Attention Network for Occluded Pedestrian Detection

Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shabbaz Khan, Ling Shao

1Tianjin University 2Inception Institute of Artificial Intelligence, UAE 3CVL, Linköping University, Sweden

{pyw,jinxie}@tju.edu.cn, {muhammad.haris,rao.anwer,fahad.khan,ling.shao}@inceptioniai.org

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 empirically demonstrate that coarse-level segmentation annotations provide reasonable approximation to their dense pixel-wise counterparts. Experiments are performed on CityPersons and Caltech datasets. Our approach sets a new state-of-the-art on both datasets. Our approach obtains an absolute gain of 9.5% in log-average miss rate, compared to the best reported results [31] on the heavily occluded HO pedestrian set of CityPersons test set. Further, on the HO pedestrian set of Caltech dataset, our method achieves an absolute gain of 5.0% in log-average miss rate, compared to the best reported results [13]. Code and models are available at: https://github.com/Leotju/MGAN.

1. Introduction

Pedestrian detection 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 [25, 10, 27, 22], to generic object detection [24, 15, 14], to pedestrian detection [23, 16, 2, 17, 1, 29].

Despite the recent progress on standard benchmarks with non-occluded or reasonably occluded pedestrians, state-of-the-art approaches still struggle under severe occlusions.

For example, when walking in close 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 baseline Faster R-CNN [24] under heavy occlusions. Handling occlusions is a key challenge; they occur frequently in real-world applications of pedestrian detection. Therefore, recent benchmarks specifically focus on heavily occluded pedestrian detection. For instance, CityPersons [31] dataset has around 70% of pedestrians depicting various degrees of occlusions.

Most existing approaches employ a holistic detection strategy [23, 16, 2, 17] 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, 18, 28,
35, 21] 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. On the other hand, methods based on joint learning of occlusion patterns are difficult to train and rely on fusion of part detection scores. Instead, we investigate occluded pedestrian detection without explicitly using part information.

In contrast to part-based approaches for handling occlusions, a few methods [33, 36] exploit visible-region information, available with standard pedestrian detection benchmarks [31, 7], to either output visible part regions for proposal generation [36] or employ as extraneous supervision to learn occlusion patterns [33]. In this work, we follow the footsteps of these recent methods to tackle the problem of occluded detection. Different to [36, 33], we make use of visible body information to produce a pixel-wise spatial attention to modulate the multichannel features in the standard full 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.

**Contributions:** 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 pixel-wise 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. Further, we empirically demonstrate that for occluded pedestrian detection, the weak approximation of dense pixel-wise annotations yields similar results.

We perform experiments on two pedestrian detection benchmarks: CityPersons [31] and Caltech [7]. On both datasets, our approach displays superior results compared to existing pedestrian detection methods. Further, our approach improves the state-of-the-art [36] from 44.2 to 39.4 in log-average miss rate on the HO set of CityPersons, which has 35-80% occluded pedestrians, using the same level of supervision, input scale and backbone network.

## 2. Related Work

**Deep Pedestrian Detection.** Recently, pedestrian detection approaches based on deep learning techniques have exhibited state-of-the-art performance [23, 16, 2, 17, 1, 29, 30, 8]. CNN-based detectors can be roughly divided into two categories: the two-stage approach comprising separate proposal generation followed by confidence computation of proposals and the one-stage approach where proposal generation and classification are formulated as a single-stage regression problem. Most existing pedestrian detection methods either employ the single-stage [23, 16, 19] or two-stage strategy [2, 17, 1, 29] as their backbone architecture. The work of [23] proposed a recurrent convolution architecture that aggregates useful contextual information among the feature maps to improve single-stage detectors. Liu et al. [16] 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 [2] 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. [17] proposed to integrate channel features (i.e., edge, heatmap, optical flow and disparity) into a two-stage deep pedestrian detector. The work of [1] introduced a multi-task approach for joint supervision of pedestrian detection and semantic segmentation. The segmentation fusion layer is employed to highlight pedestrians, thereby enabling downstream detection easier. The work of [5] employed a two-stage pre-trained 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. [29] introduced repulsion losses that prevent a predicted bounding-box from shifting to neighbouring overlapped objects to counter occlusions. Due to their superior performance on pedestrian benchmarks [31], we deploy two-stage detection strategy as backbone pipeline in our work.

**Occlusion Handling in Pedestrian Detection.** Several works investigated the problem of handling occlusions in pedestrian detection. A common strategy [18, 34, 20, 28, 35] 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 [18, 28] 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, 35] rely on joint learning of collection of parts to capture occlusion patterns.

Contrary to the aforementioned methods, recent approaches have exploited visible body information either as an explicit branch to regress visible part regions for proposal generation [36] or as external guidance to learn specific occlusion modes (full, upper-body, left-body and right-
We present an architecture that can be written as: \( \text{cls} \) depicts pedestrian detector and \( \text{reg} \) components are shown in blue in Fig. 3.1. Standard Pedestrian Detector Branch

This branch modulates the full body features and shown in the standard full body estimation branch. It takes a raw image as input, first deploys a pre-trained ImageNet model such as VGG-16 \[25\] as the standard pedestrian detector will capture occluded pedestrians and will decrease the miss rate, it would result in high false positive detections.

**3. Proposed Approach**

We propose Mask-Guided Attention Network (MGAN) that features a novel Mask-Guided Attention branch. It produces a pixel-wise attention map, highlighting the visible body part while suppressing the occluded part in the full body features. This branch is lightweight, easy to implement module and is integrated into the standard pedestrian pipeline, thereby making a single, coherent architecture capable of end-to-end training.

The overall proposed architecture comprises two main branches: a Standard Pedestrian Detector (SPD) branch that detects pedestrian \[24\] using full body information whom components are shown in blue in Fig. 2, and a novel Mask-Guided Attention (MGA) branch that produces a pixel-wise 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 and then detail the design of our MGA branch.

**3.1. Standard Pedestrian Detector Branch**

We choose Faster R-CNN\[24\] as the standard pedestrian detection branch mainly for its state-of-the-art performance. It takes a raw image as input, first deploys a pre-trained ImageNet model such as VGG-16 \[25\] and then 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 \[9\] in our experiments. This makes every proposal to have 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, 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, this and similar pipelines struggle - 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 corresponding to the occluded parts of the pedestrian. As the occlusion modifies the pedestrian appearance, the features for the occluding 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 \footnote{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 rate, it would result in high false positive detections.}. We present a mask-guided spatial attention...
tion 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 CNN branch integrated into the standard pedestrian detection network.

3.2. Mask-Guided Attention Branch

The proposed mask-guided attention branch is highlighted with 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. 4 shows three different occluded persons and their corresponding spatial attention masks. These masks accurately reveal the visible part and hide the occluded part for three variable occlusion patterns. 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.

3.2.1 MGA Architecture

The proposed MGA branch architecture is depicted in Fig. 5. The input to MGA branch are the multichannel features from RoI Align layer and the output are the modulated multichannel features. The modulated features are generated using pedestrian probability map, termed as the spatial attention mask. We denote input features as $F_r \in [H \times W \times C]$, where first two dimensions are the resolution and the last one is the depth. Firstly, two $3 \times 3$ filter size convolution layer followed by Rectified Linear Unit (ReLU) extracts features. Then, a $1 \times 1$ 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 re-weighted features $F_{r,m}$.

We achieve this by taking the element-wise product of every feature channel in $F_r$ with $F_{pm}$ as:

$$F_{r,m} = F_{r,i} \odot F_{pm}, i = 1, 2, ..., C,$$

(3)

where $i$ is the channel index and $\odot$ is the element-wise product. Instead of RoI features $F_r$, we feed modulated features $F_{r,m}$ to the classification net for scoring proposals. Fig. 6 illustrates that in contrast to RoI features, modulated features from MGA branch have visible region signified and occluded part concealed thereby leading to a relatively high confidence for occluded proposals.

3.2.2 Coarse-level Segmentation Annotation

The spatial attention mask 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 [31, 7].

The adaptation 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 coarse-level segmentation annotation. Importantly, such weakly labelled annotations have generated accurate masks in our experiments (see Fig. 4).

Description of MGA branch finishes here and the following subsection discusses the loss function optimized in proposed approach.
3.2.3 Loss Function

Here, we present our loss function for the proposed architecture MGAN. The overall loss formulation $L$ is:

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

(4)

where $L_0$ is the loss 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. $L_{mask}$ and $L_{occ}$ are defined on positive proposals. $L_{mask}$ on coarse-level (weak) supervision is formulated as a per-pixel binary cross-entropy loss (BCE loss):

$$L_{mask} = BCELoss(p_n(x,y), \tilde{p}_n(x,y)),$$

(5)

where $\tilde{p}_n(x,y)$ are the predictions produced by MGA branch and $p_n(x,y)$ represents the ground truth i.e., coarse-level segmentation annotations.

Further, to make the classification loss aware of variable occlusion levels, we introduce an occlusion sensitive loss term $L_{occ}$. It simply weights pedestrian training proposals based on their occlusion levels, derived from $p_n(x,y)$, when computing the standard cross-entropy loss (CE loss):

$$L_{occ} = \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{1}{WH} \right\} \sum_{x} \sum_{y} p_n(x,y) CELoss(p^{rcnn,cls}_n, \tilde{p}^{rcnn,cls}_n),$$

(6)

where $W$ and $H$ are the width and the height of pedestrian probability map. $\tilde{p}^{rcnn,cls}_n$ are the predictions produced by the classification branch of RCNN, and $p^{rcnn,cls}_n$ represents the ground-truth.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We perform experiments on two pedestrian detection benchmarks: CityPersons [31] and Caltech [7]. CityPersons [31] is a challenging dataset for pedestrian detection and exhibits large diversity. It consists of 2975 training images, 500 validation images, and 1575 test images. Caltech pedestrian is a popular dataset [7] 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 datasets provide box annotations for full body and visible region.

Evaluation Metrics. We report performance using standard average-log miss rate (MR) in experiments; it is computed over the false positive per image (FPPI) range of $[10^{-2}, 10^0]$ [7]. We select $MR^{-2}$ 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 [31] and report results on Reasonable (R) and Heavy (HO) sets. The visibility ratio in R set is larger than 65%, and the visibility ratio in HO set ranges from 20% to 65%. Similarly, the visibility ratio in R + HO set is larger than 20%. In all subsets, the height of pedestrians over 50 pixels is taken for evaluation, as in [33]. Note that the HO set is designed to evaluate performance in case of severe occlusions.

4.2. Implementation and Training Details

For both datasets, the networks are trained on a NVIDIA GPU and a mini-batch comprises 2 image per GPU. We select the Adam [11] solver as optimizer. We now detail settings specific to the two datasets.

CityPersons. We fine-tune the ImageNet pretrained VGG-16 [25] models on CityPersons trainset. Except we use two fully-connected layers with 1024 output dimensions instead of 4096 output dimensions, we follow the same experimental protocol as in [31]. We start with the initial learning rate of $1 \times 10^{-4}$ for the first 8 epochs and further decay it to $1 \times 10^{-5}$ and perform 3 epochs.

Caltech. We start with the model pretrained on 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.

4.3. Ablation Study on CityPersons Dataset

We evaluate our approach (MGAN) by performing an ablation study on CityPersons dataset.

Baseline Comparison. Tab. 1 shows the baseline compari-
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 ≥ 65% for the training purpose. The baseline SPD detector obtains a log-average miss rate of 13.8% and 57.0% on R and HO sets of CityPersons dataset, respectively. Our Final MGAN based on the MGA branch and occlusion-sensitive loss term significantly reduces the error on both R and HO sets. Under heavy occlusions (HO), our MGAN achieves an absolute reduction of 5.3% in log-average miss rate, compared to the baseline. The significant reduction in error on the (HO) set demonstrates the effectiveness of our MGAN against the baseline.

**Comparison with other attention strategies.** We compare our approach with attention strategies proposed by [33]. The work of [33] 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 pre-trained on MPII Pose dataset. In contrast to CA-Part, our method does not require extra annotations for part detection.

We perform an experiment integrating CA and CA-VBB attention strategies [33] in our framework. On the R and HO sets of CityPersons validation set, 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 without \( L_{occ} \) outperforms both CA and CA-VBB strategies on both R and HO sets by achieving a log-average miss rate of 11.9% and 52.7%, respectively.

**Impact of coarse-level segmentation.** As discussed in section 3.2.2, dense pixel-wise labelling is expensive to acquire. Further, such dense annotations are only available for CityPersons and not for Caltech dataset. We validate our approach using coarse-level segmentation and compare it with using dense pixel-wise labelling in Tab. 2. 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. 2 are also aligned to the prior work in instance segmentation [6]. Further, our final output is a detection box which does not require a precise segmentation mask prediction as in [6]. 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.

**Heavy Occlusion and Size Variation.** We also evaluate the effectiveness of our approach on heavily occluded pedestrians with varying sizes, especially small pedestrians. Tab. 3 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 to the baseline.

**4.4. State-of-the-art Comparison on CityPersons**

Our MGAN detector is compared to the recent state-of-the-art methods, namely Repulsion Loss [29], ATT-part [33], ALFNet [16], OR-CNN [32], TLL [26], Bi-Box [36].
on 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 [33], OR-CNN [32] and Bi-Box [36] employ both the visible bounding box (VBB) and full body information similar to our method. We therefore first compare our approach with these three methods. Tab. 4 shows the the comparison in terms of log average miss rate (MR) on the R and HO sets of CityPersons dataset. Our MGAN outperforms all three methods on both R and HO sets. When using an input scale of $1 \times 1$, the OR-CNN method [32] 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 detection results of OR-CNN [32] are improved when using an input scale of $1.3 \times 1$. Our MGAN detector outperforms OR-CNN with a significant margin on both input scales.

For an input scale of $1 \times 1$, the ATT-vbb approach [33] employing FasterRCNN detector with a visible bounding box channel attention net 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. Moreover, the recently introduced Bi-Box method [36] utilizes visible bounding box (VBB) information to generate visible part regions for pedestrian proposal generation. On the R and HO sets, the Bi-Box approach [36] yields a log-average miss rate of $11.2$ and $44.2$, respectively using an input scale of $1.3 \times 1$. Our MGAN outperforms Bi-Box on both sets by achieving a log-average miss rate of $10.5$ and $39.4$, respectively.

To summarize, the results in Tab. 4 clearly signify the effectiveness of our MGAN towards handling heavy occlusions (HO) compared to these methods [33, 32, 36] using same level of supervision, ground-truth pedestrian examples during training, input scale and backbone. Tab. 5 further shows the comparison with all published state-of-the-art methods on the CityPersons. Fig. 7 displays example det-

<table>
<thead>
<tr>
<th>Method</th>
<th>Data (visibility)</th>
<th>Scale</th>
<th>R</th>
<th>HO</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLL [26]</td>
<td>-</td>
<td>$\times1$</td>
<td>14.4</td>
<td>52.0</td>
</tr>
<tr>
<td>ATT-part [33]</td>
<td>$\geq 65%$</td>
<td>$\times1$</td>
<td>16.0</td>
<td>56.7</td>
</tr>
<tr>
<td>Rep. Loss [29]</td>
<td>$\times1$</td>
<td>13.2</td>
<td>56.9</td>
<td></td>
</tr>
<tr>
<td>MGAN</td>
<td>$\times1$</td>
<td>11.5</td>
<td>51.7</td>
<td></td>
</tr>
<tr>
<td>OR-CNN [32]</td>
<td>$\geq 50%$</td>
<td>$\times1$</td>
<td>12.8</td>
<td>55.7</td>
</tr>
<tr>
<td>MGAN</td>
<td>$\times1$</td>
<td>10.5</td>
<td>47.2</td>
<td></td>
</tr>
<tr>
<td>ALF [16]</td>
<td>$\geq 0%$</td>
<td>$\times1$</td>
<td>12.0</td>
<td>51.9</td>
</tr>
<tr>
<td>MGAN</td>
<td>$\times1$</td>
<td>11.3</td>
<td>42.0</td>
<td></td>
</tr>
<tr>
<td>Rep. Loss [29]</td>
<td>$\geq 65%$</td>
<td>$\times1.3$</td>
<td>11.6</td>
<td>55.3</td>
</tr>
<tr>
<td>MGAN</td>
<td>$\times1.3$</td>
<td>10.3</td>
<td>49.6</td>
<td></td>
</tr>
<tr>
<td>OR-CNN [32]</td>
<td>$\geq 50%$</td>
<td>$\times1.3$</td>
<td>11.0</td>
<td>51.3</td>
</tr>
<tr>
<td>MGAN</td>
<td>$\times1.3$</td>
<td>9.9</td>
<td>45.4</td>
<td></td>
</tr>
<tr>
<td>Bi-Box [36]</td>
<td>$\geq 30%$</td>
<td>$\times1.3$</td>
<td>11.2</td>
<td>44.2</td>
</tr>
<tr>
<td>MGAN</td>
<td>$\times1.3$</td>
<td>10.5</td>
<td>39.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Comparison (in terms of log-average miss rate) of MGAN with state-of-the-art methods in literature on CityPersons validation set. Our MGAN sets a new state-of-the-art by outperforming all existing methods. Best results are boldfaced in each case.

<table>
<thead>
<tr>
<th>Method</th>
<th>R</th>
<th>HO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive Faster RCNN [31]</td>
<td>12.97</td>
<td>50.47</td>
</tr>
<tr>
<td>Rep. Loss [29]</td>
<td>11.48</td>
<td>52.59</td>
</tr>
<tr>
<td>OR-CNN [32]</td>
<td>11.32</td>
<td>51.43</td>
</tr>
<tr>
<td>Our MGAN</td>
<td><strong>9.29</strong></td>
<td><strong>40.97</strong></td>
</tr>
</tbody>
</table>

Table 6. Comparison (in terms of log-average miss rate) of MGAN with state-of-the-art methods on CityPersons test set. The test set is withheld and results are obtained by sending our detection predictions to the authors of CityPersons dataset [31] for evaluation.

Table 7 compares MGAN with the state-of-the-art methods under all three occlusion subsets:

4.5. Caltech Dataset

Here, MGAN is compared to the following recent state-of-the-art methods: CompACT-Deep [3], DeepParts[28], MS-CNN [2], RPN+BF [30], SA-F.RCNN [12], MCF [4], SDSRCNN [1], F.RCNN [31], F.RCNN+ATT-vbb [33], GDFL [13], and Bi-Box [36]. Tab. 7 compares MGAN with the state-of-the-art methods under all three occlusion subsets:
Figure 8. State-of-the-art comparison on the R, HO and R+HO subsets of Caltech dataset. The legend in each plot represents the log-averaged miss rate over FPPI=[10^{-2}, 10^6]. Our approach provides superior results compared to existing approaches on all three subsets.

Detector & Occl. & R & HO & R+HO \\
--- & --- & --- & --- & --- \\
DeepParts [28] & ✓ & 11.89 & 60.42 & 22.79 \\
ATT-part [33] & ✓ & 10.33 & 45.18 & 18.21 \\
SA-F.RCNN [12] & × & 9.68 & 64.35 & 21.92 \\
SDS-RCNN [1] & × & 7.36 & 58.55 & 19.72 \\
F.RCNN [31] & × & 9.18 & 57.58 & 20.03 \\
GDFL [13] & × & 7.85 & 43.18 & 15.64 \\
Bi-Box [36] & ✓ & 7.61 & 44.40 & 16.06 \\
Our MGAN & ✓ & 6.83 & 38.16 & 13.84 \\

Table 7: Comparison (in terms of log-average miss rate) of MGAN with the state-of-the-art methods on the Caltech dataset. The second column indicates whether the method is specifically targeted to handling occlusion. Best results are in bold. Under heavy occlusions (HO), our detector outperforms the state-of-the-art GDFL detector by 5.0%. Further, our detector provides superior results compared to all published methods on both the reasonable (R) and the combined set of reasonable and heavy occlusions (R+HO).

R, HO and R+HO. Among existing methods, the SDS-RCNN approach [1] reports a log-average miss rate of 7.36 on the R set. Our MGAN achieves superior results with a log-average miss rate of 6.83 on this set. On the HO and R+HO sets, the GDFL detector [13] provides the best results among the existing methods with a log-average miss rate of 43.18 and 15.64, respectively. Our MGAN detector outperforms GDFL with an absolute gain of 5.02% and 1.80% on HO and R+HO sets, respectively. Fig. 8 shows the comparison of our detector 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 [33], and GDFL [13] in Fig. 9. All results are obtained using the same FPPI. Our MGAN accurately detects pedestrians in all five scenarios.

5. Conclusion

We proposed a mask-guided attention network (MGAN) for occluded pedestrian detection. The 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 coarse-level segmentation information for visible region. In addition to MGA, we introduced an occlusion-sensitive loss term. Experiments on two datasets clearly show the effectiveness of our approach, especially for heavily occluded pedestrians.

Acknowledgments This work was supported by National Natural Science Foundation of China (Grant # 61632018).
References


