Mask-Guided Attention Network for Occluded Pedestrian Detection

Yanwei Pang[1], Jin Xie[1], Muhammad Haris Khan[2], Rao Muhammad Anwer[2],
Fahad Shahbaz Khan[2,3], Ling Shao[2]

[1]Tianjin University [2]Inception Institute of Artificial Intelligence, UAE [3]CVL, 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.

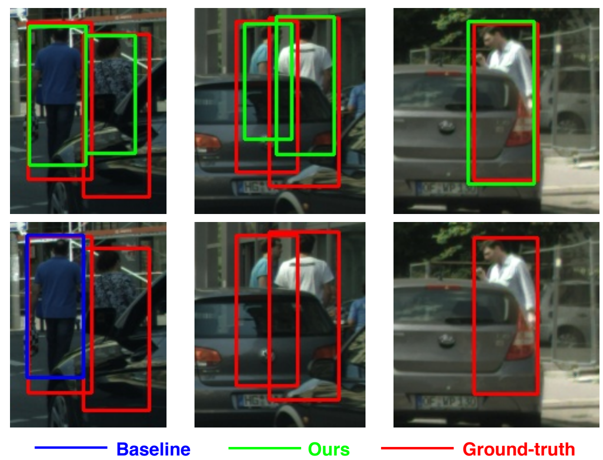


Figure 1. Detection examples using our approach (top row) and the baseline Faster R-CNN [24] (bottom row). For improved visualization, detection regions are cropped from images of CityPersons val. set [31]. All results are obtained using the same false positive per image (FPPI) criterion. Our approach robustly handles occlusions, yielding higher recall for occluded pedestrians.

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 pedestrian detector [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 rolling 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 infusion 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-

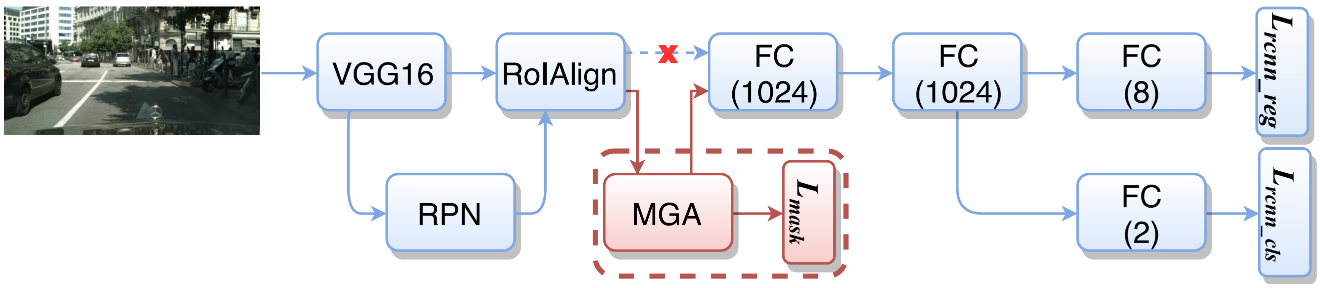


Figure 2. The overall network architecture of our Mask-Guided Attention Network (MGAN). It comprises a standard pedestrian detection (SPD) branch, whose components are shown in blue. It introduces a novel Mask-Guided Attention (MGA) module enclosed in red dashed box. Note, after RoI Align there is a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

body visible) in a supervised fashion [33]. Different to [36], we utilize the visible branch to generate a pixel-wise attention map that is used to modulate multi-channel convolutional features in the standard full body estimation branch. Unlike ATT-vbb [33], 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 approach provides a significant gain of 4.8% and 5.6% compared to [36] and [33], respectively on **HO** set of CityPersons.

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 a 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:

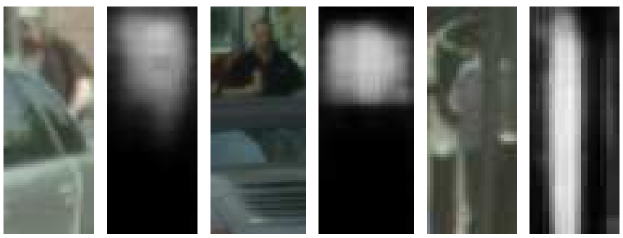
$$L_0 = L_{rpn\_cls} + L_{rpn\_reg} + L_{rcnn\_cls} + L_{rcnn\_reg}, \tag{2}$$

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 [1]. We present a **mask-guided spatial atten-**

[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 rate, it would result in high false positive detections.

Figure 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 detections. The detector cannot capture heavily occluded pedestrians and might result in high miss rate under similar circumstances.



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

Figure 4. 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.

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 $F_r \in$

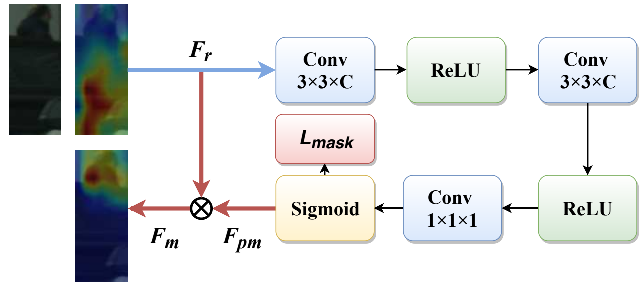


Figure 5. The network architecture of our Mask-Guided Attention (MGA) Branch. It takes RoI features and generates modulated features using a small stack of conv. operations, followed by ReLU nonlinearities.

$[H \times W \times C]$, where 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 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:

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

where i is the channel index and $\odot$ is the element-wise product. Instead of RoI features F_r, we feed modulated features F_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 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 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.

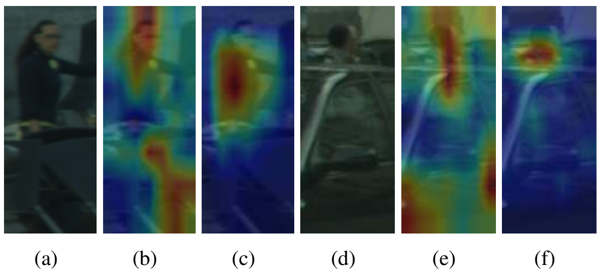


Figure 6. Visual comparison of RoI features and corresponding modulated features. (a) and (d) are two different proposals. (b) and (e) depict their corresponding RoI features. (c) and (f) show their corresponding modulated features. In contrast to RoI features, modulated features from our MGA branch have visible region signified and occluded part concealed.

3.2.3 Loss Function

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

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

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):

$$L_{mask} = BCELoss(p_n(x,y), \hat{p}_n(x,y)), \tag{5}$$

where $\hat{p}_n(x,y)$ 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):

$$\begin{aligned} L_{occ} = \frac{1}{N}\sum_{n=1}^{N}\{[1 - \frac{1}{WH}\sum_{x}^{W}\sum_{y}^{H} p_n(x,y)] \\ CELoss(p_n^{rcnn\_cls}, \hat{p}_n^{rcnn\_cls})\}, \end{aligned} \tag{6}$$

where W and H are the width and the height of pedestrian probability map. $\hat{p}_n^{rcnn\_cls}$ are the predictions produced by the classification branch of RCNN, and $p_n^{rcnn\_cls}$ 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 1×10^{-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 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-

Method	R	HO
Baseline SPD (L_0 loss in Eq.(1))	13.8	57.0
Our MGAN (L_0 + L_{mask})	11.9	52.7
Our MGAN (L_0 + L_{occ})	13.2	55.6
Our Final MGAN (L_0 + L_{mask} + L_{occ})	**11.5**	**51.7**

Table 1. Comparison (in log-average miss rates) of our MGAN with the baseline on the CityPersons. Best results are boldfaced. Beside our final MGAN (final row), we also show the performance of our MGA branch (second row) and occlusion-sensitive loss term (third row) alone. For fair comparison, we use the same training data, input scale ($\times 1$) and network backbone (VGG-16). On the heavy occlusion set (**HO**), our detector significantly reduces the error from 57.0 to 51.7, compared to the baseline.

Set	Dense Pixel-wise Annotations	Coarse-level Annotations
R	11.2	11.9
HO	51.7	52.7

Table 2. Comparison (in log-average miss rates) of our MGAN detector when using dense pixel-wise labeling with coarse-level 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 coarse-level segmentation provides a trade-off between annotation cost and accuracy.

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

Table 3. 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 CityPersons **HO** set. Best results are boldfaced in each case.

son. 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 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

Method	VBB	Backbone	Data (visibility)	Scale	R	HO
OR-CNN [32]	✓	VGG	$\geq 50\%$	$\times 1$	12.8	55.7
Our MGAN	✓	VGG	$\geq 50\%$	$\times 1$	**10.5**	**47.2**
OR-CNN [32]	✓	VGG	$\geq 50\%$	$\times 1.3$	11.0	51.3
Our MGAN	✓	VGG	$\geq 50\%$	$\times 1.3$	**9.9**	**45.4**
ATT-vbb [33]	✓	VGG	$\geq 65\%$	$\times 1$	16.4	57.3
Our MGAN	✓	VGG	$\geq 65\%$	$\times 1$	**11.5**	**51.7**
Bi-Box [36]	✓	VGG	$\geq 30\%$	$\times 1.3$	11.2	44.2
Our MGAN	✓	VGG	$\geq 30\%$	$\times 1.3$	**10.5**	**39.4**

Table 4. Comparison (in terms of log-average miss rate) with state-of-the-art methods that use both the visible bounding box (VBB) *and* full body information on CityPersons validation set. For 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 outperforms all three methods on both sets. Under heavy occlusions (**HO**), our MGAN significantly reduces the error from 44.2 to 39.4, compared to the recently introduced Bi-Box [36]. Best results are boldfaced in each case.

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]

Figure 7. Detection examples on CityPersons 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.

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 $1\times$, 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 $1.3\times$. Our MGAN detector outperforms OR-CNN with a significant margin on both input scales.

For an input scale of $1\times$, 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 $1.3\times$. 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 detections from our MGAN on CityPersons. Examples show a range of occlusion degrees *i.e.* from partial to heavy. Finally, Tab. 6 shows the state-of-the-art comparison on the CityPersons test set.

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

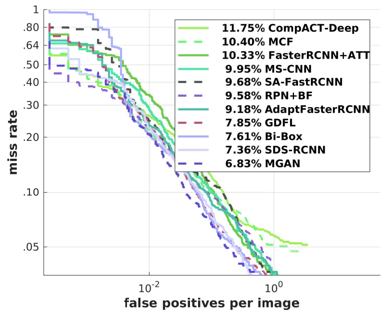
Table 5. Comparison (in terms of log-average miss rate) of MGAN with state-of-the-art methods in literature on CityPersons validation set. Our MGAN sets a new state-of-the-art by outperforming all existing methods. Best results are boldfaced in each case.

Method	**R**	**HO**
Adaptive Faster RCNN [31]	12.97	50.47
Rep. Loss [29]	11.48	52.59
OR-CNN [32]	11.32	51.43
Our MGAN	**9.29**	**40.97**

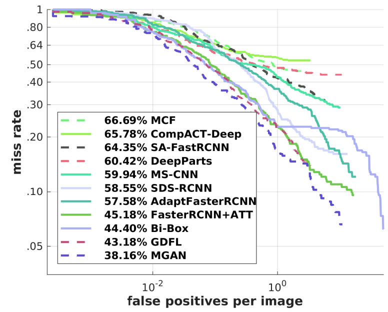
Table 6. Comparison (in terms of log-average miss rate) of MGAN with state-of-the-art methods on CityPersons test set. The test set is withheld and results are obtained by sending our detection predictions to the authors of CityPersons dataset [31] for evaluation.

4.5. Caltech Dataset

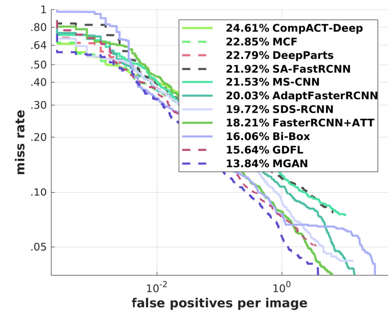
Here, MGAN is compared to the following recent state-of-art methods: CompACT-Deep [3], DeepParts[28], MS-CNN [2], RPN+BF [30], SA-F.RCNN [12], MCF [4], SDS-RCNN [1], F.RCNN [31], F.RCNN+ATT-vbb [33], GDFL [13], and Bi-Box [36]. Tab. 7 compares MGAN with the state-of-the-art methods under all three occlusion subsets:



(a) **R**



(b) **HO**



(c) **R+HO**

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=$[10^{-2}, 10^{0}]$. Our approach provides superior results compared to existing approaches on all three subsets.

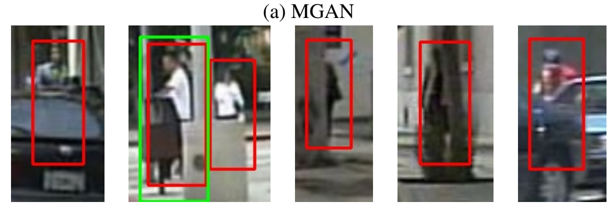
Detector	Occl.	R	HO	R+HO
CompACT-Deep [3]	×	11.75	65.78	24.61
DeepParts[28]	✓	11.89	60.42	22.79
MCF [4]	×	10.40	66.69	22.85
ATT-part [33]	✓	10.33	45.18	18.21
MS-CNN [2]	×	9.95	59.94	21.53
RPN+BF [30]	×	9.58	74.36	24.01
SA-F.RCNN [12]	×	9.68	64.35	21.92
SDS-RCNN [1]	×	7.36	58.55	19.72
F.RCNN [31]	×	9.18	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-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**).

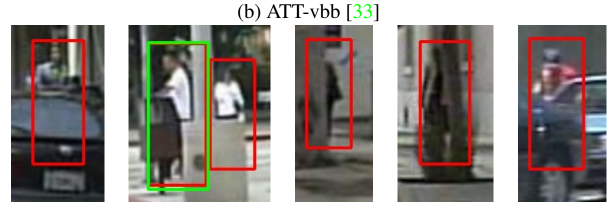
R, **HO** and **R+HO**. Among existing methods, the SDS-RCNN approach [1] reports a log-average miss rate of 7.36 on the **R** set. Our MGAN achieves superior results with a log-average miss rate of 6.83 on this set. On the **HO** and **R+HO** sets, the GDFL detector [13] provides the best results among 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.

(a) MGAN



(b) ATT-vbb [33]



(c) GDFL [13]

Figure 9. Qualitative comparison of (a) MGAN with (b) ATT-vbb [33] and (c) GDFL [13] on Caltech test set. Red boxes denote the ground-truth and green boxes indicate detector predictions. Examples images depict varying level of occlusions.

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

[1] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection & segmentation. In *ICCV*, 2017. 1, 2, 7, 8

[2] Zhaowei Cai, Quanfu Fan, Rogerio Schmidt Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016. 1, 2, 7, 8

[3] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015. 7, 8

[4] Jiale Cao, Yanwei Pang, and Xuelong Li. Learning multilayer channel features for pedestrian detection. *TIP*, 26(7):3210–3220, July 2017. 7, 8

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang1, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *ECCV*, 2018. 2

[6] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 6

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34(4):743–761, April 2012. 2, 4, 5

[8] Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. In *WACV*, 2017. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[12] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *TMM*, 20(4):985–996, April 2018. 7, 8

[13] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. Graininess-aware deep feature learning for pedestrian detection. In *ECCV*, 2018. 1, 7, 8

[14] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *TPAMI*, 2018. 1

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 1

[16] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *ECCV*, 2018. 1, 2, 6, 7

[17] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *CVPR*, 2017. 1, 2

[18] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool. Handling occlusions with franken-classifiers. In *ICCV*, 2013. 2

[19] Junhyug Noh, Soochan Lee, Beomsu Kim, and Gunhee Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *CVPR*, 2018. 2

[20] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013. 2

[21] Yanwei Pang, Jiale Cao, and Xuelong Li. Cascade learning by optimally partitioning. *TCyb*, 47(12):4148–4161, Dec 2017. 2

[22] Yanwei Pang, Manli Sun, Xiaoheng Jiang, and Xuelong Li. Convolution in convolution for network in network. *TNNLS*, 29(5):1587–1597, May 2018. 1

[23] Jimmy Ren, Xiaohao Chen, Jianbo Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai, and Li Xu. Accurate single stage detector using recurrent rolling convolution. In *CVPR*, 2017. 1, 2

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 3

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 3, 5

[26] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *ECCV*, 2018. 6, 7

[27] Hanqing Sun and Yanwei Pang. Glancenets efficient convolutional neural networks with adaptive hard example mining. *Science China Information Sciences*, 61(10):101–109, Oct 2018. 1

[28] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, 2015. 2, 7, 8

[29] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*, 2018. 1, 2, 6, 7

[30] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, 2016. 2, 7, 8

[31] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017. 1, 2, 4, 5, 7, 8

[32] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *ECCV*, 2018. 6, 7

[33] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *CVPR*, 2018. 2, 3, 5, 6, 7, 8

[34] Chunluan Zhou and Junsong Yuan. Non-rectangular part discovery for object detection. In *BMVC*, 2014. 2

[35] Chunluan Zhou and Junsong Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *ICCV*, 2017. 2

[36] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *ECCV*, 2018. 2, 3, 6, 7, 8