Progressive Refinement Network for Occluded Pedestrian Detection

Xiaolin Song[1]* Kaili Zhao[1]* Wen-Sheng Chu Honggang Zhang[1] Jun Guo[1]

[1]Beijing University of Posts and Telecommunications, Beijing, China

Abstract. We present *Progressive Refinement Network (PRNet)*, a novel single-stage detector that tackles occluded pedestrian detection. Motivated by human's progressive process on annotating occluded pedestrians, PRNet achieves sequential refinement by three phases: Finding high-confident anchors of visible parts, calibrating such anchors to a full-body template derived from occlusion statistics, and then adjusting the calibrated anchors to final full-body regions. Unlike conventional methods that exploit predefined anchors, the confidence-aware calibration offers adaptive anchor initialization for detection with occlusions, and helps reduce the gap between visible-part and full-body detection. In addition, we introduce an occlusion loss to up-weigh hard examples, and a Receptive Field Backfeed (RFB) module to diversify receptive fields in early layers that commonly fire only on visible parts or small-size full-body regions. Experiments were performed within and across CityPersons, ETH, and Caltech datasets. Results show that PRNet can match the speed of existing single-stage detectors, consistently outperforms alternatives in terms of overall miss rate, and offers significantly better cross-dataset generalization. Code is available.[1]

Keywords: Occluded pedestrian detection, Progressive Refinement Network, anchor calibration, occlusion loss, Receptive Field Backfeed.

1 Introduction

Pedestrian detection is a fundamental computer vision problem and has been widely used in broad applications such as autonomous driving [10], robotics [9], and surveillance [21]. Although promising progress was made, occluded pedestrians remain difficult to detect [18,22,32]. The major challenges involve a wide range of appearance changes due to occlusion by other pedestrians or objects (*e.g.*, cars or trees), which decrease detection accuracy to various extents.

Reviewing the literature, most methods in pedestrian detection handle occlusions by exploiting visible parts as an additional supervision to improve detection performance. These methods broadly leverage three types of designs: 1) Independent detectors trained for each occlusion pattern [6,7,20,23,25,30,34,36], 2) Attention maps to enforce learning on visible parts [27,40], and 3) Auxiliary

[*] These authors contributed equally.

[1] https://github.com/sxlpris

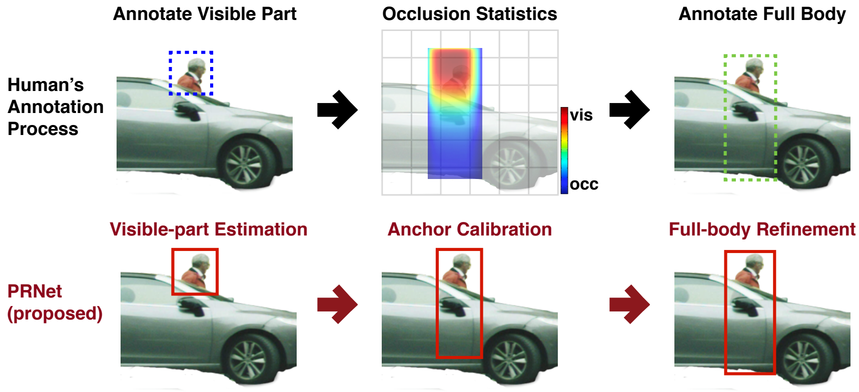


Fig. 1. Progressive Refinement Network (PRNet) imitates human's progressive annotation process on occluded pedestrians (*e.g.*, [5,39]), and gradually infers full-body regions from visible parts.

visibility classifiers to fuse prediction confidence into final scores [22,42,44]. Although these methods could benefit occluded pedestrian detection, at least three issues remain. First, independent detectors are computationally expensive, as each detector is trained for individual occlusion patterns, which are difficult to enumerate in practice. Second, attention-based methods can be slow for inference because attention modules are usually exhausted with proposals in architectures like Faster R-CNN [29]. Attention-based methods emphasize only visible parts, and thus could be suboptimal for full-body detection. Finally, detectors are usually initialized with predefined anchors, which, as will be demonstrated in Sec. 4, are suboptimal to generalize across diverse datasets.

To address the above challenges, we propose Progressive Refinement Network (PRNet), a novel single-stage detector for occluded pedestrian detection. Fig. 1 illustrates our main idea. Inspired by human's progressive annotation process of occluded pedestrians (*e.g.*, [5,39]), PRNet performs pedestrian detection in three phases. First, *visible-part estimation* generates high-confident anchors of visible parts from one single-stage detector (*e.g.*, SSD-based [14,18]). Second, *anchor calibration* adjusts the visible-part anchors to a full-body template according to occlusion statistics, which is derived from over 20,000 annotations of occluded pedestrians. Finally, we train a *full-body refiner* using the calibrated anchors and a separate detection head from the one for visible-part estimation. Using two separate detection heads allows us to fit the progressive design into a single-stage detector without adding much complexity. In addition, to improve training effectiveness, we introduce an occlusion loss to up-weigh hard examples, and a Receptive Field Backfeed (RFB) module to provide more diverse receptive fields, which help shallow layers to detect pedestrians in various sizes. Experiments on three public datasets, CityPersons [39], ETH [8], and Caltech [5], validate the feasibility of the proposed PRNet.

Our contributions in this paper can be summarized as follows:

1. Present a novel Progressive Refinement Network (PRNet) that embodies three-phase progression into a single-stage detector. With helps of the proposed occlusion loss and RFB modules, PRNet achieves competitive results with little extra complexity.
2. Analyze statistically on 20,000 visible-part and full-body regions, and derive an anchor calibration strategy that covers ∼97% occlusion patterns in both CityPersons and Caltech datasets.
3. Offer comprehensive ablation study, and experiments showing that PRNet achieves state-of-the-art within-dataset performance on **R** and **HO** subsets on CityPersons, and the best cross-dataset generalization over ETH and Caltech benchmarks.
4. Provide analysis on extreme occlusions, showing insights behind the metrics and suggesting a realistic evaluation subset for the community.

2 Related Work

CNN-based Pedestrian Detection: Along with the development of CNN-based object detection, pedestrian detection has achieved promising results. We broadly group these methods into two categories: anchor-based and anchor-free.

For anchor-based methods, two-stage detectors (*e.g.*, Faster R-CNN [29]) and one-stage detectors (*e.g.*, SSD [17]) are two common designs. Most two-stage detectors [1,2,11,13,27,35,37,40–42,44] generate coarse region proposals of pedestrians and then refine the proposals by exploiting domain knowledge (*e.g.*, hard mining [37], extra learning task [2,27,40,44], or cascaded labeling policy [1]). RPN+BF [37] used a boosted forest to replace second stage learning and leveraged hard mining for proposals. However, involving such downstream classifier could bring more training complexity. SDS-RCNN [2] jointly learned pedestrian detection and bounding-box aware semantic segmentation, thus encouraged model learning more on pedestrian regions. AR-Ped [1] exploited sequential labeling policy in region proposal network to gradually filter out better proposals. These two-stage detectors need to generate proposal in first stage, and thus are slow for inference in practice. On the other hand, single-stage detectors [14,18,22] enjoy real-time inference due to the one-shot design. GDFL [14] included semantic segmentation task from end to end, which guided feature layers to emphasize on pedestrian regions. Generally, detection accuracy and inference time are trade-offs between single-stage and two-stage detectors. To obtain both accuracy and speed, ALFNet [18] involved anchor refinement into SSD training process. The proposed PRNet takes advantage of high speed of single-stage detector, and simultaneously outperforms these conventional methods in consideration of occlusion-aware supervision.

For anchor-free methods [19,32], topological points of pedestrians and predefined aspect ratio are introduced as new annotations to replace original bbox annotations. TLL [32] predicts the top and bottom vertexes of the somatic topological line while CSP [19] predicts central points and scales of pedestrian instances. Although the above CNN-based pedestrian detectors obtains potential performance, occluded pedestrian detection is still a challenging problem.

Occluded Pedestrian Detection: Methods tackling occluded pedestrians can be broadly categorized into four types: part-based, attention-based, score-based, and crowd-specific. Part-based methods have been widely received in the community, where each detector was separately trained for individual occlusion pattern with inference done by fusing all predictions. See [6, 7, 20, 23–25, 30, 34, 36, 43] for comprehensive reviews. Moreover, exhaustively enumerating occlusion patterns is non-practical, computationally expensive, and generally infeasible. Instead of considering each occlusion pattern, [22, 41] partitioned a proposal or bounding box into fixed number parts and predict their visibility scores. Although training complexity was decreased, these methods still require manually designing the partitions.

In recent years, learning robust representations and better anchor scoring have become a popular topic. On one hand, attention-based methods [27, 40] learn robust features using guidance from attention maps. Zhang et al. [40] and MGAN [27] exploited channel-wise and pixel-wise attention maps respectively in feature layers, to highlight visible parts and suppress occluded parts. However, emphasizing visible-parts solely could be sub-optimal for full-body prediction. On the other hand, score-aware methods learn extra anchor scores by introducing additional learning task in the second stage of Faster R-CNN. For instance, Bi-box [44] constructs two classification and regression tasks for visible-part and full-body anchors, and then fuses the two anchor scores during inference. Similarly, [42] uses a separate discriminative classification by enforcing heavily occluded anchors to be close to easier anchors, and high confident scores were obtained for anchors. In addition, other studies [16, 26, 28, 33, 35] focused on tackling crowded pedestrians. RepLoss [35] designs a novel regression loss to prevent target proposals from shifting to surrounding pedestrians. The aforementioned methods are generally initialized with predefined anchors. In contrast, the proposed PRNet learns occluded pedestrian detection under confidence-aware and adaptive anchor initialization, which helps improve detection accuracy and generalization across dataset.

3 Progressive Refinement Network (PRNet)

3.1 PRNet Architecture

Motivated by human's progressive process on annotating occluded pedestrians (*e.g.*, CityPersons [39] and Caltech [5]), we construct PRNet to gradually migrate high-confident detection on visible parts toward more challenging full-body localization. For this purpose, we propose to adopt a single-stage detector with three training phases: Visible-part Estimation (VE), Anchor Calibration (AC), and Full-body Refinement (FR). Unlike most methods that detect full bodies only [18, 35] or independently with visible parts [44], we interweave them into a single-stage framework. To bridge the detection gap between visible parts and full bodies, we introduce AC to align anchors from VE to FR.

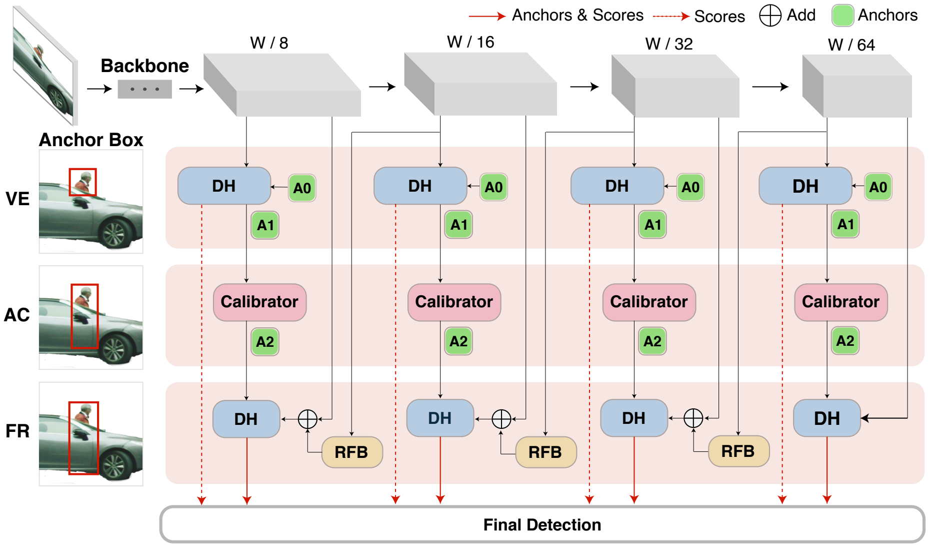
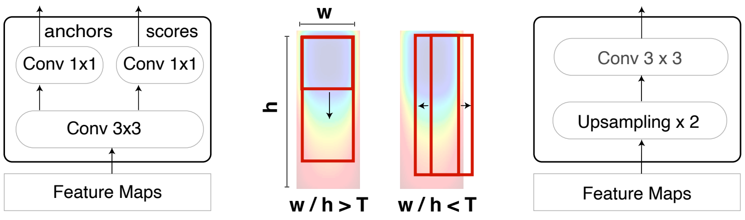


Fig. 2. Architecture of PRNet. From top to bottom, PRNet uses a detection backbone illustrated with four blocks of features maps. The network is trained in three phases: **Visible-part Estimation** (VE), **Anchor Calibration** (AC), and **Full-body Refinement** (FR). VE and FR take visible-part and full-body ground truth as references, respectively. Given initial anchors (A0), VE learns to predict visible-part anchors (A1), which are improved by AC to obtain calibrated anchors (A2). Final detection is obtained by post-processing anchors and scores from VE and FR. Detection Head (DH), Calibrator, and RFB modules are depicted in Fig. 3 and detailed in Sec. 3.1.

Fig. 2 illustrates the PRNet architecture. The top row depicts the backbone, where we truncated first 5 stages of ResNet-50 [12] with modification of appending 1 extra stage with 3x3 filters and stride 2, which provide diverse receptive fields and help capture pedestrian with various scales. Out of the 6-stage backbone, we treat the last four as detection outputs. The network is trained following three phases: Visible-part Estimation (VE), Anchor Calibration (AC), and Full-body Refinement (FR). VE and FR are trained with visible-part and full-body ground truth, respectively; AC leverages occlusion statistics to bridge the gap between visible-part anchors and full-body anchors. Details of each module are illustrated later in this section. On top of each detection layer, we attach a detection head (DH) separately for VE and FR.

Specifically, denote x as an input image, $\Phi(x)$ as feature maps from backbone, $\mathcal{A}_0$ as a set of predefined anchors (as in SSD [17]), $\mathcal{B}^*$ as the predicted bounding boxes that are obtained by post-processing anchors collected from all layers (*i.e.*, via Non-Maximum Suppression). Given an initial set of feature maps and anchors, PRNet can be formulated as a progressive detector:

$$\text{Detections} = F(E_f(C(E_v(\Phi(x), \mathcal{A}_0)))) = \{\mathcal{B}^*, s^*\}, \tag{1}$$



(a) Detection Head **(b) Calibrator** **(c) RFB**

Fig. 3. Modules used in PRNet architecture (as in Fig. 2): DH, Calibrator and RFB.

where $E_v(\Phi(x), \mathcal{A}_0)$ is the 1st-phase visible-part estimation (VE) whose outputs are a set of visible-part anchors and confidence scores $\{\mathcal{A}_1, s_1\}$, $C(\cdot)$ is the 2nd-phase anchor calibration (AC) that aligns visible-part anchors $\mathcal{A}_1$ to full-body anchors $\mathcal{A}_2$, and $E_f(\Phi(x), \mathcal{A}_2)$ is the 3rd-phase full-body refiner (FR) that outputs the final full-body anchors to compute $\mathcal{B}^*$ and their scores s^* using inference F (see Sec. 3.2). Note that $\Phi(x)$ represents different feature maps during VE and FR due to their complementary objectives. Below we discuss each phase in turn.

Visible-part Estimation (VE): To train the *visible-part estimation* $E_v(\cdot)$, we adopt a standard detection approach that learns to localize anchors $\mathcal{A}_1$ as regression (from predefined anchors $\mathcal{A}_0$), and anchor scores as classification. Fig. 3(a) depicts the detection head, whose loss can be written as:

$$\mathcal{L}_{VE} = \mathcal{L}_{focal} + \lambda_v [y = 1] \mathcal{L}_{smoothL1}, \tag{2}$$

where $\mathcal{L}_{focal}$ is focal loss [15] for classification, $\mathcal{L}_{smoothL1}$ is a smooth-L1 loss for regression (as adopted in Faster R-CNN [29]), $[y=1]$ is an indicator for positive samples, and λ_v is a tuning parameter. As VE is trained on visible parts, its prediction (*i.e.*, $\mathcal{A}_1$) on visible parts is generally more confident and accurate than detectors trained with occlusions.

Anchor Calibration (AC): After VE obtains confident visible-part anchors $\mathcal{A}_1$, we propose a simple and effective *anchor calibration* $C(\cdot)$ to migrate visible-part anchors toward full-body anchors $\mathcal{A}_2$, which are then passed to the next phase for bull-body refinement. Briefly, PRNet updates anchors as: $\mathcal{A}_0 \xrightarrow{E_v} \mathcal{A}_1 \xrightarrow{C} \mathcal{A}_2$. Three are our motivations:

1. The aspect ratio of visible-part boxes is much more diverse than that of full-body boxes [5, 39], making regression from visible-part to full-body boxes rather challenging.
2. Adaptive anchor initialization can reduce unnecessary search space and lead to better detection (*e.g.*, [3]), compared to most methods that use predefined anchors (*e.g.*, [18, 39, 41, 42]).
3. The IoU discrepancy between visible-part anchors and full-body ground truth boxes is large; proper calibration can significantly improve IoU.

Fig. 4 shows the distribution of IoU between ground truth full-body boxes and visible-part boxes before/after Anchor Calibration (AC) in CityPersons dataset [39]. The visible-part boxes before AC were taken from the annota-

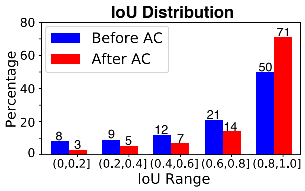


Fig. 4. IoU distribution before and after anchor calibration on the CityPersons dataset. IoU is measured between anchors and full-body ground truth.

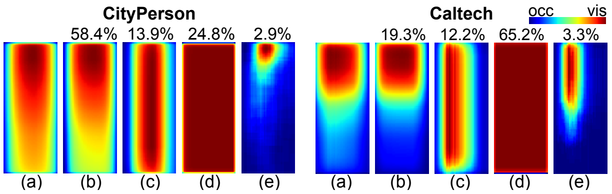


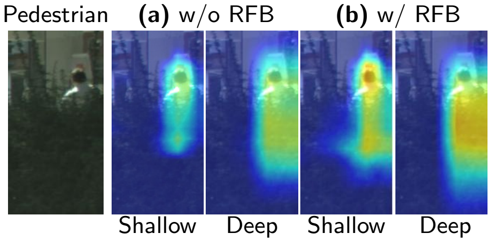
Fig. 5. Occlusion statistics from CityPersons [39] (left) and Caltech [5] (right): (a) Occlusion statistics with blue indicating occlusion; red indicates visible parts, (b) Horizontal occlusions, (c) Vertical occlusions, (d) Non-occlusion, (e) Others. Percentage (%) denotes the likelihood of each occlusion pattern.

tions in the original dataset. As can be seen, calibration significantly shifts the distribution toward higher IoU, *e.g.*, +21% for IoU in (0.8, 1.0], and thus can help detectors approximate final full-body regions. In addition, AC addresses discrepancy during anchor assignment between VE and FR, *i.e.*, without AC, a positive $\mathcal{A}_1$ could be assigned as a negative anchor for FR, making VE and FR fail to complement each other.

To achieve AC, we first derive a statistical analysis of occlusion patterns on two popular datasets CityPersons [39] and Caltech [5] using their standardized 0.41 box aspect ratio. Please see supplementary materials for detailed process. Fig. 5 illustrates occlusion distribution over a full-body box and four occlusion types (*i.e.*, horizontal, vertical, non-occlusion, and others, similar to [40]) with respective likelihood in each dataset. As can be seen in Fig. 5(a), over the two datasets, the upper box is consistently visible (*i.e.*, the head), with most occlusions appearing in the lower box (*i.e.*, the feet). This serves as strong evidence for humans and PRNet to leverage visible parts for full-body detection.

Observing the occlusion statistics, we reach two types of anchor updates according to the aspect ratio of $\mathcal{A}_1$, as depicted in Fig. 3(b). For the anchors with ratio >0.41, we vertically stretch them *downwards* until 0.41 aspect ratio, due to heads being frequently visible, as shown in Fig. 5(b) and [5]. Anchors with ratio <0.41 are horizontally extended to 0.41 w.r.t. the center of $\mathcal{A}_1$, as they likely involve vertical occlusion, as shown in Fig. 5(c). Anchors with 0.41 ratio (*i.e.*, Fig. 5(d)) remain unchanged. The anchor updates can also be rationalized with human's annotation protocol in CityPersons [39], where a full-body box is generated by fitting a fixed-ratio (0.41) box onto a line drawn from head to feet. According to Fig. 5(b)-(d), we justify the two simple updates can cover ∼97% data in both datasets, while the remaining ∼3% is shown in Fig. 5(e).

Full-body Refinement (FR): With the calibrated anchors $\mathcal{A}_2$ from AC, PRNet's last phase trains a *full-body refiner* $E_f(\cdot)$ that refines the final full-body localization. Similar to VE, FR also uses the same backbone, yet performs training on a separate detection head. Different from VE that sees only visible parts, FR starts to see hard positive samples whose anchor boxes are still far



.

Fig. 6. Saliency maps highlighted by the third FR phase: (a) w/o RFB and (b) w/ RFB. "Shallow" indicates the 2nd layer, and "Deep" indicates the 3rd layer.

from ground truth full-body region. As $\mathcal{L}_{smoothL1}$ in Eq. (2) treats every positive sample equally, it could be less effective when dealing with hard samples in FR. To encourage learning on hard positive samples, we weigh the regression loss $\mathcal{L}_{smoothL1}$ with an occlusion weight, which is defined as a reverse IoU between $\mathcal{A}_2$ and ground truth full-body boxes $\mathcal{B}_{gt}$. Given $a \in \mathcal{A}_2$ and its corresponding $b \in \mathcal{B}_{gt}$, the weighted loss, termed as *occlusion loss*, can be rewritten as:

$$\mathcal{L}_{occ} = \sum_{a \in \mathcal{A}_2} (1 - \text{IoU}(a,b)) \left\{ [|s| < 1]\, 0.5 s^2 + [|s| >= 1]\, (|s| - 0.5) \right\}, \quad (3)$$

where s is the difference between predicted offsets and ground truth offsets (see [29] for details). The less overlap between the calibrated anchors $\mathcal{A}_2$ and $\mathcal{B}_{gt}$, the higher $\mathcal{L}_{occ}$ is. As a result, the loss for FR becomes:

$$\mathcal{L}_{FR} = \mathcal{L}_{focal} + \lambda_f [y = 1] \mathcal{L}_{occ}. \quad (4)$$

Despite of up-weighting hard positive anchors, another challenge in FR regards training shallow layers, which often activate on visible parts or small-size full-body regions due to limited receptive field. In every layer of FR, we introduce a *Receptive Field Backfeed* (RFB) module to diversify receptive fields, as depicted in Fig. 3(c). RFB aims to enlarge the receptive fields of shallower layers by back-feeding features from deeper layers to the previous layer with 2X upsampling, and then summing up their feature maps in a pixelwise manner.

Fig. 6 shows the saliency maps [31] of the 2nd layer (denoted as "shallow") and the 3rd layer (*i.e.*, "deep") with/without the RFB module. As can be seen in Fig. 6(a), without RFB, visible parts are identified in the shallow layer, while the deeper layer emphasizes full-body regions. The effects of RFB can be clearly observed in Fig. 6(b). In the shallow layer, RFB not only enhances visible parts but also complements the full-body region. Similar observation can be made on the deep layer, showing that RFB can propagate larger receptive fields to shallower layers and help refine full-body detection.

3.2 Training & Inference

Training: In training, a batch of pedestrian images goes through the three phases (*i.e.*, VE, AC, and FR) sequentially–the first phase VE is trained independently and then the first detection head is frozen to train FR. Fig. 2 illustrates the architecture and examples of pedestrian annotation. Given predefined anchors $\mathcal{A}_0$ and visible-part ground truth boxes associated to the image batch, we first train VE with loss $\mathcal{L}_{VE}$ in Eq. (2), and obtain visible-part anchors $\mathcal{A}_1$. Then AC transforms $\mathcal{A}_1$ into more adaptive anchors $\mathcal{A}_2$, which better approximates full-body regions. Finally, initialized with $\mathcal{A}_2$, FR is trained with loss $\mathcal{L}_{FR}$ in Eq. (4). Note that VE and FR use two different detection heads in one single-stage detector, so they learn complementary outputs.

An anchor is assigned as positive if intersection-over-union (IoU) between an anchor bbox and ground truth bbox is above a threshold θ_p, as negative if IoU is lower than θ_n, and otherwise ignored during training. Note that VE and FR adopt different annotation of boxes, *i.e.*, VE consumes visible-part boxes, while FR uses full-body boxes.

Inference: In inference, we obtain predicted anchor boxes from FR, and associate anchor scores by multiplying the scores from VE and FR. The score fusion provides complementary guidance so to improve detection robustness (similar to [44]). We obtain the final bounding boxes $\mathcal{B}^*$ by first filtering out candidate anchor boxes with scores lower than 0.05 and then merging them with NMS (0.5 threshold is used here).

3.3 Comparisons with Related Work

The closest studies to PRNet are ALFNet [18] and Bi-box [44]. As most cascade designs, ALFNet tackles successively the same task (FR→FR), which requires occlusion patterns to be extensively illustrated in training data. Mimicking human's annotation process, PRNet exploits different tasks (VE→AC→FR), starting from detecting only visible parts (regardless of occlusion patterns as in full-body boxes), and thus relaxes training data requirements. Note that jointly tackling different tasks is non-trivial. Instead, we interweave these tasks with occlusion loss and the RFB module (Sec. 3.1) to up-weigh hard samples and facilitate training for shallow layers. As can be seen in Sec. 4.4, PRNet achieved impressive cross-dataset generalizability compared to ALFNet, showing that the PRNet structure is more effective. Similar to ALFNet [18], PRNet enjoys competitive inference time due to the use of a single-stage detector.

In terms of involving different tasks, Bi-box [44] also takes visible parts into account but by training a two-branch detector for visible parts and full body in the second stage of Faster R-CNN. During training, there is no interaction between the two branches, making their complementary benefits relatively indirect. PRNet leverages the hybrid cascade structure to progressively refine predictions from visible-part to full-body regions, providing adaptive anchor initialization to achieve the final full-body estimation.

4 Experiments

4.1 Settings

Datasets: We conducted experiments on three public datasets: CityPersons [39], ETH [8], and Caltech [5]. CityPersons [39] has high-res 2048×1024 images with

Table 1. Ablations of **three-phase components** and an alternative.

Architecture	VE	AC	FR	R	HO
PRNet-F			✓	15.6	45.7
PRNet-VA	✓	✓		11.7	51.3
PRNet-VAF	✓	✓	✓	11.4	45.3
PRNet-VRF	✓	reg	✓	12.6	44.7

Table 2. Ablations of **occlusion loss** and the **RFB module**.

Architecture	+Occ.	+RFB	R	HO
PRNet-VAF			11.4	45.3
PRNet-VAF-OCC	✓		11.0	45.7
PRNet-VAF-RFB		✓	11.6	44.9
PRNet (ours)	✓	✓	10.8	42.0

visible-part and full-body annotations, where 2,975 images are for training and 500 for validation. We trained PRNet on the training set and reported performance on the validation set in ablations and within-dataset experiments. To evaluate model generalizability, we performed cross-dataset analysis using ETH [8] and Caltech [5]. ETH dataset [8] contains 11,941 labeled persons, providing a benchmark in evaluating model's robustness to occluded pedestrians. For Caltech [5], we adopted published test set with 4,024 images with both old [5] and new annotations [38]. Both ETH and Caltech have lower-res 640×480 images that represent more cross-dataset challenges. Following [35, 40], we performed training and evaluation on pedestrians with height larger than 50 pixels.

Metrics: Evaluation was reported on the standard MR^{-2} (%) [5], which computes the log-average miss rate at 9 False Positive Per Image (FPPI). The lower MR^{-2}, the better. To ensure the results are directly comparable with the literature, we represented each test set as *6 subsets* according to visibility ratio of each pedestrian. Specifically, we reported **R** (reasonable occlusion with visibility in [0.65,1]), **HO** (heavy occlusion with [0.2, 0.65]), **R+HO** with [0.2, 1] from Zhang *et al* [40], and **Bare** with [0.9, 1.0], **Partial** with [0.65, 0.9], and **Heavy** with [0, 0.65] from [35]. To complement the visibility range covered by **R** and **HO**, we added **EO** (extreme occlusion) to represent visibility in [0, 0.2].

Implementation details: We augmented our pedestrian images following standard techniques [18, 19]. When assigning labels to anchor boxes, $\theta_p = 0.5$ and $\theta_n = 0.3$ for VE, and $\theta_p = 0.7$ and $\theta_n = 0.5$ for FR. We set $\lambda_v = 1$ and $\lambda_f = 4$ empirically. The backbone ResNet-50 is pre-trained on ImageNet [4]. PRNet is then fine-tuned with 160k iterations, a learning rate of 10^{-4}, batch size 8 and an Adam optimizer. All experiments were performed on 2 GTX 1080Ti GPUs.

4.2 Ablation Study

To analyze PRNet, we performed extensive ablations on CityPersons validation set [39] using subsets of R (reasonable) and HO (heavy occlusion).

Three-phase components: To analyze the effect of PRNet's three-phase design, we performed ablation study on each phase without occlusion loss and RFB module in FR. In Table 1, we trained a standalone FR (denoted as **PRNet-F**) initialized by predefined full-body anchors. **PRNet-VA** used only VE+AC, treating calibrated anchors $\mathcal{A}_2$ as the detection outputs. **PRNet-VAF** employed all three phases (VE+AC+FR), using calibrated anchors $\mathcal{A}_2$ to initialize FR. Comparing PRNet-F and PRNet-VA, PRNet-VA performs 3.9 points better in

R while 5.6 points worse in HO. This shows that plain calibrated anchors $\mathcal{A}_2$ in PRNet-VA can achieve better result while occlusion level is reasonable. In contrast, PRNet-F better addressed heavy occlusions. PRNet-VAF combines the benefits from both, showing a consistent improvement over both R and HO. Please see supplementary for detection examples of the three-phase progression.

Anchor calibration *vs.* box regression: A possible alternative to AC is a box regressor from the visible-part anchors $\mathcal{A}_1$ to full-body boxes. Here we reused FR for the regression task. To perform a fair comparison, we implemented **PRNet-VRF** by replacing AC with the regressor. Table 1 summarizes the results. As can be seen, PRNet-VAF consistently outperformed PRNet-VRF by 9.5% in R, showing no significant benefits of adding an extra box regressor. An explanation can be that the visible boxes change rapidly due to various occlusion types, and make the regressor hard to map the coordinates to full-body boxes with relatively constant aspect ratio. Unlike a regression network that require extra complexity and training efforts, AC provides a more generalizable strategy that better fits into the proposed three-phase approach.

Occlusion loss and RFB: Table 2 studies PRNet w/ and w/o occlusion loss and RFB module. PRNet-VAF was reused as the baseline that considers neither occlusion loss nor RFB, and compared against **PRNet-VAF-OCC** (with only occlusion loss) and **PRNet-VAF-RFB** (with only RFB). Including occlusion loss alone, PRNet-VAF-OCC improved the baseline 0.4 points on R yet lowered 0.4 points on HO. This shows that occlusion loss improves detection with reasonable occlusion (*i.e.*, over 0.65 visibility), yet could be insufficient to address heavy occlusion (*i.e.*, 0.2 to 0.65 visibility). Including RFB alone, PRNet-VAF-RFB improved the baseline 0.4 points on HO yet lowered 0.2 points on R. This suggests that the feedback from RFB could supply full-body info by enlarging receptive field, and thus offers improvement when occlusion is severe. Otherwise, when occlusion level is light, enlarging receptive field may introduce unnecessary context and hence slightly hurt. PRNet couples occlusion loss and RFB, achieving significant improvement over R and especially HO.

4.3 Within-dataset Comparisons

This section compares PRNet in a within-dataset setting against 3 types of alternatives: Occlusion-free, occlusion-aware, and closest to PRNet. We reported MR^{-2} on all 6 subsets, where R is the major evaluation criteria in CityPersons Challenge[2]. Table 3 shows comparisons with scale ×1 and ×1.3 of original resolution (2048×1024).

Occlusion-free methods: *Occlusion-free* methods aim to detect pedestrians without considering occlusion info. Adapted FasterRCNN [39] is an anchor-based benchmark, while TLL+MRF [32] and CSP [19] are anchor-free. Among the three methods, CSP achieved the state-of-the-art results without considering occlusion, as summarized in Table 3. PRNet, on the other hand, takes occlusion info into account, and provides performance gain over CSP on R, Partial, and

[2] https://bitbucket.org/shanshanzhang/citypersons/

Table 3. Comparisons on CityPersons [39]. Results of alternatives were obtained from original paper. On scale×1, bracketed and bold numbers indicate the best and the second best results, respectively. Inference time (*sec*) is measured on scale×1 images.

Method	Occ.	Scale	**R**	**HO**	**R+HO**	**Heavy**	**Partial**	**Bare**	**Time**
Adapted FasterRCNN [39]		×1	15.4	64.8	41.45	55.0	18.9	9.3	-
TLL+MRF [32]		×1	14.4	-	-	52.0	15.9	9.2	-
CSP [19]		×1	**11.0**	-	-	[**49.3**]	**10.4**	7.3	0.33
FasterRCNN+ATT [40]	✓	×1	16.0	56.7	38.2	-	-	-	-
RepLoss [35]	✓	×1	13.2	-	-	56.9	16.8	7.6	-
		×1.3	11.6	-	-	55.3	14.8	7.0	-
OR-CNN [41]	✓	×1	12.8	-	-	55.7	15.3	[**6.7**]	-
		×1.3	11.0	-	-	51.3	13.7	5.9	-
MGAN [27]	✓	×1	11.3	[**42.0**]	-	-	-	-	-
FRCN+A+DT [42]	✓	×1.3	11.1	44.3	-	-	11.2	6.9	-
ALFNet [18]		×1	12.0	43.8	**26.3**	**51.9**	11.4	8.4	0.27
Bi-box [44]	✓	×1.3	11.2	44.2	-	-	-	-	-
PRNet (ours)	✓	×1	[**10.8**]	[**42.0**]	[**25.6**]	53.3	[**10.0**]	**6.8**	0.22

Bare subsets, but not the Heavy subset. One possible reason is because CSP used box-free annotations, which is different from the original annotations and might help reduce ground truth noises in heavily occluded cases.

Occlusion-aware methods: *Occlusion-aware* methods consider occlusion information in training, including FasterRCNN+ATT [40], RepLoss [35], OR-CNN [41], FRCN+A+DT [42], and MGAN [27]. Table 3 summarizes the results. On the R subset (CityPersons' evaluation criteria), occlusion-aware methods are generally better than occlusion-free methods, except for CSP that used different box-free annotations. In contrast, PRNet consistently achieved the best MR^{-2} of (10.8, 42.0, 25.6, 10.0) on (R, HO, R+HO, Partial) and compared favorably with the best performer for Bare. The comparisons firmly validate PRNet's effectiveness by dealing with occlusion using progressive refinement.

Closest alternatives: Closest to PRNet are ALFNet [18] and Bi-box [44] per discussion in Sec. 3.3. We reported ALFNet results using the same settings and the authors' released code. We did not reproduce Bi-box due to lack of source code. Regarding inference time, PRNet performed comparably with ALFNet, as both methods are single-stage based. We infer that PRNet is substantially faster than Bi-box due to the Faster-RCNN-like design in Bi-box (*e.g.*, 2-6X speedup as demonstrated in [1, 18]). Compared to ALFNet and Bi-box, PRNet is also preferred in detection performance because of better anchor initialization and its ability to recover full-body region from confident visible parts. Due to space constraint, please refer to supplementary for examples that are mis-detected by alternative methods but successfully detected by PRNet. Observing the last three rows in Table 3, PRNet consistently outperformed Bi-box and provided performance gain upon ALFNet in all cases except for the Heavy subset.

Breakdowns in Heavy: In the Heavy subset, we noticed the occlusion-aware methods, including PRNet, were less effective than occlusion-free methods (*e.g.*, ALFNet). We performed an analysis by partitioning Heavy into HO and EO, *i.e.*, Heavy=HO $\cup$ EO. EO represents the most extreme occlusion with visibility in only [0, 0.2]. In HO, PRNet outperformed all other methods (*e.g.*, 1.8

Table 4. Cross-dataset on ETH [8].

Method	R+HO	Time
FasterRCNN [39]	35.6	-
FasterRCNN+ATT [40]	33.8	-
CSP [19]	37.2	61.3
ALFNet [18]	**31.1**	39.2
PRNet (ours)	[**27.0**]	42.1

Table 5. Cross-dataset results on Caltech [5].

Method	R (o)	R+HO (o)	R (n)	Time
ALFNet [18]	25.0	35.0	19.0	39.2
CSP [19]	**20.0**	[**27.8**]	**11.7**	61.3
PRNet (ours)	[**18.3**]	**28.4**	[**10.7**]	42.1

points better than the state-of-the-art ALFNet), while being 1.4 points worse than ALFNet in Heavy. We hypothesize that PRNet fails to compete against ALFNet only in EO, and re-evaluated their performance on EO. Not surprisingly, PRNet and ALFNet result in very high MR^2 at 80.8 and 70.2 respectively. Fig. 7 shows the distribution of visibility ratio and examples of EO from CityPersons validation set. As can be seen, visible parts are barely visible and sometimes very low-res, making it perceptually challenging even for human to detect. Ground truth boxes by human annotators in EO can thus be noisy and make performance comparisons on EO less meaningful. In addition, the proportion of EO is relatively small. As shown in top-left of Fig. 7, less than 10% are in EO and more than 90% belong to R and HO. These findings reveal that R and HO render more realistic occlusion scenarios than EO. The above analyses suggest the proposed PRNet achieved state-of-the-art performance.

4.4 Cross-dataset Generalization

To validate generalizability of the proposed method, we performed cross-dataset experiments on ETH [8] and Caltech [5] datasets. For comparison, we picked two top-performing methods, CSP [19] and ALFNet [18], where the models are available from the authors' GitHub release. For fair comparisons, PRNet was also trained on CityPersons training set and shared the same pre-processing.

Table 4 shows ETH results on the R+HO as in [40]. For reference, we also included numbers reported in the Faster-RCNN and FasterRCNN+ATT [40] without reproducing their results. CSP and ALFNet showed surprising opposite results comparing their performance within- and cross-dataset. In cross-dataset setting, ALFNet outperformed CSP by 6.1 points (from 37.2 to 31.1), while CSP reported consistently better performance in within-dataset setting (see Table 3). On the contrary, our method achieved the state-of-the-art MR^{-2} on R+HO by a significant margin. For Caltech [5], we reported R+HO and R using the old [5] (denoted as "(o)") and the new [38] annotations (denoted "(n)"), as summarized in Table 5. PRNet consistently outperformed CSP and ALFNet in R using both old and new annotations. On R+HO, PRNet performed comparably with CSP.

Rationale: PRNet's gain is evident in cross-dataset settings for two major reasons. One, PRNet's progressive structure imitates human's annotation process, which formulates the principles humans have established for annotating occluded pedestrians (*e.g.*, CityPersons, Caltech). PRNet mimics every step in human's principles, and thus fits the problem more naturally. Two, most methods (*e.g.*, ALFNet) consider only full body detection, which demands training data

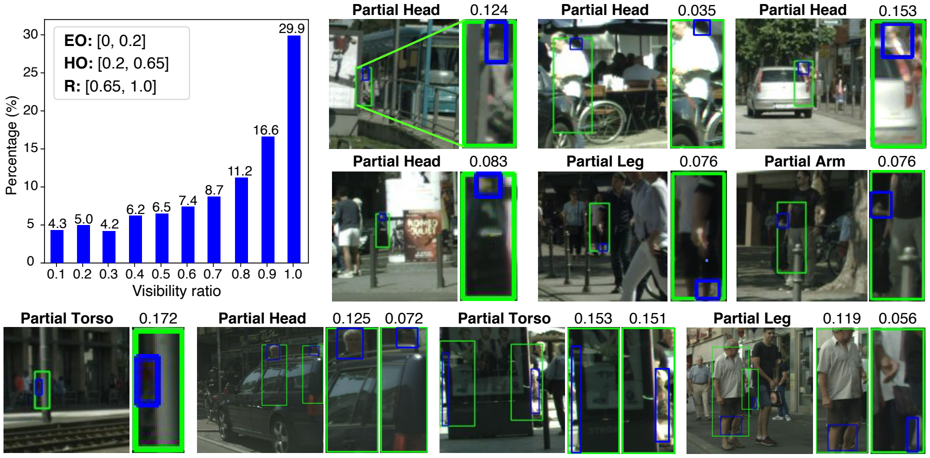


Fig. 7. Distribution of visibility ratio on CityPersons (top left), and examples of Extreme Occlusion (**EO**), such as partial head, arm, leg, and torso. Blue and green boxes indicate visible parts and full-body boxes, respectively.

with various occlusions (*e.g.*, cars, trees, other pedestrians). When the occlusion pattern is rare or unseen in training data (*i.e.*, cross-dataset settings), such methods tend to perform less favorably. As shown in supplementary, ALFNet tends to fire false positives on uncommon objects (*e.g.*, wheel, car windshield). On the contrary, PRNet propagates detection from visible parts (regardless of occlusion patterns as in full-body boxes), and thus provides better generalizability.

5 Conclusion

We have proposed PRNet, a novel one-stage approach for occluded pedestrian detection. PRNet incorporates three phases (VE, AC, and FR) to evolve anchors toward full-body localization. We introduced an occlusion loss to encourage learning on hard samples, and an RFB module to diversify receptive fields for shallow layers. We provided extensive ablation studies to justify the three-phase design. Within-dataset experiments validated PRNet's effectiveness with 6 occlusion scenarios. On cross-dataset settings, PRNet outperformed alternatives on ETH and Caltech datasets by a noticeable margin. Analysis on extreme occlusions provided insights behind metrics and suggested a more realistic choice for evaluation. Potential extensions of PRNet include providing weak annotations of visible parts for occluded pedestrian datasets.

Acknowledgement. Research reported in this paper was supported in part by the Natural Science Foundation of China under grant 61701032 to KZ, and Beijing Municiple Science and Technology Commission project under Grant No.Z181100001918005 to XS and HZ. We thank Jayakorn Vongkulbhisal for helpful comments.

References

1. Brazil, G., Liu, X.: Pedestrian detection with autoregressive network phases. In: CVPR (2019)
2. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. In: ICCV (2017)
3. Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S.Z., Zou, X.: Selective refinement network for high performance face detection. In: AAAI (2019)
4. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
5. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. TPAMI **34** (2012)
6. Duan, G., Ai, H., Lao, S.: A structural filter approach to human detection. In: ECCV (2010)
7. Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.M.: Multi-cue pedestrian classification with partial occlusion handling. In: CVPR (2010)
8. Ess, A., Leibe, B., Van Gool, L.: Depth and appearance for mobile scene analysis. In: ICCV (2007)
9. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32** (2013)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
11. Girshick, R.: Fast R-CNN. In: ICCV (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. TMM **20** (2018)
14. Lin, C., Lu, J., Wang, G., Zhou, J.: Graininess-aware deep feature learning for pedestrian detection. In: ECCV (2018)
15. Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: CVPR (2017)
16. Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: CVPR (2019)
17. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: ECCV (2016)
18. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: ECCV (2018)
19. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: CVPR (2019)
20. Mathias, M., Benenson, R., Timofte, R., Gool, L.V.: Handling occlusions with franken-classifiers. In: ICCV (2013)
21. Nascimento, J.C., Marques, J.S.: Performance evaluation of object detection algorithms for video surveillance. TMM **8** (2006)
22. Noh, J., Lee, S., Kim, B., Kim, G.: Improving occlusion and hard negative handling for single-stage pedestrian detectors. In: CVPR (2018)
23. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR (2012)
24. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV (2013)

25. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship in pedestrian detection. In: CVPR (2013)
26. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: CVPR (2013)
27. 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)
28. Pepikj, B., Stark, M., Gehler, P., Schiele, B.: Occlusion patterns for object class detection. In: CVPR (2013)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015)
30. Shet, V.D., Neumann, J., Ramesh, V., Davis, L.S.: Bilattice-based logical reasoning for human detection. In: CVPR (2007)
31. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 (2013)
32. 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)
33. Tang, S., Andriluka, M., Schiele, B.: Detection and tracking of occluded people. IJCV **110** (2014)
34. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: ICCV (2015)
35. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. In: CVPR (2018)
36. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV (2005)
37. Zhang, L., Lin, L., Liang, X., He, K.: Is faster r-cnn doing well for pedestrian detection? In: ECCV (2016)
38. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR (2016)
39. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR (2017)
40. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in CNNs. In: CVPR (2018)
41. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In: ECCV (2018)
42. Zhou, C., Yang, M., Yuan, J.: Discriminative feature transformation for occluded pedestrian detection. In: ICCV (2019)
43. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: ICCV (2017)
44. Zhou, C., Yuan, J.: Bi-box regression for pedestrian detection and occlusion estimation. In: ECCV (2018)