Count- and Similarity-aware R-CNN for Pedestrian Detection

Jin Xie[1⋆], Hisham Cholakkal[2,3⋆], Rao Muhammad Anwer[2,3], Fahad Shahbaz Khan[2,3], Yanwei Pang[1⋆⋆], Ling Shao[2,3], and Mubarak Shah[4]

[1] Tianjin Key Laboratory of Brain-Inspired Artificial Intelligence, School of Electrical and Information Engineering, Tianjin University, China
[2] Mohamed bin Zayed University of Artificial Intelligence, UAE
[3] Inception Institute of Artificial Intelligence, UAE
[4] University of Central Florida, USA
{jinxie, pyw}@tju.edu.cn,
{hisham.cholakkal,rao.anwer,fahad.khan,ling.shao}@mbzuai.ac.ae,
shah@crcv.ucf.edu

Abstract. Recent pedestrian detection methods generally rely on additional supervision, such as visible bounding-box annotations, to handle heavy occlusions. We propose an approach that leverages pedestrian count and proposal similarity information within a two-stage pedestrian detection framework. Both pedestrian count and proposal similarity are derived from standard full-body annotations commonly used to train pedestrian detectors. We introduce a count-weighted detection loss function that assigns higher weights to the detection errors occurring at highly overlapping pedestrians. The proposed loss function is utilized at both stages of the two-stage detector. We further introduce a count-and-similarity branch within the two-stage detection framework, which predicts pedestrian count as well as proposal similarity. Lastly, we introduce a count and similarity-aware NMS strategy to identify distinct proposals. Our approach requires neither part information nor visible bounding-box annotations. Experiments are performed on the CityPersons and CrowdHuman datasets. Our method sets a new state-of-the-art on both datasets. Further, it achieves an absolute gain of 2.4% over the current state-of-the-art, in terms of log-average miss rate, on the heavily occluded (**HO**) set of CityPersons test set. Finally, we demonstrate the applicability of our approach for the problem of human instance segmentation. Code and models are available at: https://github.com/Leotju/CaSe.

Keywords: Pedestrian detection, Human instance segmentation.

1 Introduction

Pedestrian detection is a challenging computer vision problem and serves as an important component in many vision systems. Despite recent progress, detecting

⋆ The first two authors contribute equally to this work.
⋆⋆ Corresponding author.

heavily occluded pedestrians remains a key challenge in real-world applications due to the frequent occurrence of occlusions. The most common type of occlusion in pedestrian detection is *crowd occlusion* caused by other pedestrians. This is evident in recent benchmarks, such as CityPersons [30], where crowd occlusion alone accounts for around 49%. In this work, we tackle the problem of heavily occluded pedestrian detection.

Most existing pedestrian detection approaches either rely on part information [25,36] or exploit visible bounding-box annotations [31,37,20] to handle occlusions. Typically, part-based approaches are computationally expensive and require a large number of part detectors. Recent approaches relying on visible bounding-box supervision, in addition to standard full-body annotations, have shown superior performance for occluded pedestrian detection. However, this reliance on visible bounding-box annotations introduces another level of supervision. In this work, we propose an approach for occluded pedestrian detection that requires neither part information nor visible bounding-box supervision.

State-of-the-art pedestrian detectors [3,16,2,27,20,5,6] are mostly based on two-stage detection framework. One of the most commonly used two-stage object detection frameworks is that of Faster R-CNN [21], later adapted for pedestrian detection [30]. Here [21,30], a region proposal network (RPN) is employed in the first stage to generate pedestrian proposals. The second stage, also known as Fast R-CNN, consists of an RoI (region-of-interest) feature extraction from each proposal followed by classification confidence prediction and bounding-box regression. During inference, a post-processing strategy, such as non-maximum suppression (NMS), is used to remove duplicate bounding-box predictions.

While promising results have been achieved when adapting Faster R-CNN for standard pedestrian detection [30], its performance on heavily occluded pedestrians is far from satisfactory. This is likely due to the fact that the number of overlapping pedestrian instances in an RoI pooled region are not explicitly taken into account, during either training or inference. In this work, we argue that pedestrian count and proposal similarities are useful cues for tackling crowd occlusion with no additional annotation cost. Pedestrian count information within an RoI is readily available with full-body annotations that are typically used in pedestrian detection training. During training, a higher pedestrian count within an RoI indicates a high level of crowd occlusion. In such crowd occlusion scenarios, multiple highly overlapping pedestrians need to be detected from a large number of spatially adjacent duplicate proposals. A proposal similarity embedding is desired to identify distinct proposals from multiple duplicate proposals for each pedestrian. Count and similarity predictions at inference are expected to aid accurate detection of highly overlapping pedestrians (crowd occlusion).

Contributions: To the best of knowledge, we are the first to leverage pedestrian count *and* proposal similarity information in a two-stage framework for occluded pedestrian detection. Our contributions are: **(i)** a count-weighted detection loss is introduced for the classification and regression parts of both the RPN and Fast R-CNN modules, during training. As a result, a higher weight is assigned to proposals with a large number of overlapping pedestrians, improving the perfor-

mance during heavy-occlusion. **(ii)** We introduce a count-and-similarity branch to accurately predict both pedestrian count and proposal similarity, leading to a novel multi-task setting in Faster R-CNN. **(iii)** Both the predicted count and proposal similarity embedding are utilized in our count and similarity-aware NMS strategy (CAS-NMS), to identify distinct proposals in a crowded scene. **(iv)** Extensive experiments are performed on CityPersons [30] and CrowdHuman [22]. Our count- and similarity-aware, R-CNN based pedestrian detection approach, dubbed as *CaSe*, achieves state-of-the-art results on both datasets. On heavily occluded (**HO**) set of the CityPersons test set, our detector improves the state-of-the-art results [20], reducing the log-average miss rate from 41.0% to 38.6%. Note that [20] requires both full-body and visible bounding-box annotations. In contrast, our approach only utilizes full-body annotations. **(v)** Additionally, we validate our proposed components by integrating them into Mask R-CNN framework for person instance segmentation, achieving consistent improvement in performance on OCHuman [33].

2 Related Work

Several pedestrian detectors apply a part-based approach [17,35,19,25,36], where a set of part detectors is learned with each one designed for handling a specific occlusion pattern. Different from these approaches, more recent works aim at exploiting additional visible bounding-box (VBB) supervision to either output visible part regions [37] or provide support for learning occlusion scenarios [20]. Here, we look into an alternative approach that neither uses part information nor requires additional VBB annotation for occluded pedestrian detection.

Generally, object detectors [21,18] employ non-maximum suppression (NMS) as a post processing strategy. Several previous works have investigated improving NMS for the generic object detection [1,26,12,11]. Despite being extensively investigated for generic object detection, less attention has been paid to improve NMS in the context of occluded pedestrian detection [10,13]. The work of [13] proposes an approach that learns to predict the threshold according to the instance-specific density. The work of [10] introduces an approach based on the joint processing of detections and penalization for double detections. Improving NMS for occluded pedestrian detection is an open problem, as most existing pedestrian detectors [14,27,20] still employ traditional post-processing strategy.

Recent works have investigated problem of improving bounding-box regression for crowd occlusion [27,32]. The work of [27] introduces repulsion losses that penalize predicted boxes from shifting towards other ground-truth objects, requiring each predicted box to be away from those with different ground-truths. The work of [32] proposes an approach that learns to adapt the predicted boxes closer to the corresponding ground-truths. Both of these approaches are employed on the regression part of the pedestrian detector. Instead, our proposed count-weighted detection loss is designed for both classification and regression parts of the two-stage Faster R-CNN detector. In addition to the count-weighted detection loss, we introduce a parallel count-and-similarity branch within the

Fast R-CNN module of Faster R-CNN to accurately predict both pedestrian count and proposal similarity. Further, we use both predicted count and proposal similarity embedding for distinct proposal identification during inference.

3 Baseline Two-Stage Detection Framework

In this work, we base our approach on the popular Faster R-CNN framework [21] that is adopted in several two-stage pedestrian detectors [30,20,28] as their base architecture. Faster R-CNN employs a region proposal network (RPN), during the first stage, to generate class-agnostic proposals and their confidence scores, respectively. In the second stage, also known as Fast R-CNN, RoI (region-of-interest) features are extracted from each proposal, followed by a detection branch that generates classification score (*e.g.*, probability of a proposal being a pedestrian) and regressed bounding-box coordinates for each proposal.

The detection problem can be formulated as a joint minimization of the classification and regression losses, in both the RPN and Fast R-CNN modules [21], where $L_{det} = L_{rpn} + L_{frc}$. Here, both L_{rpn} and L_{frc} are computed as an accumulation of the average classification and regression loss L_c and L_r, in their respective modules. L_c and L_r are given by:

$$L_c = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) \quad L_r = \lambda \frac{1}{N_{reg}} \sum_i L_{reg}(l_i, l_i^*), \tag{1}$$

where i represents index of a proposal in a mini-batch, p_i represents the predicted probability of proposal i being a pedestrian, and p_i^* is the ground-truth label of the proposal. The predicted location of a positive proposal i is denoted by l_i and l_i^* denotes the associated ground-truth location. N_{cls} and N_{reg} are the total number of proposals during classification and regression, respectively. λ is the parameter to balance the two loss terms. The classification loss (L_{cls}), is a cross-entropy loss for RPN and Fast R-CNN modules. The regression loss (L_{reg}), for both RPN and Fast R-CNN modules, is Smooth-L1 loss function.

4 Our Approach

Motivation: The above-mentioned two-stage Faster R-CNN baseline is trained using full-body pedestrian annotations. In recent methods [20,31,32,37,34], this two-stage standard pedestrian detection framework, or Faster R-CNN, has been extended to incorporate additional visible bounding-box annotations. Here, we propose an approach that does not rely on additional visible bounding-box supervision and instead utilizes pedestrian count information within an RoI, which is readily available with standard full-body annotations.

Overall Architecture: Fig. 1 shows the overall network architecture. It consists of a detection branch and a count-and-similarity branch. Both these branches take the RoI feature of a proposal as input. Our approach leverages count information within the baseline pedestrian detection framework at two different

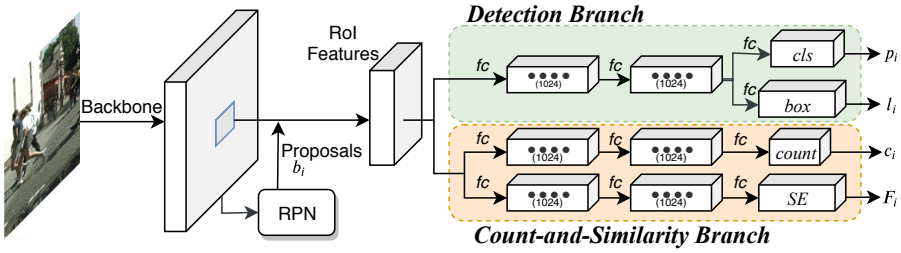

Fig. 1. Overall architecture of the proposed count- and similarity-aware pedestrian detector (CaSe). Our CaSe consists of a Detection Branch (sec. 4.1) and a Count-and-Similarity branch (sec. 4.2). We introduce a count-weighted detection loss that employ a count-weighting in regression and classification losses of both the RPN and Fast R-CNN stages of Faster R-CNN. The detection branch predicts the pedestrian location (l_i) and its probability (p_i). The count-and-similarity branch introduced in our CaSe consists of two outputs: the count prediction (c_i) and similarity embedding (F_i). The count-prediction estimates the number of ground-truth instances in a given RoI, whereas the similarity embedding measures the similarity between all overlapping proposals. Both these outputs are further used for distinct proposal identification during inference.

stages. First, our count-weighted detection loss integrate a count weighting into the classification and regression losses of both modules (RPN and Fast R-CNN). Second, a parallel network, which we call the count-and-similarity branch, is introduced to improve bounding-box prediction by estimating the number of pedestrian instances for a given pedestrian proposal and its similarity with overlapping proposals. Further, we use both predicted count and proposal similarity embedding to identify distinct proposals during inference, by introducing a novel post-processing strategy. Both the detection and count-and-similarity branches in our network are jointly trained with the loss function $L = L^{'}_{det} + L_{cas}$. Here, L_{cas} is the training loss for the count-and-similarity branch and $L^{'}_{det}$ is the proposed count-weighted detection loss, employed in *both* the RPN and Fast R-CNN modules. Next, we describe the detection branch and the associated count-weighted detection loss. Then, the proposed count-and-similarity branch is presented in Sec. 4.2. Finally, inference of the proposed framework and our novel post-processing strategy are described in Sec. 4.3.

4.1 Detection Branch

As described earlier(Sec. 3), our framework is based on two-stage Faster R-CNN, employed in several pedestrian detection methods [30,20]. Next, we introduce a novel count-weighted detection loss to improve both the localization and classification performance of the Faster R-CNN under heavy occlusion. Our count-weighted detection loss is integrated in both RPN and Fast R-CNN modules.

Count-weighted Detection Loss: Different to detecting isolated pedestrians in a sparse scene, pedestrian detection in crowded scenes is a more challenging problem due to the presence of multiple highly overlapping pedestrians. To

counter this issue, we introduce a weight w_i proportional to the ground-truth count of a proposal in the classification and regression loss terms in Eq. 1. This implies that a higher weight is assigned to detection errors occurring at highly overlapping pedestrians (crowd occlusions). Our count-weighted detection loss function $L_{det}^{'}$ (CW-loss) has the following terms:

$$L_c^{'} = \frac{1}{N_{cls}} \sum_i w_i L_{cls}(p_i, p_i^*) \quad L_r^{'} = \lambda \frac{1}{N_{reg}} \sum_i w_i L_{reg}(l_i, l_i^*), \tag{2}$$

where w_i is the loss weight that assigns a higher weightage to a proposal overlapping with a large number of ground-truth boxes. The weight w_i of each proposal box b_i can be obtained from its ground-truth count c_i^* as

$$w_i = 1 + \alpha \cdot \mathbf{max}(c_i^* - 1, 0), \tag{3}$$

where α is a balancing factor, empirically set to 0.5. It can be observed that, if a positive proposal b_i overlaps with multiple ground-truth bounding-boxes, a higher weight w_i will be assigned to that sample. This implies that the proposals at crowded image regions will be assigned with a higher weightage *during training*, compared to the proposals from less crowded regions. Next, we explain how to compute the ground-truth count using the full-body bounding-box annotations which are readily available during training.

Ground-truth Count of a Proposal: The ground-truth count c_i^* of a proposal b_i depends on the number of overlapping full-body (ground-truth) bounding-boxes. First, we compute the intersection-over-union (IoU) between b_i and all its overlapping ground-truth bounding boxes. Then, c_i^* is computed as the number of ground-truth bounding boxes with an $IoU \geq th$. Here, th is empirically set to 0.5. During training, the ground-truth count c_i^* of a proposal is used to compute the loss weight (Eq. 3). Further, it is used as a ground-truth count to train our count-and-similarity branch.

4.2 Count-and-Similarity Branch

Combined Use of Count and Similarity: In the presence of crowd occlusion, many highly overlapping duplicate proposals are generated and assigned a higher classification score by the detector. This is problematic when using a fixed overlap threshold to remove duplicate proposals. Fig. 2 shows an example with two highly overlapping pedestrians (crowd occlusion). In such a case, a count prediction for an RoI can be used to obtain the number of overlapping pedestrians. This count prediction can be utilized to adapt the overlap threshold, thereby removing duplicate proposals based on their classification scores. However, count alone is sub-optimal for identifying distinct proposals in crowd occlusion scenarios since several proposals with higher classification scores may belong to the same pedestrian instance. Therefore, it is desired to identify distinct proposals belonging to different (overlapping) pedestrians. To this end, we utilize a similarity embedding that projects RoI features into a low-dimensional representation,

where the euclidean distance is inversely proportional to similarity between proposals (see Fig. 2). Instead of calculating the euclidean distance for all pairs of overlapping proposals (above a certain threshold), we use count prediction as an indicator to compute the distance for only a subset of pairs having a predicted count more than one. This leads to speed-up during inference, compared to exhaustively computing the distance for all pairs of overlapping proposals.

Our count-and-similarity branch is shown in Fig. 1. It has two parallel sub-branches, where, the first predicts the number of pedestrians present within a proposal (RoI), and the second outputs a similarity embedding for estimating the proposal similarity. During training, we define the loss L_{cas} on each RoI as $L_{cas} = L_{cp} + L_{se}$. Here L_{cp} denotes the counting loss for the first sub-branch and L_{se} denotes the similarity embedding loss for the second sub-branch.

Proposal Count: The pedestrian count of a given proposal b_i is predicted by a sub-branch that consists of three fully-connected (fc) layers, where the last layer outputs the pedestrian count c_i. The three fc layers are separated by a ReLU layer. The count loss L_{cp} is a mean-squared error (MSE) loss that penalize the deviation of the predicted count from its ground truth. *i.e.*, $L_{cp} = \frac{1}{N_{cp}} \sum_{i=1}^{N_{cp}} \|c_i - c_i^*\|_2^2$, where N_{cp} denotes number of proposals used when training with the count loss, and c_i^* represents the ground-truth count of a proposal, described in Sec. 4.1.

Proposal Similarity: As mentioned earlier, the predicted count alone is sub-optimal to identify distinct proposals in a crowded scene. To address this issue, we introduce a similarity embedding sub-branch that projects RoI feature of a proposal b_i to a low-dimensional feature embedding F_i. The similarity embedding sub-branch has a structure similar to its parallel (count) sub-branch, except its final fc layer outputs a 64-dimensional feature embedding F_i. The euclidean distance between the feature embedding of two proposals is proportional to their dissimilarity. Proposals with no pedestrians can be removed based on their predicted count. Hence, the euclidean distance between two overlapping proposals only needs to be computed if both proposals contain at least one pedestrian.

For a given proposal b_i, we first select its overlapping proposals with an $IoU \geq 0.5$. Let b_j be one of the selected proposals that has a ground-truth count $c_j^* \geq 1$, and a feature embedding F_j. We train the similarity embedding sub-branch with proposals having a ground-truth count of at least one, using the contrastive loss:

$$L_{se} = \frac{\sum_{ij}(y_{ij} d_{ij}^2 + (1 - y_{ij})\ \mathbf{max}(\delta - d_{ij}, 0)^2)}{N_{se}} \tag{4}$$

where $d_{ij} = \|F_i - F_j\|_2$ indicates the distance between the feature embeddings F_i and F_j. The binary label y_{ij} indicates the ground-truth similarity, where proposals of the same ground-truth bounding box are labelled as similar, *i.e.*, $y_{ij} = 1$. N_{se} is the number of proposals used when training with the similarity embedding loss. The margin δ is set to 2. Training of our similarity embedding sub-branch using contrastive loss (L_{se} helps in projecting RoI features to a low-dimensional feature embedding, where the distance between proposals of

Fig. 2. An illustrative example showing operations during inference of our CaSe detector. On left: Predictions from the detection and count-and-similarity branches. On right: final output of our CaSe framework including a count and similarity-aware post-processing step. The red box (b_H) indicates the proposal with highest classification confidence score, and all its overlapping proposals (b_j) are shown with dotted boxes. Although the yellow box b_2 has a large count prediction ($c_2 = 1.7$), it is highly similar to b_H, as indicated by the distance $d_{2H} = 0.01$, and is thus removed by our count and similarity-aware post-processing step. Similarly, cyan box is removed due to low count prediction $c_3 = 0.4$. The green box belonging to another pedestrian is predicted with a high count ($c_1 = 1.6$) and a higher distance $d_{1H} = 1.6$, hence not removed.

two distinct overlapping instances is large.) Next, we describe the procedure to identify distinct proposals, during inference.

4.3 Inference

During inference, our approach predicts both the count and similarity embedding in addition to pedestrian classification and regression. This is followed by a count and similarity-aware post-processing step for removing duplicate proposals. In crowded scenes, there are multiple ground-truth boxes with a very high overlap. Hence, the detected proposals are also expected to be highly overlapped. Traditional post-processing involves an NMS strategy where a fixed overlap threshold is employed to eliminate overlapping bounding-box predictions. This often results in a loss of correct target bounding-boxes. To counter this issue, we introduce a post-processing step, named count and similarity-aware NMS (CAS-NMS), that considers both the count and similarity between proposals.

Fig. 2 shows the procedure involved in our CAS-NMS. We first sort the proposals based on their classification confidence scores. Let b_H be the proposal with the highest classification score. Similar to traditional NMS, all the proposals overlapping with b_H are selected as possible duplicate proposals (*i.e.*, $IoU \geq 0.5$). Let b_j be one such selected proposal that has $IoU \geq 0.5$ with b_H. In scenarios where b_j corresponds to a distinct pedestrian (*i.e.*, different to the one localized by b_H), it is common that (i) there are more than one pedestrians in b_H, (ii) there is at least one pedestrian in b_j, and (iii) both b_H and b_j are dissimilar. Our CAS-NMS uses predicted count and similarity embedding of

both b_H and b_j, and categorizes b_j as a duplicate proposal, when any of the three criteria mentioned above are not fulfilled. *i.e.*, If predicted counts c_H, c_j of the proposals are below thresholds t_2, t_1, or the distance $d_{jH} = \|F_j - F_H\|_2$ between the similarity embeddings of both proposals is below a threshold N_{st}. Since, the distance between proposals is required only in the third criterion, d_{jH} is computed only for proposals satisfying the first two criteria. Our CAS-NMS removes duplicate proposals and preserves only distinct proposals containing at least one pedestrian. More details are available at https://github.com/Leotju/CaSe.

5 Experiments

5.1 Datasets and Evaluation Metrics

Datasets: We perform experiments on two challenging datasets: CityPersons [30] and CrowdHuman [22]. CityPersons contains 2975 training, 500 validation, and 1525 test images. It is suitable to evaluate performance on occluded pedestrians, as around 70% of the pedestrians in the dataset depict various levels of occlusions [30]. CrowdHuman contains crowded scenes and is therefore also suitable to evaluate performance on crowd occlusions. It consists of 15000, 4370, and 5000 images in the training, validation and test sets, respectively. Further, it contains more than 470K human instances in the training and validation sets, with many images containing more than 20 person instances.
Evaluation Metrics: For both datasets, we report pedestrian detection performance using average-log miss rate (MR), computed over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ [8]. We select MR^{-2} to report the results and its lower value to mirror better detection performance. On CityPersons dataset, following [30,20], we report results on two different degrees of occlusions: Reasonable (**R**), and Heavy Occlusion (**HO**) to evaluate our approach. In the **R** set, the visibility ratio is larger than 65%, whereas the visibility ratio in the **HO** set ranges from 20% to 65%. And the height of pedestrians over 50 pixels is taken for detection evaluation, as in [31,20]. On CrowdHuman dataset, we follow the same evaluation protocol as in [22].

5.2 Implementation Details

Our framework utilizes an ImageNet pre-trained backbone (*e.g.* VGG-16 [23]). On CityPersons datasets, we follow the same experimental protocol as in [20]. On CrowdHuman datasets, we follow the same experimental protocol as in [22]. In case of CityPersons, the ($\times$1) input scale is 1024×2048 and $\times$1.3 input scale is 1344×2688. For CrowdHuman, the scale of input images is resized such that the shorter side is at 800 pixels while the longer side does not exceed more than 1400 pixels. In our experiments, the hyper-parameters $t_2 = 1.5$, $t_1 = 1$ and $N_{st} = 1.5$ are fixed for all datasets. Further, for the similarity sub-branch, the number of dissimilar and similar pairs are set to 16 and 32, respectively, for all the experiments. Our network is trained on NVIDIA GPUs and a mini-batch comprises 2 images per GPU.

Table 1. State-of-the-art comparisons (in terms of log-average miss rate) on the CityPersons validation set. Best results are boldfaced in each case. For fair comparison, we select the same set of ground-truth pedestrian examples and input scale when comparing our CaSe with each existing method. We also compare with existing methods using additional visible bounding-box (VBB) supervision. Our CaSe detector sets a new-state-of-the-art on both sets. Under heavy occlusions (**HO** set), our CaSe outperforms the state-of-the-art MGAN [20], reducing the error from 39.4% to 37.4%, without using additional VBB supervision.

Method	VBB	Training Setting		**R**	**HO**
		Visibility	Input Scale		
TLL [24]	×	-	×1	14.4	52.0
F.RCNN+ATT-vbb [31]	✓	≥ 65%	×1	16.4	57.3
F.RCNN+ATT-part [31]	×		×1	16.0	56.7
Repulsion Loss [27]	×		×1	13.2	56.9
Adaptive-NMS [13]	×		×1	11.9	55.2
MGAN [20]	✓		×1	11.5	51.7
CaSe (Ours)	×		×1	**11.0**	**50.3**
OR-CNN [32]	✓	≥ 50%	×1	12.8	55.7
MGAN [20]	✓		×1	10.8	46.7
CaSe (Ours)	×		×1	**10.1**	**45.2**
ALFNet [14]	×	≥ 0%	×1	12.0	51.9
CSP [15]	×		×1	11.0	49.3
MGAN [20]	✓		×1	11.3	42.0
CaSe (Ours)	×		×1	**10.5**	**40.5**
Repulsion Loss [27]	×	≥ 65%	×1.3	11.6	55.3
Adaptive-NMS [13]	×		×1.3	10.8	54.0
MGAN [20]	✓		×1.3	10.3	49.6
CaSe (Ours)	×		×1.3	**9.6**	**48.2**
OR-CNN [32]	✓	≥ 50%	×1.3	11.0	51.3
MGAN [20]	✓		×1.3	9.9	45.4
CaSe (Ours)	×		×1.3	**9.1**	**43.6**
Bi-box [37]	✓	≥ 30%	×1.3	11.2	44.2
FRCN +A +DT [34]	✓		×1.3	11.1	44.3
MGAN [20]	✓		×1.3	10.5	39.4
CaSe (Ours)	×		×1.3	**9.8**	**37.4**

5.3 CityPersons Dataset

State-of-the-art Comparison: We compare our CaSe detector with the recent state-of-the-art methods, namely Repulsion Loss [27], F.RCNN+ATT-vbb [31], F.RCNN+ATT-part [31], OR-CNN [32], TLL [24], Bi-Box [37], Adaptive-NMS [13], FRCN +A +DT [34], and MGAN [20] on the CityPersons validation set. Tab. 1 shows the state-of-the-art comparison on the validation set. Note that different set of ground-truth pedestrian examples are used for training by existing state-of-the-art pedestrian detection methods. For fair comparison, we therefore select the same set of ground-truth pedestrian examples that are at least 50 pixels tall with different visibility (mentioned in 'Training Setting' column of Tab 1) and an input scale, when comparing with each existing method.

Table 2. State-of-the-art comparison (in terms of log-average miss rate) on the CityPersons test set. Note that the test set is withheld. The results are obtained by sending our detection predictions to the authors of CityPersons [30] for evaluation. Our approach achieves state-of-the-art results on both the **R** and **HO** sets. On the **HO** set, our approach significantly outperforms the recently introduced MGAN [20], reducing the error from 41.0% to 38.6%.

Method	**R**	**HO**
Adaptive Faster RCNN [30]	13.0	50.5
MS-CNN [4]	13.3	51.9
Rep. Loss [27]	11.5	52.6
OR-CNN [32]	11.3	51.4
Cascade MS-CNN [4]	11.6	47.1
Adaptive-NMS [13]	11.4	-
MGAN [20]	9.3	41.0
CaSe (Ours)	**9.2**	**38.6**

Among recently introduced pedestrian detectors, ATT-vbb [31], OR-CNN [32], Bi-Box [37], FRCN+A+DT [34], and MGAN [20] utilize full-body and additional visible bounding-box (VBB) supervision. Our CaSe outperforms all these approaches on both the **R** and **HO** sets, *without* using VBB supervision. When using an input scale of $\times 1$, Repulsion Loss [27] achieves a log-average miss rate of 13.2% and 56.9% on the **R** and **HO** sets, respectively. Our CaSe provides superior detection performance compared to [27] with a log-average miss rate of 11.0% and 50.3% on the **R** and **HO** sets, respectively. Similarly, a consistent improvement in performance is obtained over [27], when using an input scale of $\times 1.3$ and the same training settings.

On the validation set, the best reported result for the **HO** subset is 39.4%, in terms of a log-average miss rate, obtained by the recently introduced MGAN [20] with an input scale of $\times 1.3$. Our CaSe sets a new state-of-the-art on the **HO** set with a log-average miss rate of 37.4%. Our detector also outperforms existing methods on the **R** set. Additionally, we present the results on the test set in Tab 2. Note that the test set is withheld and the results are obtained by sending our detector predictions to the authors of CityPersons [30]. Our detector outperforms all reported methods on both sets of the test set. Fig. 3 shows the qualitative detection comparisons between Repulsion Loss approach [27] and our CaSe. Note that similar to our approach, Repulsion Loss method [27] also specifically targets at handling occlusions.

Comparison with PedHunter and APD: Other than methods exploiting additional VBB information, the work of [7], termed as PedHunter, utilizes extra head annotations. PedHunter [7] integrates three novel training strategies to the pedestrian detector training stage, achieving promising results. We integrate the PedHunter training strategies by re-implementing them in our framework and observe this to further improve the performance[‡]. Through the integration of

[‡] Thanks to the PedHunter [7] authors for sharing head annotation on CityPersons validation set through email correspondence.

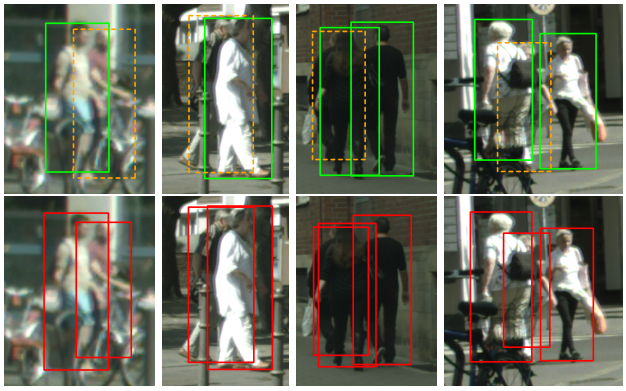

Fig. 3. Qualitative detection results on CityPersons validation set using Repulsion loss [27] (top) and our proposed CaSe (bottom). Detection results from the Repulsion loss and our CaSe are shown in green and red bounding-boxes, respectively. False negative detection results are shown with dashed orange bounding-boxes.

Table 3. Comparison (in log-average miss rate) of our CaSe with the baseline on the CityPersons validation set. We report results using two training settings: visibility $\geq$ 65% with $\times$1.3 and $\times$1.0 input scales. In both settings, we observe a consistent improvement in performance due to progressively integrating one contribution at a time. Note that Baseline + CW-loss + CSB just shows the impact of joint training the detection branch (using CW-loss) and the count-and-similarity branch (CSB). The predictions from CSB are further utilized in CAS-NMS, resulting in a significant improvement in our overall results (CaSe: Baseline + CW-loss + CSB + CAS-NMS).

Input Scale	Baseline	CW-Loss	CSB	CAS-NMS	**R**	**HO**
$\times$1.3	✓				12.2	53.5
	✓	✓			11.3	51.5
	✓	✓	✓		10.8	49.3
	✓	✓	✓	✓	**9.6**	**48.2**
$\times$1.0	✓				13.8	57.0
	✓	✓	✓	✓	**11.0**	**50.3**

PedHunter modules, our CaSe can achieve a log-average miss rate of 8.0% and 41.2% on **R** and **HO** subsets of CityPersons validation set. APD [29] uses a stronger backbone (DLA-34), instead of VGG-16. For a fair comparison with APD, we re-train our model using DLA-34 backbone. Our method outperforms APD on both **R** and **HO** sets of CityPersons validation set and achieves log-average miss rates of 8.3% and 43.2%, respectively.

Ablation Study: Here, we analyze our CaSe approach on the CityPersons benchmark by demonstrating impact of progressively integrating our contributions: count-weighted detection loss (CW-loss), count-and-similarity branch (CSB), and count and similarity-aware NMS (CAS-NMS). Tab. 3 shows the results. For an extensive comparison, we report results using two standard settings.

Table 4. Comparison (in log-average miss rate) with other loss function on CityPersons Val. set. Our CW-loss outperforms other approaches on both **R** and **HO**.

	Scale	Visiblity	**R**	**HO**
Agg. Loss [32]	×1.3	≥ 50%	11.4	52.6
CW Loss (Ours)	×1.3	≥ 50%	10.8	47.1
Rep. Loss [27]	×1.3	≥ 65%	11.6	55.3
CW Loss (Ours)	×1.3	≥ 65%	11.3	51.5

Table 5. Comparison (in log-average miss-rate) with state-of-the-art methods [13] that improves NMS on CityPersons validation sets.

	Scale	R	HO
Adaptive NMS [13]	×1.0	11.9	55.2
Our CaSe	×1.0	11.0	50.3
Adaptive NMS [13]	×1.3	10.8	54.0
Our CaSe	×1.3	9.6	48.2

We use pedestrians with a height larger than 50 and visibility larger than 65% as training data and an input image scale of 1.3× and 1.0×. Note that we use same network backbone (VGG) for all the experiments in Tab. 3. Our approach yields a significant improvement in performance over the baseline.

We further compare our CW-loss with other loss function [27,32] on CityPersons validation sets. For fair comparison, we use the same set of ground-truth pedestrian examples (visibility) and input scale for training our CaSe when comparing with each method. Tab. 4 shows that our CW-loss outperforms both the Rep. Loss [27] and Agg. loss [32], on **R** and **HO** sets. The results in Tab. 5 demonstrate the effectiveness of our method compared to the Adaptive-NMS[13] using *same* ground-truth pedestrian examples, input scale and backbone.

As described in Sec. 4.2, both the count prediction and similarity embedding are crucial for our CAS-NMS. To validate the impact of the similarity embedding, we perform an experiment by removing the similarity prediction from our CAS-NMS. This leads to inferior results (53.1), likely due to multiple false positive detections, compared to using both count prediction and similarity embedding (50.3) on the **HO** set. Removing the count prediction in our CAS-NMS leads to lower inference speed, highlighting the importance of count [§].

Inference time: For a 1024 × 2048 input, baseline and our CaSe operates at an inference time of 305, 330 milliseconds, respectively. There is only a slight increase in inference time of our CaSe compared to baseline, thanks to the combined utilization of predicted count and similarity embedding in our CAS-NMS. For a fair comparison, both methods are evaluated on a single Titan X GPU.

5.4 CrowdHuman Dataset

Tab. 6 shows the state-of-the-art comparison on the recently introduced CrowdHuman dataset. We use the same protocol to report the results as used in the original dataset [22]. Note that all the methods in Tab. 6 employ the same backbone (ResNet50 + FPN). The Adaptive-NMS [13] and MGAN methods [20] obtain a log-average miss rate of 49.7% and 49.3%, respectively. Our approach sets a new state-of-the-art on this dataset by outperforming both Adaptive-NMS and MGAN methods with a log-average miss rate of 47.9%.

[§] More results are available at https://github.com/Leotju/CaSe.

Table 6. State-of-the-art comparison (in log-average miss rate) on the CrowdHuman dataset. Note that all methods employ the same network backbone (ResNet50 + FPN). Best results are boldfaced. Our detector significantly outperforms the state-of-the-art MGAN, achieving a log-average miss rate of 47.9%.

Method	FPN [22]	Adaptive NMS [13]	MGAN [20]	**CaSe (Ours)**
MR^{-2}	50.4	49.7	49.3	**47.9**

Table 7. Comparison on OCHuman for person instance segmentation. Here, AP_M indicates accuracy (AP) on moderately overlapped ground-truths (IoU with other ground-truths are between 0.5 and 0.75), while AP_H indicates accuracy on heavily overlapped ground-truths (IoU with other ground-truths are larger than 0.75). Our CaSe achieves consistent improvements, on both *val* and *test* sets, over Mask R-CNN.

Method	*val* sets			*test* sets		
	AP	AP_M	AP_H	AP	AP_M	AP_H
Mask RCNN [9,33]	16.3	19.4	11.3	16.9	18.9	12.8
CaSe (Ours)	17.5	20.2	13.0	18.0	20.1	13.9

5.5 Results on Person Instance Segmentation

Finally, we also evaluate our approach for the person instance segmentation task. We integrate our novel components (CW-loss, CSB and CAS-NMS) into Mask-RCNN [9]. We report the results on OCHuman [33], following the same protocol as in [33]. Note that the state-of-the-art [33] for person instance segmentation relies on additional human pose annotation. Tab. 7 shows the comparison of our approach with the baseline Mask RCNN on OCHuman. The results for the baseline and our approach are shown without using human pose information. Our approach outperforms the baseline, in terms of mask AP.

6 Conclusion

We propose an approach by leveraging pedestrian count and proposal similarity information within a two-stage pedestrian detection framework. We introduce a count-weighted detection loss for both the RPN and Fast R-CNN modules of two stage Faster R-CNN. Further, we propose a count-and-similarity branch that predicts both pedestrian count and proposal similarity. Lastly, we introduce a count and similarity-aware NMS strategy to remove duplicate proposals in crowded scenes. Experiments are performed on CityPersons and CrowdHuman datasets. Our results clearly show the effectiveness of our pedestrian detection approach towards handling heavy occlusions. Additionally, we demonstrate the applicability of our components for the problem of human instance segmentation.
Acknowledgment: The work is supported by the National Key R&D Program of China (Grant # 2018AAA0102800 and 2018AAA0102802) and National Natural Science Foundation of China (Grant # 61632018).

References

1. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS – improving object detection with one line of code. In: ICCV (2017) 3
2. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. In: ICCV (2017) 2
3. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: ECCV (2016) 2
4. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. arXiv preprint arXiv:1906.09756 (2019) 11
5. Cao, J., Pang, Y., Han, J., Gao, B., Li, X.: Taking a look at small-scale pedestrians and occluded pedestrians. IEEE Transactions on Image Processing **29**, 3143–3152 (2020) 2
6. Cao, J., Pang, Y., Zhao, S., Li, X.: High-level semantic networks for multi-scale object detection. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2019) 2
7. Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S.Z.X.Z.: Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In: AAAI (2020) 11
8. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. T-PAMI (2012) 9
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017) 14
10. Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: CVPR (2017) 3
11. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR (2018) 3
12. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: ECCV (2018) 3
13. Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: CVPR (2019) 3, 10, 11, 13, 14
14. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: ECCV (2018) 3, 10
15. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: CVPR (2019) 10
16. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: CVPR (2017) 2
17. Mathias, M., Benenson, R., Timofte, R., Gool, L.V.: Handling occlusions with franken-classifiers. In: ICCV (2013) 3
18. Nie, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., Shao, L.: Enriched feature guided refinement network for object detection. In: ICCV (2019) 3
19. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV (2013) 3
20. Pang, Y., Xie, J., Khan, M.H., Anwer, R.M., Khan, F.S., Shao, L.: Mask-Guided attention network for occluded pedestrian detection. In: ICCV (2019) 2, 3, 4, 5, 9, 10, 11, 13, 14
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015) 2, 3, 4
22. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018) 3, 9, 13, 14

23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9
24. Song, T., Sun, L., Xie, D., Sun, H., Pu, S.: Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: ECCV (2018) 10
25. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: ICCV (2015) 2, 3
26. Tychsen-Smith, L., Petersson, L.: Improving object localization with fitness nms and bounded iou loss. In: CVPR (2018) 3
27. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. In: CVPR (2018) 2, 3, 10, 11, 12, 13
28. Xie, J., Pang, Y., Cholakkal, H., Anwer, R.M., Khan, F.S., Shao, L.: Psc-net: Learning part spatial co-occurrence for occluded pedestrian detection. arXiv preprint arXiv:2001.09252 (2020) 4
29. Zhang, J., Lin, L., Li, Y., chen Chen, Y., Zhu, J., Hu, Y., Hoi, S.C.H.: Attribute-aware pedestrian detection in a crowd. arXiv preprint arXiv:1910.09188 (2019) 12
30. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR (2017) 2, 3, 4, 5, 9, 11
31. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in cnns. In: CVPR (2018) 2, 4, 9, 10, 11
32. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In: ECCV (2018) 3, 4, 10, 11, 13
33. Zhang, S.H., Li, R., Dong, X., Rosin, P.L., Cai, Z., Xi, H., Yang, D., Huang, H.Z., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: CVPR (June 2019) 3, 14
34. Zhou, C., Yang, M., Yuan, J.: Discriminative feature transformation for occluded pedestrian detection. In: ICCV (2019) 4, 10, 11
35. Zhou, C., Yuan, J.: Non-rectangular part discovery for object detection. In: BMVC (2014) 3
36. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: ICCV (2017) 2, 3
37. Zhou, C., Yuan, J.: Bi-box regression for pedestrian detection and occlusion estimation. In: ECCV (2018) 2, 3, 4, 10, 11