Supplementary Material for 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. This material includes more results that cannot be fitted into the main paper due to space limitation. We first illustrate details of occlusion statistics including methods and more statistics on different partitions of datasets. Then, we show examples progressively detected by three components of PRNet and the ones of full PRNet, varifing that PRNet can handle various occlusions. In addition, we display examples that are mis-detected by alternative methods but successfully detected by PRNet. At last, we provide qualitative results of cross-dataset generalization on ETH [2] and Caltech [1] dataset.

1 Occlusion statistics

Here, we illustrate how we obtained occlusion statistics in Sec. 3.1 of the main paper, and study more occlusion statistics on the remaining sets of CityPersons [6] and Caltech [1]: val/test and train + val/test, aside from only training set or training+val set in main paper.

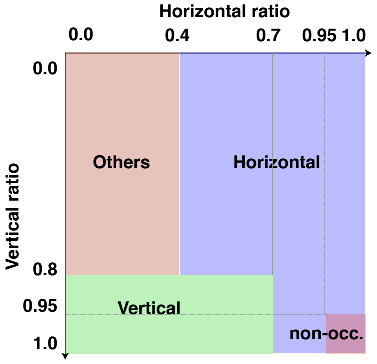

Fig. 1. Overlapping ratio designs for four types. With the designs, the obtained occlusion distributions are consistent with [1, 5, 6].

⋆ These authors contributed equally.

There are three steps to obtain occlusion statistics for a given pedestrian dataset. For each annotated pedestrian, we first pad zero outside of its visible box to re-scale its size as same as a fixed full-body box, whose w/h equals to standardized 0.41 [1, 5, 6]. Then we generated an image from such the padded box, where each pixel will be assigned to be 1 if it locates inside original visible box, otherwise will be 0. Last, we accumulated all the generated images, and thus occlusion statistics (*i.e.* the accumulated image) were obtained. In this way, each pixel's value of occlusion statistics map represents likelihood of visibility, the higher the value is, the more frequently objects in this pixel can be visible.

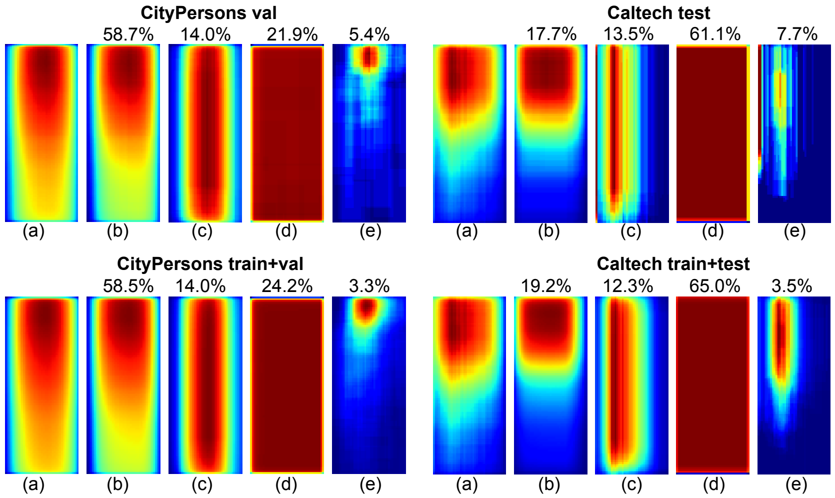

Fig. 2. Occlusion statistics from CityPersons [6] (left) and Caltech [1] (right) datasets. The first and second lines separately show statistics of test set and complete set. Note that val set is regarded as test set in existing publications. We show four main occluded types as follows: (a) Occlusion statistics with blue indicating occlusion and red indicating visible parts, (b) Horizontal occlusions, (c) Vertical occlusions, (d) Non-occlusion, and (e) Others. Percentage (%) denotes the likelihood of each occlusion pattern. Color on each map indicates visibility fraction from low high.

To further understand occlusion distributions, we partitioned occlusion statistics into four types: horizontal, vertical, non-occlusion (or bare occlusion), and others, similar to [1, 5, 6]. Here we explain the rules of generating these four occluded types. The definition of specific occluded type depends on two directional overlapping ratios, which measure overlapping fractions of visible box over full-body box horizontally and vertically. For abbreviation, we denote the two overlapping ratio ranges as $\mathcal{R}_x$ and $\mathcal{R}_y$ respectively. Given a pedestrian, we confirm its occluded type according to the following empirical rules. The principles are illustrated in Fig.1, where (1) For horizontal type, $\mathcal{R}_x = (0.4, 1]$ and $\mathcal{R}_y = (0, 0.8]$, or $\mathcal{R}_x = (0.7, 1]$ and $\mathcal{R}_y = (0.8, 0.95]$, or $\mathcal{R}_x = (0.7, 0.95]$ and

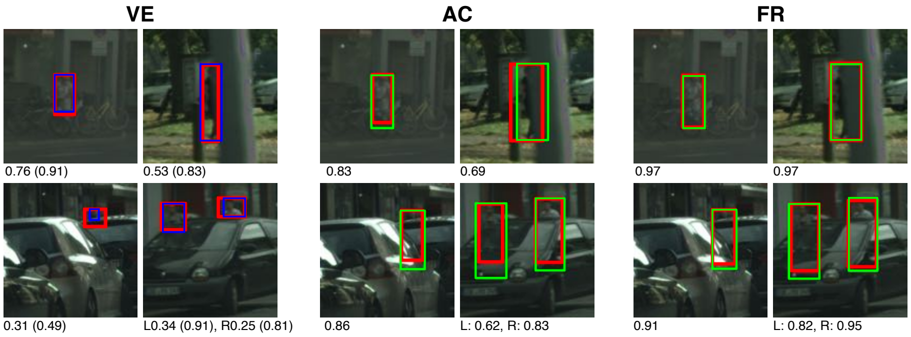

Fig. 3. Examples of PRNet's three-phase progression on varying occlusion types. Blue and green boxes indicate visible-part and full-body ground truth; Red boxes denote predictions in each phase. Numbers (with parentheses) indicate IoU between green (blue) boxes and predicted boxes.

$\mathcal{R}_y = (0.95, 1.0]$; (2) For vertical type, $\mathcal{R}_x = (0, 0.7]$ and $\mathcal{R}_y = (0.8, 1]$; (3) For non-occlusion, *i.e.* bare occlusion type, $\mathcal{R}_x = (0.95, 1]$ and $\mathcal{R}_y = (0.95, 1]$; (4) For others, $\mathcal{R}_x = (0, 0.4]$ and $\mathcal{R}_y = (0, 0.8]$. With the designs, we obtained the consistent observations with [1] that over 95% pedestrians can fit to 3 types including horizontal, vertical, and non-occlusion.

Fig. 2 shows more occlusions statistics on CityPersons and Caltech which follow the conclusions made in Sec. 3.1 of the main paper. Similarly, Fig. 2(a) shows head regions are highlighted and most visible regions are concentrated on upper parts of pedestrians. On the other hand, average ∼95% pedestrians in Fig. 2 (b)-(c) belong to the three types we exploited in **Anchor Calibration**.

2 Three components of PRNet and *full* PRNet

Fig. 3 illustrates examples of the three-phase progression. From left to right, the IoU against full-body ground truth box shows that PRNet gradually approaches the answer on varying occlusion types. These findings suggest that the progressive design involving VE, AC, and FR yields advanced results for occluded pedestrian detection. Across these illustrated examples and quantitative results in the main paper, we conclude the following observations:

Visible-part Estimator (VE) in PRNet can recover visible-part boxes in various occluded situations with high confidence. For example, our VE can successfully detect the clutter pedestrians which are occluded by grass as in second row of Fig. 3. Even for the low-resolution pedestrians, VE also achieved satisfactory detection.

Anchor Calibration (AC) in PRNet fits well on existing pedestrian detection datasets. As discussed in Sec. 1 and Sec. 3.1 of the main paper, over 95% pedestrians can fit to 3 types including horizontal, vertical, and non-occlusion in occlusion statistics. We can see that most examples shown in Fig. 3 belong

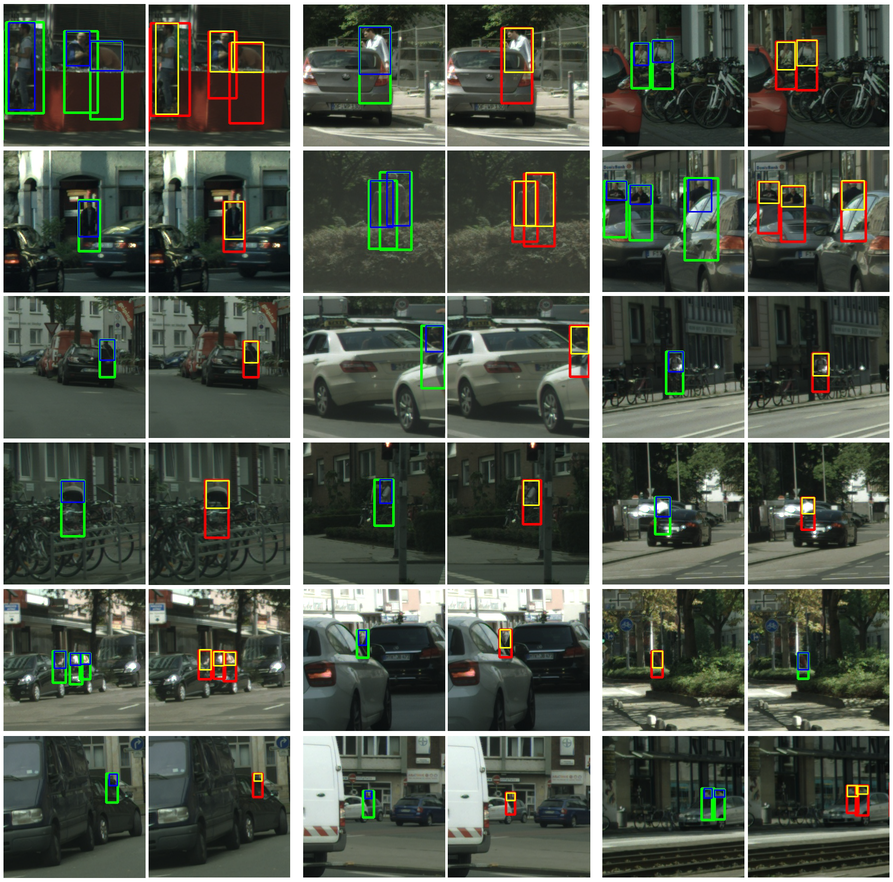

Fig. 4. Detection examples (from bare to heavy occlusions) by the proposed PRNet. Green box □ indicates full-body annotations and blue box □ corresponds to visible-part annotations. Red box □ and yellow box □ represent detected full-body and visible-part boxes respectively.

to the proposed 3 occlusions. In addition, the fixed-aspect ratio used in AC can help approach toward human annotations. One interesting case can be observed in Fig. 3 that only VE+AC could fit full-body regions to a large extent.

Fig. 4 illustrates detection examples of the *full* PRNet performed on CityPersons validation dataset at FPPI equals to 1. Observing detections, we see PRNet can successfully predict full-body predictions from bare to heavy occlusions involving gardens, cars, grass, bicycle, *etc.* Full PRNet can achieve promising results in the cases that even human are hard to detect, such as low-resolution and clutters shown in Fig. 4.

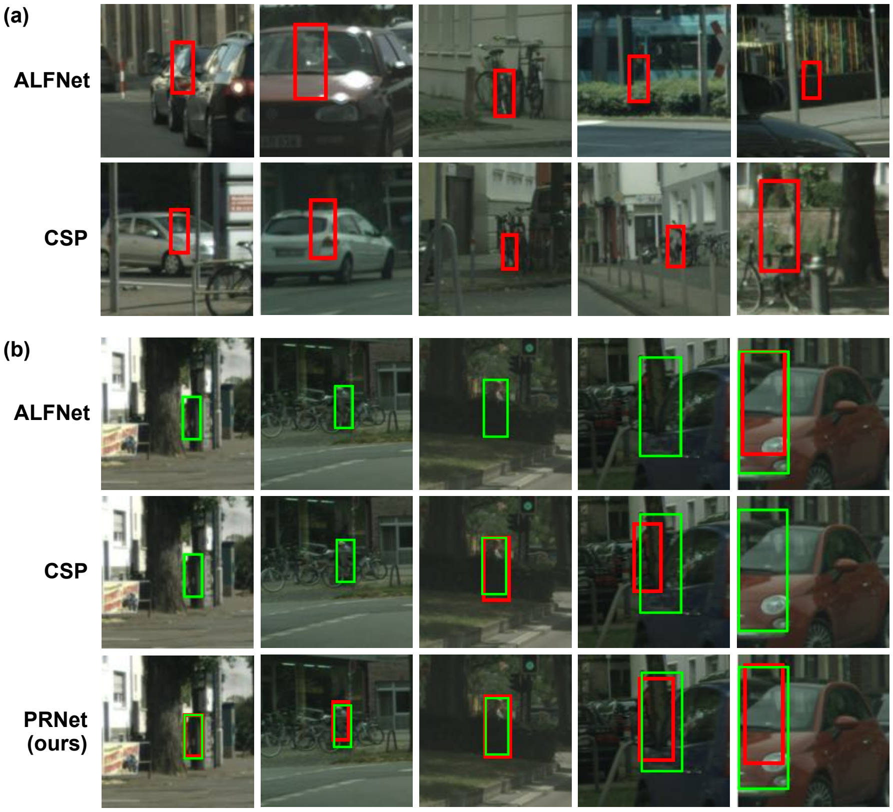

Fig. 5. Examples (at FPPI=1) from alternative methods performed on CityPersons validation set: (a) **false positives** and (b) **false negatives**. Green and red boxes denote ground truth and detection results, alternatively. In these examples, PRNet not only mitigates the false positives, but also correctly recovers the false negatives.

3 False detection by alternatives

Fig. 5 shows false negatives and false positives miss-detected by ALFNet [3] and CSP [4] where the proposed PRNet successfully predicted. Without considering occlusion supervision, ALFNet and CSP has limitations to learn discrimination between pedestrian and occlusions, such as trees, bicycle, cars and so on, thus false positives and false negatives are hard to reduce.

4 Cross-dataset generalization

PRNet provides adaptive anchor initialization thus generalizes well across diverse dataset. We train PRNet model on CityPersons dataset and then generalize to both ETH [2] and Caltech [1] dataset. Figs. 6 and 7 provide generalization results of ETH [2] and Caltech [1] dataset individually. Observing the detection, we can summarize several aspects as follows:

- PRNet model can successfully generalize to most occluded cases varying from bare (*e.g.*, Fig. 6, Fig. 7(a)) to heavy (*e.g.*, Fig. 7(b)-(d)).
- The generalized PRNet model can correct some annotations in original dataset. For instance, in Figs. 6(a), there is a pedestrian below streetlight but no annotations are provided in ETH dataset. Our PRNet correctly localizes it with visible-part and full-body predictions. The similar situations appear in crowded pedestrians in Figs. 6(b) and (d). In addition, we found some visible-part bounding boxes (bboxes) are mis-annotated in Caltech dataset. Observing one fully-visible pedestrian who stands near the traffic light shown in Fig. 7(b), visible-part annotation was only labelled around upper-body while our PRNet predicts visible-part box toward "real" visible full-body region. The similar case can be referred to the pedestrian showing in center of Fig. 7(c).
- The generalized PRNet model potentially provides weak annotations of visible parts in this community shown in Fig. 6, in which original visible-part bboxes are not given. We can see that predicted visible-part boxes are relatively reliable thus tightly fit the visible pedestrians in Fig. 6 and Fig. 7.

In all, PRNet not only outperforms the state-of-the-art methods illustrated in main paper, but also generalize better over diverse dataset across quantitative evaluations of the main paper and qualitative results illustrated in this section.

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.

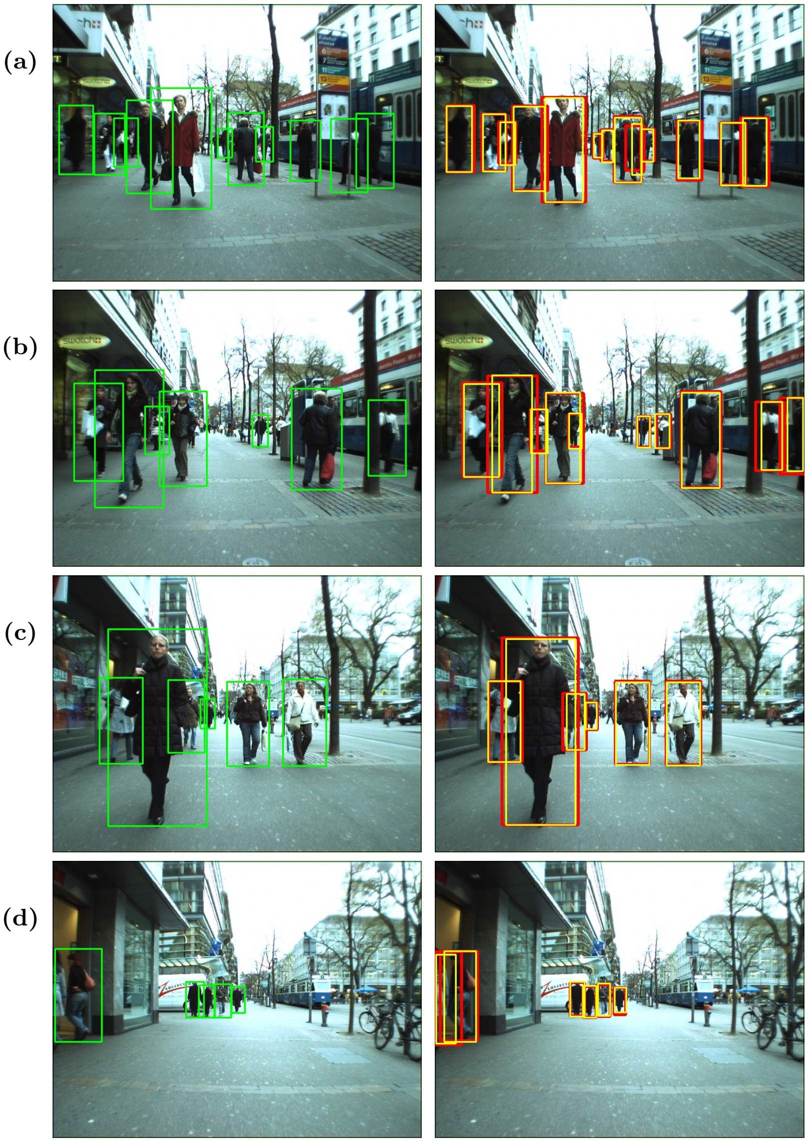

Fig. 6. Examples of generalizations on ETH dataset [2]. Green box 🟩 indicates annotated full-body box and visible-part box is not provided in ETH dataset. Red box 🟥 and yellow box 🟨 separately represent predicted full-body and visible-part boxes.

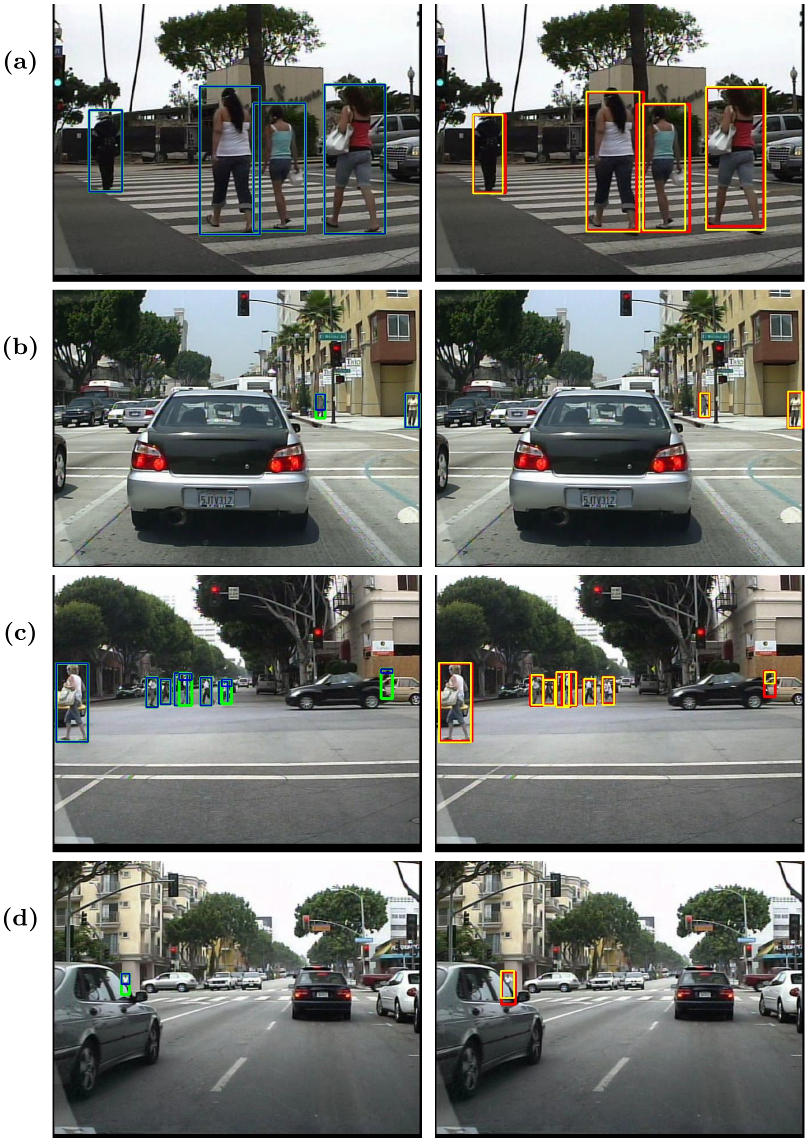

Fig. 7. Examples of generalizations on Caltech dataset [1]. Green box □ indicates full-body annotations and blue box □ corresponds to annotated visible-part box. Red box □ and yellow box □ represent predicted full-body and visible-part boxes respectively.

References

1. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. TPAMI **34** (2012)
2. Ess, A., Leibe, B., Van Gool, L.: Depth and appearance for mobile scene analysis. In: ICCV (2007)
3. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: ECCV (2018)
4. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: CVPR (2019)
5. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR (2016)
6. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR (2017)