Neurocomputing 482 (2022) 98–115

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

PRNet++: Learning towards generalized occluded pedestrian detection via progressive refinement network

Xiaolin Song, Binghui Chen, Pengyu Li, Biao Wang, Honggang Zhang *

Beijing University of Posts and Telecommunications, Beijing, China

ARTICLE INFO

Article history:
Received 16 September 2021
Revised 26 November 2021
Accepted 14 January 2022
Available online 21 January 2022
Communicated by Zidong Wang

Keywords:
Generalized Occluded Pedestrian Detection
Progressive Refinement Network
Dual-stream architecture
Unsupervised domain adaptation
Dynamic Iterative adaptation strategy
Multi-experts adaptation strategy

ABSTRACT

Pedestrian detection has achieved significant progress in recent years. Though promising results have been obtained on standard pedestrians, it remains challenging to detect pedestrians in various occlusion situations. In this paper, we propose *Progressive Refinement Network (PRNet)*, a novel single-stage detector for occluded pedestrian detection. Inspired by human's progressive process on annotating occluded pedestrians, PRNet perform sequential refinement in a single-stage detection framework by three phases: finding anchors of visible parts with high confidence, calibrating these anchors with a full body template derived from occlusion statistics, and adjusting the calibrated anchors to target full-body regions. Unlike conventional methods that utilize predefined anchors directly for full-body estimation, the proposed confidence-aware anchor calibration offers an adaptive anchor initialization for detection with occlusions, while helps reduce the gap between visible-part and full-body detection. We also propose an occlusion loss, which automatically up-weights heavily occluded pedestrian samples. In addition, a Receptive Field Backfeed (RFB) module is introduced to diversify receptive fields in early layers that commonly fire only on visible parts or small-size full-body regions. To further learn generalized representations of pedestrians in different occlusion states, we propose to establish a new single-stage detector with dual-stream architecture namely PRNet++. An Easy-branch and a Hard-branch are designed to learn complementary representation that are more robust to various occlusions. Moreover, the generalization ability of models for other domains are critical in real-world applications, but it has not attracted too much attention. To address this issue, we introduce the unsupervised domain adaptation setting. To handle occlusion situations in unknown domain better, we especially design a Dynamic Iterative adaptation strategy and a Multi-experts adaptation strategy for our occlusion-aware detectors PRNet and PRNet++. Extensive supervised within-dataset experiments and unsupervised domain adaptation experiments were performed on CityPersons, Caltech, ECP-night, KITTI and INRIA datasets, which can validate the effectiveness of our proposed methods in occluded pedestrian detection.

1. Introduction

Pedestrian detection is a fundamental computer vision task and serves as a crucial component of various real-world applications such as robotics[1–4], intelligent surveillance [5–7] and autonomous driving [1,8]. Though promising results were achieved in recent advances, occluded pedestrian detection remains a difficult problem [9–11]. The main challenges involve a wide variation in appearances of pedestrians due to occlusions caused by background objects (*e.g.*, cars, bicycles or trees) or other pedestrians, which occur frequently in real-world scenes and deteriorate the detection performance to different degrees. According to recent pedestrian detectors [12–16,10], the detection accuracy decreases significantly from "reasonable" occlusions to "heavy" occlusions (*i.e.*, from 10–16 points to 45–60 points in terms of miss rate).

Reviewing the literature, a common strategy for occlusion handling is utilizing annotations of visible parts as additional supervision cues to guide the full-body estimation, which is exploited by many existing pedestrian detectors. These approaches can be generally clustered into three groups: 1) A set of independent detectors trained for different occlusion patterns [19–26], 2) Attention maps to guide the learning of visible parts [27,28], and 3) Auxiliary visibility learning to refine final confidence scores of anchors [9,12,29]. Though these methods achieved some performance gains, there exist at least three limitations. First, training a group

* Corresponding author.
E-mail addresses: sxlshirley@bupt.edu.cn (X. Song), chenbinghui@bupt.cn (B. Chen), lipengyu007@gmail.com (P. Li), wangbiao225@foxmail.com (B. Wang), zhhg@bupt.edu.cn (H. Zhang).

https://doi.org/10.1016/j.neucom.2022.01.056
0925-2312/© 2022 Elsevier B.V. All rights reserved.

of detectors is computationally expensive, as each detector is trained for one specific occlusion pattern, which is difficult to enumerate in practice. Second, attention-based methods usually establish attention modules that are exhausted with proposals in architectures like Faster R-CNN [30], which result in a slow inference speed. Also, emphasizing visible parts solely could be sub-optimal for full-body detection. Finally, these full-body detectors are usually initialized with predefined anchors, which could be sub-optimal to learn robust representations of occluded pedestrians.

To tackle above challenges, we propose to establish a novel single-stage detector namely Progressive Refinement Network (PRNet) for occluded pedestrian detection. Fig. 1(a)(b) presents our main idea. Motivated by human's progressive labelling process of occluded pedestrians (*e.g.*, [17,18]), PRNet performs pedestrian detection in three phases, which gradually infers final full-body regions from visible parts. First, *visible-part estimation* generates high-confidence anchors of visible parts with a set of detection heads. To reduce the detection gap between visible parts and full-body regions, *anchor calibration* adjusts the visible-part anchors to a full-body template based on occlusion statistics, which is obtained from over 20,000 annotated bounding boxes of occluded pedestrians. For full-body detection, the calibrated anchors serve as a more adaptive initialization than predefined anchors (*e.g.*, [27,16,12,31,11]) since they have been refined twice to gradually approach full-body regions (via visible-part estimation and anchor calibration). Finally, we train a *full-body refiner* with the calibrated anchors and an extra group of detection heads. The novel progressive design is fitted into a single-stage detector by using one shared backbone network and separate detection heads for visible-part and full-body estimation, which involves limited extra complexity. Moreover, we introduce an occlusion loss to up-weight hard samples and a Receptive Field Backfeed (RFB) module that provides more diverse receptive fields to help shallow layers detect pedestrians in various sizes, which further improve training effectiveness.

To further improve the overall ability and learn more generalized models to handle various occlusion situations, we built a novel detector PRNet++ based on PRNet. Fig. 1(c) shows our idea. PRNet ++ is designed with dual-stream structure, where an Easy-branch for easy data (via weak augmentation) learning and a Hard-branch for harder data (via strong augmentation) learning. These two branches complement each other and the mixture of their results can raise a further performance gain. Especially, we propose a novel post-processing method called Support-areas Voting, which aims to further refine results after NMS. Considering the model architecture, PRNet++ just has several more detection heads than PRNet since the aforementioned two branches in PRNet++ share the same backbone network. Thus, PRNet++ still remain the single-stage design without introducing too much computation complexity.

Supervised experiments on CityPersons [17], Caltech [18] and ECP-night [32] have validated the feasibility of the proposed PRNet and PRNet++. Besides, we notice the limitation of supervised learning: the assumption that test data has the same distribution as train data. In reality, domain shift frequently occurs in many practical applications. For example, variances of pedestrian appearance, viewpoints, backgrounds, illumination, and weather condition can easily degenerate the performance of a pedestrian detector. When there exists a gap between the distributions of train data and test data, supervised learning methods perform poorly. One possible solution is collecting labeled data for a new domain, but it is usually expensive and time-consuming. Since the cross-domain generalization in occluded pedestrian detection is a critical issue,we introduce the unsupervised domain adaptation setting into occluded pedestrian detection task and conduct several domain adaptation explorations, which transfer knowledge from the train data domain (source domain) to the test data domain (target domain). Fig. 1(d) presents the ideas of our proposed unsupervised domain adaptation strategies. Based on PRNet, we propose a Dynamic Iterative adaptation pipeline with mean-teacher mechanism and adaptive thresholds for sample selection, which aims to generate more reliable pseudo labels and effectively reduce the annotation errors. On this basis, we further establish a Multi-experts adaptation pipeline, where different experts learn the transferable knowledge from different points of views and the mixture of their results can provide more precise pseudo labels. Extensive unsupervised domain adaptation experiments are conducted on CityPersons [17], Caltech [18], ECP-night [32], KITTI [1] and INRIA [33], which validate the effectiveness of our proposed adaptation strategies.

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) On the basis of PRNet, propose PRNet++ with dual-stream architecture to learn more robust representations to various occlusion situations while remaining the single-stage design.
(3) Introduce a novel task, unsupervised domain adaptive pedestrian detection. Propose a Dynamic Iterative adaptation strategy. Introduce Mean-teacher mechanism to the task and propose an adaptive threshold strategy for sample selection so as to generate pseudo annotations of high quality. Propose Multi-experts adaptation strategy, where different transferable knowledge from multiple experts guides the domain adaptation learning process.
(4) Extensive experiments including supervised experiments and unsupervised domain adaptation experiments validate the effectiveness of our proposed methods.

This work is an extension to our earlier publication [34]. Compared with [34], it has several improvements as below:

(1) We propose an advanced single-stage PRNet++ detector with dual-stream architecture to tackle data of different difficulty levels by different branches.
(2) We introduce the unsupervised domain adaptive pedestrian detection task to make some explorations on the cross-domain generalization capacity of our models.
(3) Under the unsupervised domain adaptation setting, we propose a Dynamic Iterative adaptation strategy and a Multi-experts adaptation strategy.
(4) We achieve significant performance improvements over our previous work [34].

2. Related Work

CNN-based Pedestrian Detection: Recently, pedestrian detection has achieved promising performance with the rapid progress of CNN-based general object detection [35]. Existing CNN-based pedestrian detectors can be roughly categorized into anchor-based and anchor-free detectors. For anchor-based approaches, two-stage detector (*e.g.*, Faster R-CNN [30]) and single-stage detector (*e.g.*, SSD [36]) are two typical designs. Most two-stage pedestrian detectors [37–40,28,27,16,12,29,15,41] generate coarse region proposals of pedestrian instances at first stage and then refine the proposals at second stage by leveraging some domain knowledge which learned in different ways (*e.g.*, extra learning task [38,28,29,27], hard mining [37], or cascaded labeling policy

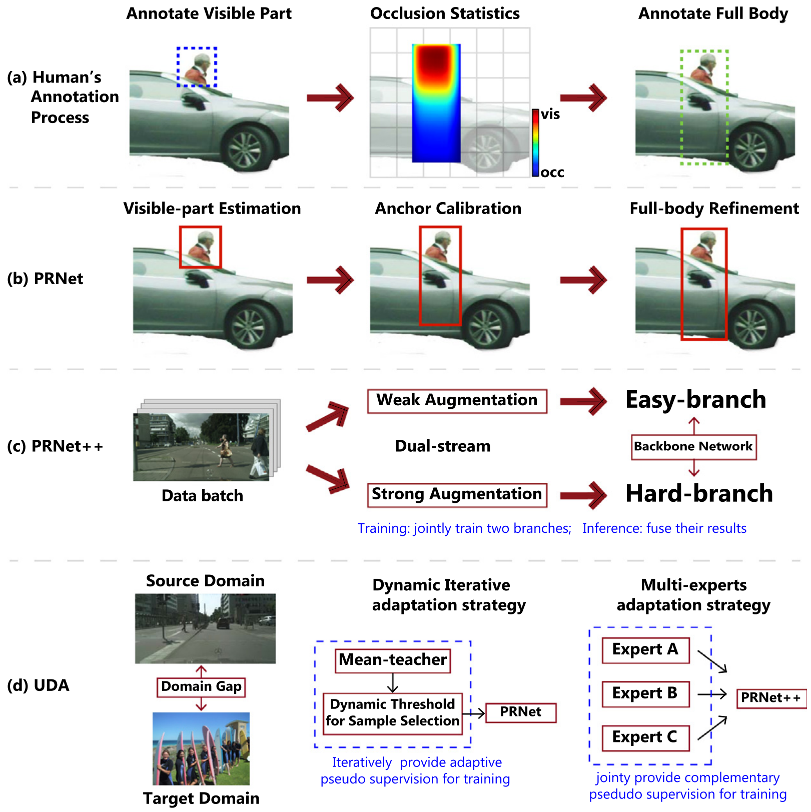


Fig. 1. (a) An overview of human's annotation process. (b) PRNet (proposed) imitates human's progressive annotation process on occluded pedestrians (*e.g.*, [17,18]), and gradually infers full-body regions from visible parts. (c) PRNet++ (proposed) is established with dual-stream structure, where Easy-branch and Hard-branch learn complementary knowledge from data generated by different augmentation modules. (d) Illustration of two proposed Unsupervised domain adaptation (UDA) strategies. Dynamic Iteration adaptation strategy guides the adaptation process by dynamic sample selection and mean teacher. Multi-experts adaptation strategy provide reliable pseudo annotations by fusing complementary knowledge from di.fferent experts.

[40]). For instance, RPN + BF [37] replaces the second stage refinement with a boosted forest and applies hard mining on proposals. However, the hybrid architecture could increase the training complexity. SDS-RCNN [38] jointly learns pedestrian detection and semantic segmentation so as to encourage the model pay more attention to pedestrian regions, where bounding-boxes serve as weak supervision for segmentation branch. Since two-stage detectors generate regions of interest (RoIs) at first stage and refine detections with cropped RoI features at second stage, the entire inference pipeline is slow and unable to meet the time efficiency requirements in real-world applications. In contrast, single-stage detectors [31,11,9] show faster inference speed owning to the one-shot design. Generally speaking, inference speed and detection accuracy are trade-offs between two-stage and single-stage detectors. To make a balance between accuracy and speed, ALFNet [11] gradually evolves anchor boxes with cascaded detection heads while keeping the single-stage design. The proposed PRNet and PRNet++ exceed these conventional methods by introducing occlusion–handling strategies while maintaining the high speed of single-stage detection framework.

As for anchor-free methods [10,14], instead of using original bounding box annotations, keypoints of pedestrians are introduced as new annotation forms. Thus, pedestrian detection problem is treated as keypoints detection problem, where anchors are no longer needed. TLL [10] predicted pedestrians' top and bottom vertexes and then grouped paired keypoints into individual pedestrian instances. And, CSP [14] detects pedestrians by predicting central points and scales. Though these CNN-based detectors have achieved promising advances, most detectors show limited detection performance on heavily occluded pedestrians.

Occluded Pedestrian Detection: Up to now, many works have been proposed to handle the occlusion issue in pedestrian detection. The Part-based method [19–25,42,26,43] is a typical solution, which learns series of specific detectors for various occlusion patterns and fuse their predictions during inference. Nevertheless, it is impractical to exhaustively enumerate occlusion patterns due to the complex network design and expensive computation. Instead of learning each occlusion pattern, [16,9] divide a proposal or bounding box into a fixed number of blocks and predict the visible probability of each block. Though computation complexity is optimized, these approaches are lack of flexibility since they rely on manually designed partitions. In recent years, some works aim to learn robust pedestrian representations and anchor scoring strategy. On one hand, [28,27] introduce the attention mechanism to learn robust feature representations under the guidance of attention maps extracted from proposals mappings to occlusion-aware semantics. Pang et al. [27] and Zhang et al. [28] learn pixel-wise and channel-wise attention maps respectively so as to highlight visible body parts and suppress occluded parts. Whereas, overemphasizing visible body parts might result in sub-optimal learning for full-body estimation. On the other hand, some methods [29,12] learn extra confidence scores for proposals or anchors by introducing extra learning task. Bi-box [29] establishes two parallel branches for visible-part and full-body detection respectively, and fuses confidence scores from two branches during inference. [12] constructs a separate discriminative classification branch to learn more reliable confidence scores for proposals, where heavily occluded proposals are enforced to be close to non-occluded or slightly occluded ones. Besides, some other works [44–46,15,47–51] focus on pedestrian detection in crowded scenes. For instance, RepLoss [15] proposes a novel regression loss, which hinders proposals from shifting to surrounding targets with additional penalty terms. In addition, a few works begin to tackle occlusion in other perspectives. *W^3Net* [52] introduces bird-eye view maps and [53] resorts to temporal information. AEVB [54] reformulates pedestrian detection as a variational inference problem and optimizes it by a Auto-Encoding Variational Bayes algorithm. These novel attempts achieve potential performance, but the generalization ability needs to be further studied. Unlike most aforementioned methods that are initialized with pre-defined anchors, the proposed PRNet takes confidence-aware and adaptive anchor initialization to learn occluded pedestrian detection, which help improve detection performance. Moreover, the advanced version PRNet++ gives a new perspective of constructing two branches to tackle data in different occlusion levels and get two groups of complementary detections.

Unsupervised Domain Adaptation: Unsupervised domain adaptation (UDA) [55] aims to tackle domain-shift between source domain with labeled images and target domain with unlabeled images. It has attracted much attention because of its capacity of saving the cost of manual annotations. Many UDA works has been proposed in image classification [56–64] and segmentation [65–67]. In comparison, unsupervised domain adaptive object detection has received less attention. Most works [68–74] focus on the cross-domain alignment between the source domain and target domain. DA-Faster [68] reduces the domain gap in both image level and instance level by adversarial learning. [74] employs style transfer methods to generate multiple domains and uses a multi-domain discriminator to adapt all domains simultaneously. Considering the distance divergence between different domains, [69] introduces an intermediate domain to tackle the difficulty in mapping two domains with big distribution gap. However, these methods may overlook some potential important local regions, which may contain objects of interests and be crucial to object detection. To address this issue, Region-level alignment[71] and CR-DA-DET [72] propose to mine discriminative local regions for alignment. SW-Faster [70] focus on aligning local receptive fields at low-level features along with weak alignment on global regions. However, it is hard to apply these methods to our single-stage detection framework since they utilize Faster R-CNN based two-stage design. As for single-stage object detection, [75] introduces the self-training strategy and adversarial background score regularization.

In this paper, we give insight into unsupervised domain adaptive pedestrian detection. As a sub-problem of object detection, pedestrian detection under unsupervised domain adaptation setting was rarely discussed. [76] also focuses on cross-domain problem, but their work is under the directly cross-dataset evaluation setting without any adaptation. [77] is the closest work, which randomly samples negative instances from source domain and positive instances with high confidence from target domain for training. However, as domain shifts, background clutters are often different from ones in source domain. In other words, negative instances in target domain may be never seen by a model in source domain. Since [77] ignores negative instances training in target domain, false positive errors may break out. Consequently, false positive annotations mislead the optimization process and the resulted model generates more false positives. Differently, in this paper, we make a lot of efforts to alleviate bad effects of false pseudo annotations.

3. Methods

In this section, we first describe the details of PRNet (see Section 3.1), as well as the advanced detector PRNet++ (see Section 3.2). Then, we give the problem definition of the unsupervised domain adaptive pedestrian detection task and introduce the proposed adaptation pipelines (see Section 3.3).

3.1. PRNet Architecture

Motivated by human's progressive process on annotating occluded pedestrians (*e.g.*, CityPersons [17] and Caltech [18]), we construct PRNet to gradually migrate highly confident detection on visible parts toward more challenging full-body detection. For this purpose, we propose to construct a single-stage detector with three training phases: Visible-part Estimation (VE), Anchor Calibration (AC), and Full-body Refinement (FR). Unlike most methods that learn full-body only [11,15] or independently with visible parts [29], we integrate both of them into a single-stage detection framework. To bridge the gap between visible-parts and full-body detection, we introduce AC to align anchors from VE to FR.

Fig. 2 presents the PRNet architecture. The top row depicts the backbone network, where we truncated first 5 stages of ResNet-50 [78] with modification of appending an 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 acquire detection from the outputs of last four. Detection heads for VE and FR are attached on top of each detection layer. The network is trained following three phases: VE, AC, and 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.

Specifically, denote x as an input image, $\Phi(x)$ as feature maps from backbone network, $\mathscr{A}_0$ as a set of predefined anchors (as in SSD [36]), $\mathscr{B}^*$ 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:

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

where $E_v(\Phi(x), \mathscr{A}_0)$ is the 1st-phase visible-part estimation (VE) whose outputs are a set of visible-part anchors and confidence scores $\{\mathscr{A}_1, s_1\}$, $C(\cdot)$ is the 2nd-phase anchor calibration (AC) that aligns visible-part anchors $\mathscr{A}_1$ to full-body anchors $\mathscr{A}_2$, and $E_f(\Phi(x), \mathscr{A}_2)$ is the 3rd-phase full-body refiner (FR) that outputs the final full-body anchors to compute $\mathscr{B}^*$ and their scores s^* using inference F. 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 $\mathscr{A}_1$ as a regression task (from predefined anchors $\mathscr{A}_0$), and anchor scores as a classification task. Fig. 3 (a) depicts the detection head, whose loss can be written as:

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

where $\mathscr{L}_{focal}$ is focal loss [79] for classification, $\mathscr{L}_{smoothL1}$ is a smooth-L1 loss for regression (as adopted in Faster R-CNN [30]), $[y = 1]$ is an indicator for positive samples, and λ_v is a tuning parameter. As VE is trained on visible parts, its prediction (*i.e.*, $\mathscr{A}_1$) on visible parts is generally more confident and accurate than detectors trained with occlusions.

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

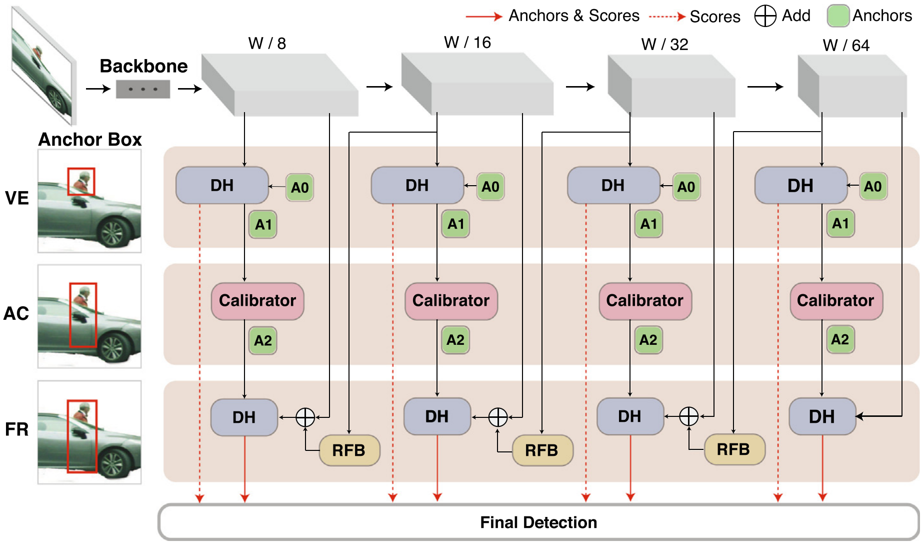


Fig. 2. Architecture of PRNet. From top to bottom, PRNet uses a detection backbone illustrated with four blocks of features maps. The network is trained in three phases: **Visible-part Estimation** (VE), **Anchor Calibration** (AC), and **Full-body Refinement** (FR). VE and FR take visible-part and full-body 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 Section 3.1.

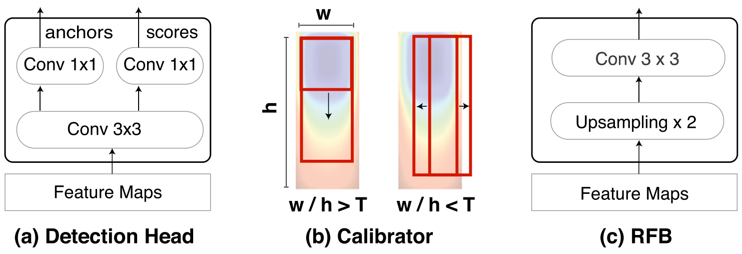


Fig. 3. Modules used in PRNet architecture (as in Fig. 2): DH, Calibrator and RFB.

1. The aspect ratio of visible-part boxes is much more diverse than that of full-body boxes [17,18], 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.*, [80]), compared to most methods that use predefined anchors (*e.g.*, [17,16,11,12]).
3. The IoU discrepancy between visible-part anchors and full-body ground truth boxes is large; proper calibration can significantly improve IoU.

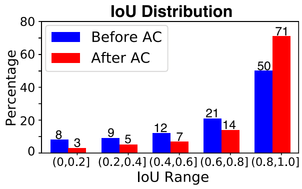


Fig. 4. IoU distribution before and after anchor calibration on the CityPersons dataset. IoU is measured between anchors and full-body ground truth.

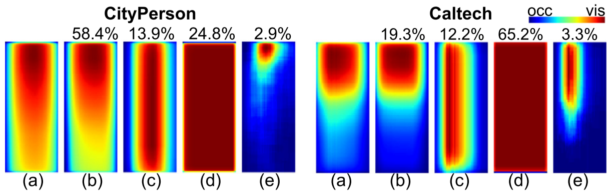


Fig. 5. Occlusion statistics from CityPersons [17] (left) and Caltech [18] (right): (a) Occlusion statistics with blue indicating occlusion; red indicating visible parts, (b) Horizontal occlusions, (c) Vertical occlusions, (d) Non-occlusion, (e) Others. Percentage (%) denotes the likelihood of each oc.clusion pattern.

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 [17], where visible-part boxes before AC were taken from the annotations in the original dataset. As can be seen, calibration significantly shifts the distribution toward higher IoU, *e.g.*, +21% for IoU in $(0.8, 1.0]$, which can help approximate final full-body regions. In addition, AC addresses discrepancy during anchor assignment between VE and FR, *i.e.*, without AC, a positive $\mathscr{A}_1$ could be assigned as a negative anchor for FR, making VE and FR fail to complement each other.

To achieve AC, we first study occlusion statistics on two popular datasets CityPersons [17] and Caltech [18] using their standardized 0.41 box aspect ratio. Fig. 5 illustrates occlusion distribution over a full-body box and four occlusion types (*i.e.*, horizontal, vertical, non-occlusion, and others, similar to [28]) with respective likelihood in each dataset. As can be seen in Fig. 5(a), over the two datasets, the occlusion statistics show the upper bbox, which usually includes heads, are consistently visible, with most occlusion appearing in the lower portion (*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 $\mathscr{A}_1$, as depicted in Fig. 3(b). For the anchors with ratio >0.41, we vertically stretch them *downwards* until reach 0.41 aspect ratio, due to heads being frequently visible, as shown in Fig. 5b) and and [18]. Anchors with ratio <0.41 are horizontally extended to 0.41 w.r.t. the center of $\mathscr{A}_1$, 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 [17], 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 $\mathscr{A}_2$ 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 network, but exploits separate detection heads. Different from VE that sees only visible parts, FR starts to see hard positive samples whose anchor boxes are still far from ground-truth full-body region. The $\mathscr{L}_{smoothL1}$ in Eq. (2) treats every positive sample equally, which could be less effective to deal with hard samples. To encourage learning on hard positive samples, we reformulate the regression loss $\mathscr{L}_{smoothL1}$ with an occlusion weight, which is defined as a reverse IoU between $\mathscr{A}_2$ and ground truth full-body boxes $\mathscr{B}_{gt}$. Given $a \in \mathscr{A}_2$ and its corresponding $b \in \mathscr{B}_{gt}$, the weighted loss, termed as *occlusion loss*, can be rewritten as:

$$\mathscr{L}_{occ} = \sum_{a \in \mathscr{A}_2} (1 - \mathrm{IoU}(a, b))\{[|s| < 1]0.5s^2 + [|s| >= 1](|s| - 0.5)\}, \qquad (3)$$

where s is the difference between predicted offsets and ground truth offsets (see [30] for details). The less overlap between the calibrated anchors $\mathscr{A}_2$ and $\mathscr{B}_{gt}$, the higher $\mathscr{L}_{occ}$ is. As a result, the loss for FR becomes:

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

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 pixel-wise manner.

Fig. 6 shows the saliency maps [81] 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.

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

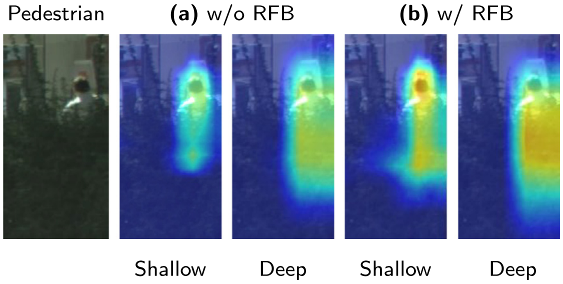


Fig. 6. Saliency maps highlighted by the third FR phase: (a) w/o RFB and (b) w/ RFB. "Shallow" indicates the 2nd layer, and "Deep" indicates the 3rd layer.

An anchor is labeled as positive if intersection-over-union (IoU) between an anchor box and a ground-truth box is above a threshold θ_p, as negative if the IoU is lower than θ_n, and otherwise will be ignored during training.

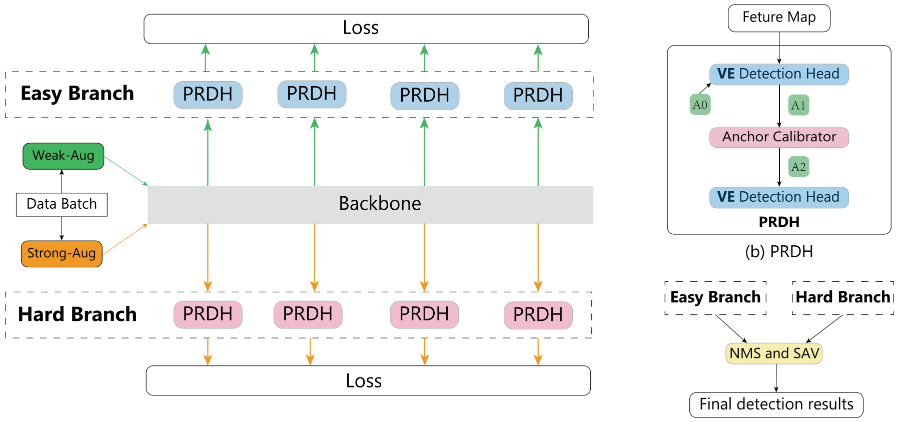
Inference: During inference, we obtain predicted anchor boxes from FR, and associate anchor scores by multiplying the scores from VE and FR. The score fusion provides complementary guidance so as to improve detection robustness (similar to [29]). We obtain the final bounding boxes $\mathscr{B}^*$ by first filtering out candidate anchor boxes with scores lower than 0.05 and then merging them with NMS.

3.2. PRNet++ Architecture

We observe that it is hard for a pedestrian detector to obtain consistent good results on test subsets in different occlusion levels. In other words, when a model can precisely detect occluded instances, it may miss some non-occluded ones due to the appearance discrepancy. On the contrary, if a model does very well in detecting non-occluded or slightly occluded instances, it may generate low confidence scores for heavily occluded ones. To further improve the ability of handling various occlusion situations, we propose a more advanced network PRNet++, which is built upon PRNet and the entire pipeline is illustrated in Fig. 7(a). For clarity, We integrate detection heads of VE and FR as well as AC into a module named PRNet detection head (PRDH) as shown in Fig. 7(b).

Inspired by Divide-and-conquer algorithm, we propose a dual-stream strategy, which is the core idea of PRNet++. The general thought is to establish an Easy-branch and a Hard-branch which focus on easy situations and hard situations respectively. By training these two branches jointly, we can enforce two branches to focus on different parts of feature maps from the backbone. The representations learned by the easy branch are more precise for easy instances but may miss some hard instances due to the appearance discrepancy. The hard branch performs better on hard subsets but may be not robust enough to detect some easy instances. Two branches generate complementary detection results and the mixture of them can be more robust to various occlusion situations.

Owning to the flexible architecture of PRNet, we can easily build these two branches while remaining the single-stage design. The two branches have the same architecture with four PRDHs and share one backbone. As shown in Fig. 7(a), a green data stream and an orange data stream go through Easy-branch and Hard-branch respectively. How to obtain these two stream is a crucial. In this paper, we generate two streams by controlling the proportion of hard examples in the data batch. Since occlusion occurs frequently in pedestrian detection task, the occlusion level of one pedestrian instance can be viewed as its degree of difficulty. To generate more hard examples and increase the difficulty of training



(b) PRDH

(a) PRNet++ architecture

(c) PRNet++ in Inference phase

Fig. 7. (a) Architecture of PRNet++. PRNet++ establish two branches (*i.e.*, Easy-branch and Hard-branch) to tackle training data in different occlusion levels. The green stream denotes easy data stream via weak augmentation. The orange stream denotes hard data stream via strong augmentation. (b) Architecture of PRNet Detection Head. The detection pipeline of PRNet via VE, AC and FR in a single detection layer is integrated into a PRNet detection heads. Given initial anchors (A0), a VE detection head learn to predict visible-part anchors (A1), which are calibrated by AC to generate calibrated anchors (A2). A FR detection head takes A2 as input and generates final full body bounding boxes. (c) In inference phase, detection results from Easy-branch and Hard-branch are processed by NMS and SAV (Supporting-areas voting) to obtain final detection results.

data, we adapt Random-erasing into our data augmentation module. Random-erasing is widely used in other human related tasks like person re-identification, but hasn't been exploited in pedestrian detection yet. Given a pedestrian sample, Random-erasing randomly selects a rectangle region in the visible parts and erases its pixels with random values. With Random-erasing, we can convert an easy pedestrian instance into a hard pedestrian one. Here, we introduce the concept of weak-strong augmentation. In this paper, We take augmentation methods in [14,11] as the weak augmentation including random flipping, random color distortion, random paving and random cropping methods. As for the strong augmentation, other than the weak augmentation methods, we randomly select some easy examples to apply Random-erasing. As shown in Fig. 7(a), the batch data via the weak augmentation is the input of Easy-branch while the one via the strong augmentation is the input of Hard-branch.

In training phase, the two branches are trained jointly with the same loss function. In inference phase, we fuse results from the two branches by NMS firstly. To fully utilize the complementarity of the two branch, we propose a Support-areas voting algorithm, which is illustrated in Algorithm 1 and apply it after NMS to obtain the final detection results.

Algorithm 1:Supporting-areas voting.

```
Input: Bounding Box set generated by $Model_1$ and $Model_2$
  (fused by NMS): $B = \{b_1, b_2, \ldots, b_N\}$; Box set generated by
  $Model_1$ before NMS: $B^1$; Box set from $Model_2$ before NMS:
  $B^2$; Supporting-areas selection IOU threshold $N_t$; Voting
  threshold $N_v$.
$B^* \leftarrow \{\}$
$count \leftarrow 0$
for$b$ in $B$ do
  for$b^1$ in $B^1$ do
    if $iou(b, b^1) \geqslant N_t$ then
      $count \leftarrow count + 1$
    end if
  end for
  for$b^2$ in $B^2$do
    if $iou(b, b^2) \geqslant N_t$ then
      $count \leftarrow count + 1$
    end if
  end for
  if$count \geqslant N_v$then
    $B^* \leftarrow B^* \bigcup b$
  end if
end for
Output: $B^*$
```

3.3. Domain Adaptation for Pedestrian Detection

Preliminary: Unsupervised domain adaptation aims to transfer the model trained from source domain to target domain. The most challenging problem lies in no labels are available in target domain. We assume that source data (x^s, y^s) is drawn from source domain X_s and unlabeled target data (x^t) is drawn from the target domain X_t, where x represents an image and y represents corresponding ground-truth labels. We denote the distribution of domain X as $P(X)$ and $P(X_s) \neq P(X_t)$.

Pseudo Labeling Since there is no access to manual annotations in target domain, we obtain supervision cues by applying a source domain pre-trained model M_s to x_t, which can be formulated as follows:

$$\{(b,s)\} = M_s(x_t), \tag{5}$$

where $\{(b,s)\}$ is a detected bounding boxes set which has been processed by Non-Maximum Suppression (NMS) and b denotes a detected bounding box attached with a confidence score s. We take the only supervision cues $\{(b,s)\}$ to generate pseudo labels y^t for x^t from target domain, which can be formulated as:

$$y^t = \begin{cases} positive & \mathrm{s} > \mathrm{h_p} \\ ignore & \mathrm{s} \leqslant \theta_\mathrm{p} \quad \text{and} \quad \mathrm{s} \geqslant \mathrm{h_n} \\ negative & \mathrm{s} < \mathrm{h_n} \end{cases} \tag{6}$$

According to the value of y^t, we can obtain $\{b\}_p, \{b\}_{ig}$ and $\{b\}_n$, which denote the positive samples set, ignore areas set and negative samples set respectively divided from $\{(b,s)\}$ by the sample

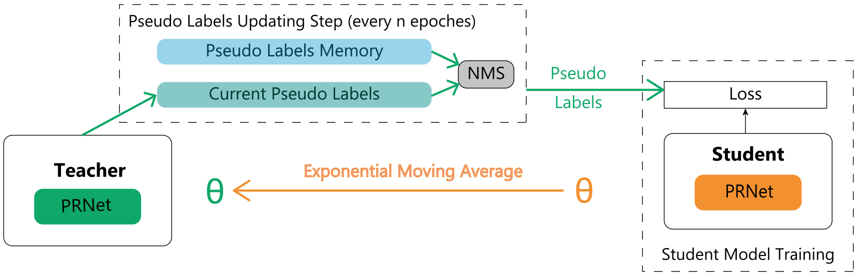


Fig. 8. Pipeline of Dynamic Iterative adaptation strategy. Teacher model and student model have the same model architecture. Teacher model is an ensemble of student models in different iterations via Eq. 7. In Pseudo label updating step, the last pseudo labels saved in Pseudo Labels Memory and the current pseudo labels generated by the up-to-date teacher model are fused by NMS to provide the supervision for the training of student model. θ represents model parameters.

selection threshold h_p and h_n. Thus, we can exploit (x^t, y^t) to fine-tune the source domain pre-trained model M^s and finally get a model M^t for target domain. The objective loss is the same as Eq. 2 and Eq. 4.

Below, we introduce the Mean-teacher mechanism, which are exploited in all of our domain adaptation strategies.

Mean-teacher mechanism Mean-teacher mechanism [82] has been widely studied in semi-supervised learning and unsupervised learning, where the core idea is to generate consistent training supervision for unlabeled data. To improve the stability of the estimated pseudo labels during training process and facilitate optimization of the framework, we introduce Mean-teacher mechanism to our pipeline, where teacher parameters are the exponential moving average (EMA) of student parameters at different training iterations, which can be formulated as:

$$\theta_{tea} = \alpha\theta_{tea}^{t-1} + (1-\alpha)\theta_{stu}^{t}, \tag{7}$$

where α is a smoothing coefficient hyperparameter and t represents the $t-th$ epoch, θ_{tea} and θ_{stu} denotes parameters of teacher model and student model respectively. The slowly progressing teacher model can be regarded as an ensemble of student models across different training epochs. The quality of pseudo labels produced by teacher models is much higher than those produced by student models since teacher model aggregates complementary information through temporal ensemble. In all of our experiments, we exploit up-to-date teacher models to generate pseudo labels in every label updating steps.

Dynamic Iterative adaptation strategy: To further improve the adaptation performance, we propose a Dynamic iterative adaptation strategy. The entire pipeline is illustrated in Algorithm 2 and Fig. 8.

Algorithm 2: Dynamic Iterative adaptation strategy.

```
Input: Source domain data $(x^s, y^s)$; Source domain unlabeled
    image data $x^t$; Source domain pre-trained model $M_s$;
    Pseudo labels updating interval: $N^p$; Initial sample selection
    confidence threshold $h_p$ (for positive samples) and $h_n$ (for
    negative samples).
1: Train the source model $M_s$ on the source stream with
    abundant manual annotated data;
2: Use $M_s$ to initialize a model on target steam as $M_0$ and
    generate pseudo labels $y^t$;
3: for $i = 1 : N$ do
4:     if $i \geqslant 0$ and $i \bmod N^p = 0$ then
5:     Begin the $j_{th}$ pseudo labels updating step, $j = i/N^p$.
6:         Generate up-to-data pseudo labels $y_j = \{(b^j, s^j)\}$
    with $M_i$;
```

* (*continued*)

Algorithm 2: Dynamic Iterative adaptation strategy.

```
7:         Get the history pseudo labels $y_{j-1} = \{(b^{j-1}, s^{j-1})\}$
    from the pseudo label memory;
8:         $\{(b^t, s^t)\}$=NMS(CAT($y_j$,$y_{j-1}$));
9:         Update $h_p$ and $h_n$: $h_p = h_p + 0.01 \times i$,
    $h_n = h_n + 0.01 \times i$;
10: Take updated $h_p$, $h_n$ and $\{(b^t, s^t)\}$ to generate pseudo
    labels $y_j^t$ by Eq. 6
11:         Update the pseudo label memory with $y_j$;
12:         Randomly select training images with manually
    annotations $(x_j^s, y_j^s)$ from source domain;
13:         $X = \{x_j^s, x^t\}, Y = \{y_j^s, y_j^t\}$;
14:     end if
15:     Take $(X, Y)$ as training data to update $M_i$ into $M_{i+1}$
    with Eq. 2 and 4
16: end for
Output: final model $M_N$
```

Firstly, we train a model M_s on source domain data with abundant manual annotations. Then, M_s is applied to unlabeled training data in target domain and detection results will be stored in memory. We can obtain initial pseudo labels with these detection results by Eq. 6 and train a model in target domain. A pseudo labels updating step is set every $n(n > 1)$ epoch. At this step, we firstly apply the up-to-date model on training data in target domain. To smooth the detection results and leverage the discrepancies among the outputs from models in different epochs, the up-to-date detection results and the last ones stored in memory are fused by NMS and the fused results are used to generate pseudo labels. Then, we update the memory with up-to-date detection results. As the pseudo labels contain agnostic noises, the performances may be affected by the wrongly annotated samples including the false negative instances and false positive ones. How to select reliable positive samples is a challenging problem. A common solution is filtering out samples with a high confidence threshold, but some hard positive samples, especially those in heavy occlusion level, may become false negative errors and cannot be learned by the model. And, a lower threshold will introduce more false positive bounding boxes and degrade the performance. What's more, as a hyper-parameter, an invariable threshold will make the model sensitive to its value. To handle this problem, we exploit dynamic thresholds as shown in the 9_{th} line of Algorithm 2, where h_p and h_n are increased over time. In other words, we progressively increase the selection criterion of positive samples. Thus, models trained by samples with different false positive noise are ensembled by teacher model and the generated up-to-data pseudo labels

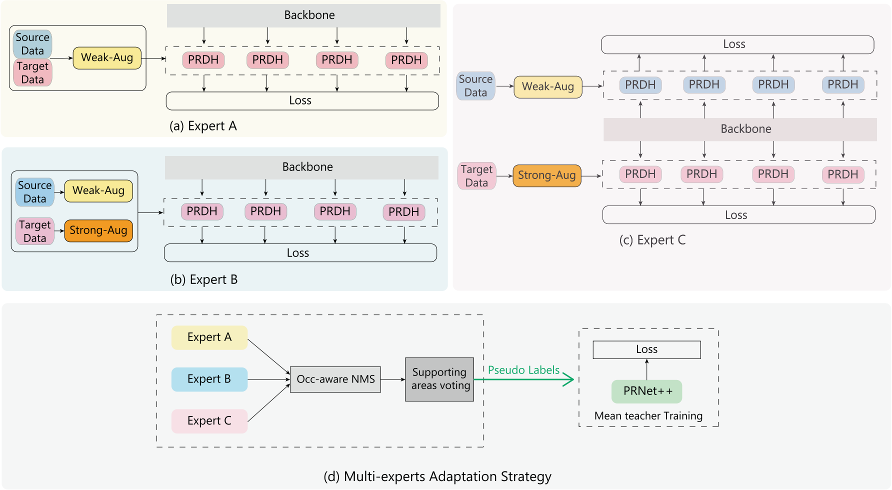


Fig. 9. (a) Architecture of Expert A. (b) Architecture of Expert B. (c) Architechture of Expert C. (d) Pipeline of Multi-experts adaptation strategy. Experts of different architectures with different data augmentations provide complementary supervision for the down-stream training of PRNet++.

and history ones may also have different noise cases. The process of model ensemble and pseudo labels fusion can counteract the bad effects of different kinds of false positive and false negative errors.

Multi-experts adaptation strategy: The Dynamic Iterative adaptation strategy establishes a basic pipeline for unsupervised domain adaptive pedestrian detection. Although dynamic thresholds and label fusing strategy can alleviate some false pseudo annotations, those remained ones may be further strengthened through incorrect supervision. It is difficult for the model itself to rectify pseudo labeling errors. To tackle this problem and further improve the adaptation performance, we propose a Multi-experts adaptation strategy, where different experts learn to handle different occlusion situations and provide complementary pseudo labels. The entire pipeline is presented in Fig. 9.

How to make the model learn more useful information from pseudo labels generated by the model itself is a crucial problem under the unsupervised domain adaptation setting. Weak-strong data augmentation scheme is an effective solution, which has been validated in semi-supervised image classification and semi-supervised object detection, where the model is enforced to make consistent predictions on the weakly augmented data and the strongly augmented data. We introduce the weak-strong data augmentation scheme into our unsupervised domain adaptive pedestrian detection task. In Section 3.2, we have exploited it to generate easy and hard data streams. Here, we use the same weak and strong data augmentation methods. As we all know, there is a big detection accuracy gap between occluded pedestrian samples and non-occluded ones. When domain shifts, it is very difficult for the model to detect all the occluded pedestrians. Hence, some occluded samples may be missed during pseudo labeling process, which leads to the lack of occluded training samples and a poor adaptation performance on the heavily occlusion subset finally. Strong augmentation with Random-erasing can convert some slightly-occluded samples into heavily-occluded ones. In this way, the proportion of data in different occlusion level can be changed and the sample inbalance problem can be alleviated. Besides, Random-erasing can also reduce the risk of over-fitting and improve the model's robustness to label noise.

Based on the weak-strong augmentation scheme, we introduce three expert architectures as shown in Fig. 9. All of them are variants of PRNet. Expert A (see Fig. 9 (a)) utilizes weak augmentation for both source data and target data, which focus on samples in relatively slight occlusion level. Expert B (see Fig. 9) (b)) and Expert C (see Fig. 9 (c)) apply strong augmentation on target data, which aims to tackle heavier occlusion situations since the percentage of heavily-occluded samples is increased. The difference of Expert B and Expert C is the network architecture. The source data stream and target data stream are processed by different groups of detection heads in Expert C, while they share the same detection heads in Expert B. These experts of different architectures with different data augmentations can learn training samples from different views and handle different situations during inference. We use the advanced PRNet++ as the downstream detector. The mixture of multiple experts by NMS and Support-areas voting can provide a more reliable supervision for the downstream training.

4. Experimental Settings

Datasets: We conducted experiments on five public datasets: CityPersons [17], Caltech [18], ECP-night [32], KITTI [1] and Inria [33]. **CityPersons** [17] is a challenging dataset, which was recorded in many cities and countries across Europe and includes various occlusion cases. We use the standard training and validation split, which consists of 2,975 and 500 images respectively. To ensure the results are directly comparable with the literature, we report results on 6 subsets with various occlusion degrees: **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. [28], and **Bare** with [0.9, 1.0], **Partial** with [0.65, 0.9], and **Heavy** with [0, 0.65] from [15]. Note that we only consider pedestrians with height larger than 50 pixels. **Caltech** [18] is a widely used pedestrian dataset that contains 10 h 640 × 480 30 Hz video taken from an urban driving environment. We use the Caltech10x (4250 images) for training and the original test set (4024 images) for evaluation. Results on **R**, **HO**, **Heavy** and **R + HO** are reported, where the subset definition is the same as CityPersons. **ECP-night** is the night-time subset of EuroCity Persons (ECP) dataset [32], which was collected in 31 cities across 12 countries in Europe. The training and validation sets contain 4222 and 770 images respectively. We report results on 3 subsets: reasonable set (persons with height larger than 40 pixels which are occluded less than 40%), occluded set (persons with height lager than 40 pixels which are occluded between 40% and 80%) and all set (persons with height lager than 20 pixels which are occluded less than 80%). **KITTI**[1] is a popular urban object detection dataset. Following [83], the 7,481 training images is split into a training set (3,712 images) and a validation set (3,769 images). Evaluation is done for the pedestrian class in three regimes: Easy, Moderate and Hard, which contain objects of different occlusion and truncation levels. **INRIA**[33] contains images of high resolution pedestrians collected mostly from holiday photos taken by mobile platforms, which consists of 2,120 images, including 1,832 images for training and 288 images for testing. Specifically, there are 614 positive images and 1,218 negative images in the training set. Following [16], we use the 614 positive images for training and the 288 testing images for evaluation. We reports results on the **All** set, which contains all testing images.

Metrics: For CityPersons, Caltech and INRIA, evaluation was reported on the standard MR^{-2} (%) [18], which computes the log-average miss rate at 9 False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^{0}]$. The lower MR^{-2} is the better. For KITTI, we report results on Average Pecision (AP), which is the metric in the official evaluation server of KITTI[1]. The higher AP is the better. For CityPersons, ECP-night and KITTI, we evaluate the performance of our methods on validation sets since annotations of test sets are unavailable and submissions frequency to each official evaluation server is limited.

Implementation details: We implement our methods based on the Keras [84] with 2 V100 GPUs. 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 IOU threshold for NMS is set to 0.5 in all experiments. The α in Eq. 7 is set to 0.999. In Algorithm 1, we set Supporting-areas selection IOU threshold $N_t = 0.5$ and voting threshold $N_v = 25$. In Algorithm 2, we set pseudo labels updating interval $N^p = 5$ and the initial sample selection confidence thresholds $h_p = 0.5, h_n = 0.1$.

During training, we use an Adam optimizer with a learning rate of 10^{-4}. A mini-batch contains 10 images per GPU. The backbone network ResNet-50 is pre-trained on ImageNet [85]. During evaluation, we test on the original image size unless otherwise stated.

5. Supervised Experiment

5.1. State-of-the-art Comparisons on CityPersons

In this section, we compare PRNet and PRNet++ under a supervised within-dataset setting on CityPersons against two groups of state-of-the-art methods: Occlusion-free and occlusion-aware. Since detection speed is important in real-world applications, inference time based on the original image resolution (2048×1024) are also compared. Following the literature, MR^{-2} on all 6 subsets are reported. Specially, MR^{-2} on **R** is the official major evaluation criteria in CityPersons Challenge[2]. Table 1 presents the comparisons.

Occlusion-free methods: *Occlusion-free* methods perform pedestrian detection without any specific strategies for occlusion handling. Adapted FasterRCNN [17] is an anchor-based baseline detector. Both TLL + MRF [10] and CSP [14] are anchor-free methods. As shown in Table 1, CSP achieved the best performance among the three methods without considering occlusions. The proposed PRNet and PRNet++, on the other hand, take occlusion handling into account, and provide performance gain over CSP on **R**, Partial, and Bare subsets except the Heavy subset. A potential explanation is that CSP exploited box-free annotations, which is different from the original ground truth and might help alleviate the bad effect of annotation noises in heavy occlusion cases.

Occlusion-aware methods: *Occlusion-aware* methods take various occlusion–handling strategies during training, including FasterRCNN + ATT [28], RepLoss [15], OR-CNN [16], FRCN + A+DT [12], and MGAN [27]. Generally, occlusion-aware methods show better performance than occlusion-free methods, except for CSP that used different box-free ground-truth annotations. Among all occlusion-aware alternatives, PRNet++ consistently achieves the best MR^{-2} of (10.7, 40.9, 51.2, 25.4, 9.9) on (**R**, **HO**, Heavy, **R + HO**, Partial). The comparisons firmly validate PRNet++'s effectiveness in occlusion handling by taking progressive refinement and dual-stream processing.

Inference time comparisons: Regarding inference time, we firstly divide these methods into three categories: Faster-RCNN-like two-stage design (including FasterRCNN [17], FasterRCNN + ATT [28], RepLoss [15], OR-CNN [16], MGAN [27], FRCN + A+DT [12], Bi-box [29]), single-stage design (including ALFNet [11], PRNet and PRNet++) and anchor-free design(including CSP [14] and TLL + MRF [10]). According to [11,40], we infer that single-stage methods raise a 2-6x speedup than Faster-RCNN-like methods. Among single-stage methods, PRNet performs comparably with ALFNet. Also, PRNet++ dosen't involve too much computation since it keeps the single-stage design of PRNet. Due to lack of source code, we can not figure out the inference time of TLL + MRF [10]. As the best anchor-free pedestrian detector up to now, CSP is slower than our PRNet and PRNet++. In conclusion, our methods achieve an excellent balance between accuracy and speed.

5.2. State-of-the-art Comparisons on Caltech

This section compares our PRNet and PRNet++ with state-of-the-art methods on Caltech under three occlusion subsets: **R**, **HO** and **R + HO**, which is presented in Table 2. Our PRNet++ consistently achieves the best MR^{-2} of (5.2, 35.3, 13.0) on (**R**, **HO**, **R + HO**), which outperforms the previous best results of MGAN [27] by (1.6, 2.9, 0.6) MR^{-2}. What's more, our PRNet achieves the second best MR^{-2} of (5.76, 35.59, 13.59), which also exceeds MGAN [27] on all subsets. Such a big gap demonstrates the superiority of our PRNet and PRNet++.

5.3. State-of-the-art Comparisons on ECP-night

It is challenging to detect occluded pedestrians under poor lighting conditions due to the low contrast, color loss and motion blur. To further validate the effectiveness of our methods, we conducted experiments on ECP-night [32], which was recorded at night with various poor lighting scenes. For comparison, we pick two excellent methods, CSP [14] and ALFNet [11], which achieve the best and third best performance on **Heavy** occlusion subset of CityPersons as shown in Table 1 (only take into account results based on the original image resolution). All methods are trained with ResNet-50 as backbone and share the same pre-pocessing. Table 3 summarize the comparisons. It can be seen that Our PRNet++ consistently achieves the best MR^{-2} of (18.7, 43.3, 31.7) on (reasonable, occluded, all) set, which outperforms other methods by a big margin. The comparisons show the superiority of PRNet++ in occlusion handling under poor lighting conditions.

5.4. Qualitative Results

We present some qualitative detection examples of PRNet++ on CityPersons and ECP-night in Fig. 10, where we can see PRNet++ produces robust detections for different occlusion patterns and lighting conditions.

[1] http://www.cvlibs.net/datasets/kitti/

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

Table 1
Comparisons on CityPersons [17]. Results of alternatives were obtained from original paper. On scale×1, bracketed and bold numbers indicate the best and the second best results, respectively. The scale means the enlarger factor of original image resolution (2048×1024) during inference. Inference time (*sec*) is measured on scale×1 images.

Method	Occ.	Scale	**R**	**HO**	**R + HO**	**Heavy**	**Partial**	**Bare**	**Time**
FasterRCNN [17]		×1	15.4	64.8	41.45	55.0	18.9	9.3	-
TLL + MRF [10]		×1	14.4	-	-	52.0	15.9	9.2	-
CSP [14]		×1	11.0	-	-	[**49.3**]	10.4	7.3	0.21
FasterRCNN + ATT[28]	✓	×1	16.0	56.7	38.2	-	-	-	-
		×1	13.2	-	-	56.9	16.8	7.6	-
RepLoss [15]	✓	×1.3	11.6	-	-	55.3	14.8	7.0	-
OR-CNN [16]	✓	×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[12]	✓	×1.3	11.1	44.3	-	-	11.2	6.9	-
ALFNet [11]		×1	12.0	43.8	26.3	51.9	11.4	8.4	0.12
Bi-box [29]	✓	×1.3	11.2	44.2	-	-	-	-	-
PRNet (ours)	✓	×1	**10.8**	**42.0**	**25.6**	53.3	**10.0**	**6.8**	0.14
PRNet++(ours)	✓	×1	[**10.7**]	[**40.9**]	[**25.4**]	**51.2**	[**9.9**]	6.9	0.20

Table 2
Comparisons on Caltech [18]. Results of alternatives were obtained from [27]. Best results are in bold.

Method	Occ.	**R**	**HO**	**R + HO**
ComACT-Deep [86]		11.75	65.78	24.61
DeepParts [26]	✓	11.89	60.42	22.79
MCF [87]		10.40	66.969	22.85
FasterRCNN + ATT [28]	✓	10.33	45.18	18.21
MS-CNN [88]		9.95	59.94	21.53
RPN + BF [37]		9.58	74.36	24.01
SA-FRCNN []		9.68	64.35	21.92
SDS-RCNN [38]		7.36	58.55	19.72
FasterRCNN [17]		9.18	57.58	20.03
GDFL [31]		7.85	43.18	15.64
Bi-Box [29]	✓	7.61	44.40	16.06
MGAN [27]	✓	6.83	38.16	13.84
PRNet(ours)	✓	5.76	35.59	13.59
PRNet++(ours)	✓	**5.2**	**35.3**	**13.0**

Table 3
Comparisons on ECP-night[32]. Best results are in bold.

Method	**reasonable**	**occluded**	**all**
CSP [14]	19.6	44.5	33.1
ALFNet [11]	24.5	49.0	37.1
PRNet(ours)	19.4	45.7	32.8
PRNet++(ours)	**18.7**	**43.3**	**31.7**

5.5. Ablation Study of PRNet

We run a number of ablations to analyze PRNet. Results were reported on CityPersons [17] validation set using subsets of **R** (reasonable) and **HO** (heavy occlusion).

Three-phase components: To demonstrate the effectiveness of PRNet's three-phase design, we performed ablation study on each phase without exploiting occlusion loss and RFB module in FR. We list the experimental results in Table 4. **PRNet-F** was a standalone FR detector that was initialized by predefined full-body anchors. **PRNet-VA** used only VE + AC, treating calibrated anchors $\mathscr{A}_2$ as final detection outputs. **PRNet-VAF** included all three phases (VE + AC + FR), where calibrated anchors $\mathscr{A}_2$ are used to initialize FR. Comparing **PRNet-F** and **PRNet-VA**, **PRNet-VA** performs 3.9 points better on **R** but 5.6 points worse on **HO**. This shows that plain calibrated anchors $\mathscr{A}_2$ in **PRNet-VA** can achieve better results when occlusion level is reasonable. In contrast, **PRNet-F** performs better in heavy occlusions. **PRNet-VAF** combines the benefits from both, showing a consistent improvement over both **R** and **HO**.

Anchor calibration vs box regression: A possible alternative to AC is an extra box regression from visible-part anchors $\mathscr{A}_1$ to full-body bounding boxes. Here we reused FR for the regression task. For a fair comparison, we implemented **PRNet-VRF** by replacing AC with an extra box regression learning branch. Table 4 summarizes the results. As can be seen, **PRNet-VAF** consistently outperformed **PRNet-VRF** by 9.5% in **R**, which shows no significant benefits of adding an extra learning branch. An explanation can be that the visible boxes change rapidly due to various occlusion types, and make it hard to map the coordinates to full-body boxes with relatively constant aspect ratio. Unlike a regression learning branch that require extra complexity and training efforts, AC offers a more generalizable strategy that better fits into the proposed three-phase framework.

Occlusion loss and RFB: Table 5 studies PRNet w/ and w/o occlusion loss and RFB module. **PRNet-VAF** was reused as the baseline that exploits 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 0.4 points over the baseline on **R**, yet lowered 0.4 points on **HO**. This illustrates that occlusion loss can improve detection performance in reasonable occlusion (*i.e.*, over 0.65 visibility), but 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 information by enlarging the receptive field, and thus yields improvement when occlusion is severe. Otherwise, when occlusion level is slight, enlarging receptive field may introduce unnecessary context and hence slightly hurt. PRNet couples occlusion loss and RFB, achieving significant improvement on **R** and especially **HO**.

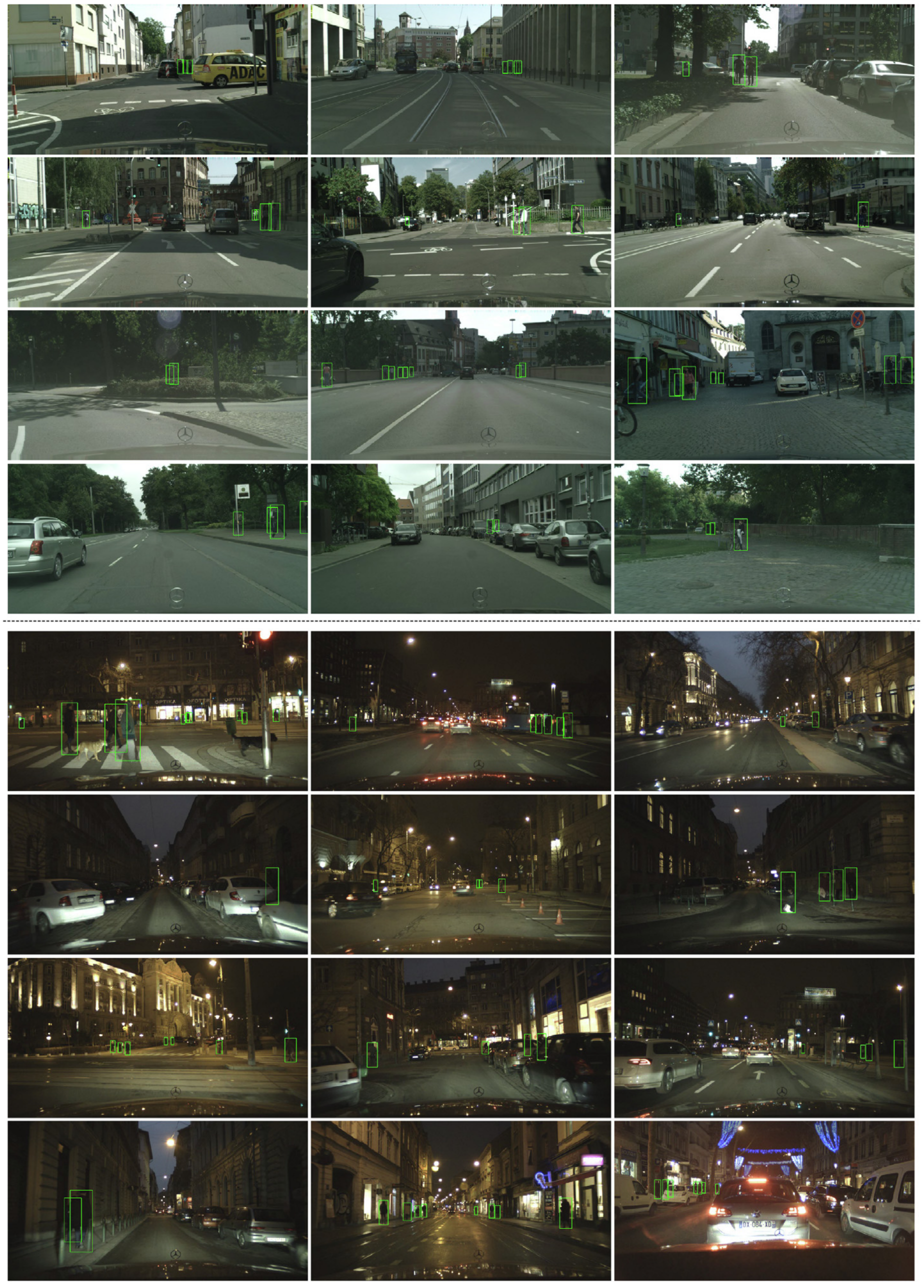


Fig. 10. Qualitative detection results of PRNet++ at FPPI of 0.3 on CityPersons (top 4 rows) and ECP-night (bottom 4 rows)..

Table 4
Ablations of **three-phase components** and an alternative. reported in MR^{-2}.

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 5
Ablations of **occlusion loss** and the **RFB module**. reported in MR^{-2}.

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

Table 6

Within-dataset comparisons of PRNet and PRNet++ on CityPersons, Caltech and INRIA. Numbers denote MR^{-2}. ⇓ menas lower is better).

Method	Citypersons				Caltech				INRIA
	R⇓	**HO⇓**	**Heavy⇓**	**R + HO⇓**	**R⇓**	**HO⇓**	**Heavy⇓**	**R + HO⇓**	**R⇓**
PRNet	10.8	42.0	53.3	25.6	5.8	35.6	40.4	13.6	4.4
PRNet++ Easy-branch	10.9	41.1	51.3	25.8	5.6	35.8	39.1	13.3	4.0
PRnet++ Hard-branch	11.0	41.2	51.7	25.7	5.9	34.9	38.7	13.3	4.1
PRNet++	10.7	40.9	51.2	25.4	5.2	35.2	39.0	13.0	4.0

Table 7

Within-dataset comparisons of PRNet and PRNet++ on ECP-night. Numbers denote MR^{-2}. ⇓ menas lower is better).

Method	**reasonable⇓**	**occluded⇓**	**all⇓**
PRNet	19.4	45.7	32.8
PRNet++ Easy-branch	17.7	47.7	31.9
PRNet++ Hard-branch	19.1	42.9	32.0
PRNet++	18.7	43.3	31.7

Table 8

Within-dataset comparisons of PRNet and PRNet++ on KITTI. Numbers denote *AP*. ⇑ means higher is better.

Method	**Easy⇑**	**Moderate⇑**	**Hard⇑**
PRNet	83.4	77.19	71.2
PRNet++ Easy-branch	84.4	78.1	72.1
PRNet++ Hard-branch	83.7	78.0	72.3
PRNet++	84.3	78.1	72.3

5.6. PRNet vs PRNet++:

PRNet++ is a new proposed framework that split the detection heads into easy and hard branch to handle data in different occlusion levels. To validate the effectiveness of dual-stream structure, we compare the performance of PRNet and PRNet++ in this section.

We conducted supervised within-dataset experiments with PRNet and PRNet++ on five datasets. Following [27,16,28], models on Caltech, INRIA and KITTI are finetuned from the model pretrained on CityPersons. Table 6, Table 7 and Table 8 summarize the camparisons. It can be seen that PRNet++ exceeds PRNet in almost every subsets of all datasets. To further analysis, we also present results from Easy-branch and Hard-branch of PRNet++. As expected, Hard-branch shows better performance on subsets in heavy occlusion level than Easy-branch. The whole PRNet++ with fused results by NMS and Support-areas voting show a large improvement over PRNet, which illustrates that the two branches complement each other.

6. Unsupervised Domain Adaptation Experiment

In this section, we conducted experiments under unsupervised domain adation setting to validate the effectiveness of our proposed domain adaptation methods on CityPersons [17], Caltech [18], ECP-night [32], KITTI [1] and Inria [33].

6.1. Results and Comparisons

As introduced in Section 2, [77] is the only existed work in unsupervised domain adaptive pedestrian detection task. Since it is published in 2016, the benchmarks, experiments settings and the network architecture are very different with works in recent years. Also, it cannot be embedded into the state-to-the-art detection frameworks. Hence, we cannot fairly compare it with our proposed methods quantitatively. In this section, We compare our adaptation methods against the following common methods in unsupervised domain adaptive object detection:

PL in [89] fine-tunes the source domain model with pseudo labels, where pseudo labels are generated only once before training.

ST [75] is the naive approach of self-training that utilizes pseudo labels as ground-truth without localization loss. During training, pseudo labels are updated by the up-to-date model every single iteration.

WST in [75] is built upon ST, which aims to generate more reliable pseudo labels. To reduce false negatives, the gradients of some background examples are masked out. To reduce false positives, a Supporting Region-based Reliable Score strategy (SRRS) is proposed to give more reliable confidence scores to detected boxes. Note that WST also doesn't utilize the localization loss during training.

We also introduce the performance lower bound and upper bound of unsupervised domain adaptation methods:

Lower bound We directly apply pre-trained models from source domain on the test set of target domain without finetuning, which can be seen as a lower bound of unsupervised domain adaptation methods.

Upper bound We use ground-truth labels of training set on target domain for supervised fine-tuning. It can reflect the performance upper bound for unsupervised domain adaptation methods.

For fair comparisons, we implement the above methods based on our PRNet with the same configurations and employ the mean-teacher training strategy. As CityPersons is a relatively diverse and dense dataset, we establish four unsupervised domain adaptation tasks: $CityPersons \rightarrow Caltech$, $CityPersons \rightarrow ECP - night$, $CityPersons \rightarrow KITTI$ and $CityPersons \rightarrow INRIA$, where $A \rightarrow B$ represents A is a source domain and B is a target domain. Tables 9–12 summarize the comparisons in three adaptation tasks respectively. It can be seen that our proposed Multi-experts adaptation strategy clearly outperforms other methods in all tasks and subsets, which demonstrates its effectiveness of tackling cross-domain pedestrian detection problem without any labels in target domains. Also, our proposed Dynamic Iterative adaptation strategy performs better than the naive methods PL [89] and ST [75]. On the other hand, WST [75] shows a competitive performance in $CityPersons \rightarrow KITTI$ but the worst overall performance in **R + HO** of $CityPersons \rightarrow KITTI$ and **All** set of $CityPersons \rightarrow INRIA$. The inconsistent performance in different tasks reflect its limitation. When being utilized in pedestrian detection, WST [75] cannot provide a stable guideline of optimization process. It is mainly due to the complex occlusion situations in pedestrian detection, which is different from general object detection. In comparison, by taking occlusion issue into consideration, our proposed adaptation strategies can effectively handle different occlusions in different domains.

6.2. Ablation Study

To validate the effectiveness of our proposed domain adaptation strategies, ablation experiments are conducted on CityPersons

Table 9
Comparison of various methods in terms of MR^{-2} from CityPersons to Caltech. ⇓ means lower is better. Descriptions of each methods is in Section 6.1.

Method	Source domain	Target domain	**R**⇓	**HO**⇓	**Heavy**⇓	**R + HO**⇓
Lower bound	CityPersons	Caltech	10.7	44.2	48.1	19.2
Upper bound	CityPersons	Caltech	5.8	35.6	40.4	13.6
PL [89]	CityPersons	Caltech	8.7	42.5	47.7	17.3
ST [75]	CityPersons	Caltech	9.0	40.8	46.1	17.1
WST [75]	CityPersons	Caltech	9.6	40.8	45.4	17.5
Dynamic Iterative strategy (ours)	CityPersons	Caltech	7.8	40.2	44.4	16.0
Multi-experts strategy (ours)	CityPersons	Caltech	7.7	39.6	43.7	16.0

Table 10
Comparison of various methods in terms of MR^{-2} from CityPersons to ECP-night. ⇓ means lower is better. Descriptions of each methods is in Section 6.1.

Method	Source domain	Target domain	**reasonable**⇓	**occluded**⇓	**all**⇓
Lower bound	CityPersons	ECP-night	65.2	83.2	72.5
Upper bound	CityPersons	ECP-night	19.4	45.7	32.8
PL [89]	CityPersons	ECP-night	56.3	79.2	65.5
ST [75]	CityPersons	ECP-night	55.1	80.0	64.8
WST [75]	CityPersons	ECP-night	60.5	82.7	68.9
Dynamic Iterative strategy (ours)	CityPersons	ECP-night	49.8	75.9	60.2
Multi-experts strategy (ours)	CityPersons	ECP-night	49.5	75.3	59.8

Table 11
Comparison of various adaptation methods from CityPersons to KITTI in terms of *AP*. ⇑ means higher is better. Descriptions of each methods is in Section 6.1.

Method	Source domain	Target domain	**Easy**⇑	**Moderate**⇑	**Hard**⇑
Lower bound	CityPersons	KITTI	77.6	71.6	65.4
Upper bound	CityPersons	KITTI	83.4	77.2	71.2
PL [89]	CityPersons	KITTI	78.4	72.5	66.3
ST [75]	CityPersons	KITTI	79.0	73.0	66.9
WST [75]	CityPersons	KITTI	79.5	73.1	67.2
Dynamic Iterative strategy (ours)	CityPersons	KITTI	79.3	73.1	66.5
Multi-experts strategy (ours)	CityPersons	KITTI	79.6	73.7	67.7

Table 12
Comparison of various adaptation methods from CityPersons to INRIA in terms of MR^{-2}. ⇓ means lower is better. Descriptions of each methods is in Section 6.1.

Method	Source domain	Target domain	**All**⇓
Lower bound	CityPersons	INRIA	24.5
Upper bound	CityPersons	INRIA	4.4
PL [89]	CityPersons	INRIA	8.4
ST [75]	CityPersons	INRIA	11.0
WST [75]	CityPersons	INRIA	11.8
Dynamic Iterative strategy (ours)	CityPersons	INRIA	7.2
Multi-experts strategy (ours)	CityPersons	INRIA	6.9

[17], Caltech [18], ECP-night [32], KITTI [1] and Inria [33]. Note that $A \rightarrow B$ represents A is a source domain and B is a target domain. The lower bound and upper bound were introduced in Section 6.1. The baseline in this section refers to PL [75].

Effectivness of Dynamic Iterative adaptation strategy. From results presented in Table 13 (*CityPersons → Caltech*), Table 14 (*CityPersons → ECP − night*),Table 15 (*CityPersons → KITTI*) and Table 16 ((*CityPersons → INRIA*)), we can find that our Dynamic Iterative strategy raises a consistently improvement over the baseline in all subsets. It proves that iteratively updating pseudo labels by teacher model with dynamic sample selection threshold can facilitate the training of student models. Since teacher model is the ensemble of student models in different epochs, a positive circle is formed.

Effectiveness of Multi-experts adaptation strategy. To validate the effectiveness of our Multi-experts adaptation strategy, we presents results of every experts when they are exploited as a stand-alone detector. As shown in Table 13 (*CityPersons → Caltech*), Experts B and C exceed Experts A by a big margin on **HO** and **H**, but cannot get better results on **R**. The similar phenomenon can be found in Table 14 (*CityPersons → ECP − night*). Since Expert B and C all exploit strong augmentation to convert some slightly-occluded samples to heavily-occluded ones, they learn more heavier occluded samples than Expert A and do better in occlusion situations naturally. In Table 15 (*CityPersons → KITTI*), it can be seen that Experts B and C outperform Experts A in all subsets. A possible reason is that strong augmentation can also alleviate the over-fitting problem and reduce some false detections, which contribute to the overall results. As no occlusion subsets can be split in INRIA, we can only get the overall results in Table 16 (*CityPersons → INRIA*) and the phenomenon is similar to the above tasks. Although experts perform differently on these datasets, the common conclusion is that experts with different architecture and augmentations generate different detection results and perform well in different occlusion scenes. Finally, let's see the final results comparisons. Multi-experts adaptation strategy outperform Dynamic Iterative adaptation strategy by a remarkable margin, which proves that the mixture of different experts can provided more reliable supervision information.

Different domain combinations. To investigate the robustness of our proposed adaptation strategies, we establish more adaptation tasks with different domain combinations and present results in Table 17 (*KITTI → INRIA*), Table 18 (*KITTI → Caltech*), Table 19 (*Caltech → KITTI*), Table 20 (*Caltech → INRIA*). Since both of Caltech and KITTI don't have enough training instances to be a source domain alone, the initial source domain models are fine-tuned from CityPersons so as to make up the problem of source domain

Table 13
Results of adaptation from CityPersons to Caltech. Numbers refer to MR^{-2}. ⇓ means lower is better.

Method	Source domain	Target domain	**R**⇓	**HO**⇓	**Heavy**⇓	**R + HO**⇓
Lower bound	CityPersons	Caltech	10.7	44.2	48.1	19.2
Upper bound	CityPersons	Caltech	5.8	35.6	40.4	13.6
Baseline	CityPersons	Caltech	8.7	42.5	47.7	17.3
Dynamic Iterative strategy (Experts A)	CityPersons	Caltech	7.8	40.2	44.4	16.0
Experts B	CityPersons	Caltech	9.3	38.5	42.2	16.9
Experts C	CityPersons	Caltech	9.3	36.6	41.4	16.5
Multi-experts strategy	CityPersons	Caltech	7.7	39.6	43.7	16.0

Table 14
Results of adaptation from CityPersons to ECP-night. Numbers refer to MR^{-2}. ⇓ means lower is better.

Method	Source domain	Target domain	**reasonable**⇓	**occluded**⇓	**all**⇓
Lower bound	CityPersons	ECP-night	65.2	83.2	72.5
Upper bound	CityPersons	ECP-nighth	19.4	45.7	32.8
Baseline	CityPersons	ECP-night	56.3	79.2	65.5
Dynamic Iterative strategy (Experts A)	CityPersons	ECP-night	49.8	75.9	60.2
Experts B	CityPersons	ECP-night	50.0	75.3	60.1
Experts C	CityPersons	ECP-night	49.8	75.3	59.9
Multi-experts strategy	CityPersons	ECP-night	49.5	75.3	59.8

Table 15
Results of adaptation from CityPersons to KITTI. Number refers to AP. ⇑ means higher is better.

Method	Source domain	Target domain	**Easy**⇑	**Moderate**⇑	**Hard**⇑
Lower bound	CityPersons	KITTI	77.6	71.6	65.4
Upper bound	CityPersons	KITTI	83.4	77.2	71.2
Baseline	CityPersons	KITTI	78.4	72.5	66.3
Dynamic Iterative strategy (Experts A)	CityPersons	KITTI	79.3	73.1	66.5
Experts B	CityPersons	KITTI	79.6	73.5	67.6
Experts C	CityPersons	KITTI	79.6	73.7	67.1
Multi-experts strategy	CityPersons	KITTI	79.6	73.7	67.7

Table 16
Results of adaptation from CityPersons to INRIA. Numbers refer to MR^{-2}. ⇓ means lower is better.

Method	Source domain	Target domain	**All**⇓
Lower bound	CityPersons	INRIA	24.5
Upper bound	CityPersons	INRIA	4.4
Baseline	CityPersons	INRIA	8.4
Dynamic Iterative strategy (Experts A)	CityPersons	INRIA	7.2
Experts B	CityPersons	INRIA	7.1
Experts C	CityPersons	INRIA	7.5
Multi-experts strategy	CityPersons	INRIA	6.9

Table 17
Results of adaptation from Kitti to Inria. Numbers refer to MR^{-2}. ⇓ means lower is better.

Method	Source domain	Target domain	**All**⇓
Lower bound	KITTI	INRIA	16.7
Upper bound	KITTI	INRIA	4.4
Baseline	KITTI	INRIA	8.3
Dynamic Iterative strategy	KITTI	INRIA	7.6
Multi-experts strategy	KITTI	INRIA	7.3

data insufficiency. Results of different adaptation tasks consistently show the effectiveness of our proposed Dynamic Iterative strategy and Multi-experts strategy.

7. Conclusion

In this paper, we aim to learn towards generalized occluded pedestrian detection. We have proposed Progessive Redinement Network(PRNet), a novel single-stage detector to tackle occluded pedestrian detection. PRNet incorporates three phases, Visible-part Estimation (VE), Anchor Calibration (AC), and Full-body Refinement (FR), to gradually evolve anchors towards target localization. To encourage learning on occluded samples, we introduced an occlusion loss and a Receptive Field Backfeed (RFB) module. Extensive ablation studies are conducted to justify the need of each component. To learn more generalized representations to handle various occlusions, we further propose a new single-stage detector PRNet++ with dual-stream structure, which consists of an Easy-branch and a Hard-branch to generate complementary results. Supervised within-dataset experiments validated the effectiveness of PRNet and PRNet++ in various occlusion scenarios. Furthermore, we give insight into the novel unsupervised domain adaptive pedestrian detection task as the generalization ability to other domains is also crucial. A Dynamic Iterative adaptation strategy

Table 18
Results of adaptation from KITTI to Caltech. Numbers refer to MR^{-2}. ⇓ means lower is better.

Method	Source domain	Target domain	**R**⇓	**HO**⇓	**Heavy**⇓	**R + HO**⇓
Lower bound	KITTI	Caltech	10.8	44.9	48.8	19.3
Upper bound	KITTI	Caltech	5.8	35.6	40.4	13.6
Baseline	KITTI	Caltech	8.7	41.3	45.5	17.1
Dynamic Iterative strategy	KITTI	Caltech	8.7	41.3	45.5	17.1
Multi-experts strategy	KITTI	Caltech	8.5	41.6	46.2	16.9

Table 19
Results of adaptation from Caltech to KITTI. Number refers to AP. $\Uparrow$ means higher is better.

Method	Source domain	Target domain	**Easy**$\Uparrow$	**Moderate**$\Uparrow$	**Hard**$\Uparrow$
Lower bound	Caltech	KITTI	79.1	72.8	66.7
Upper bound	Caltech	KITTI	83.4	77.2	71.2
Baseline	Caltech	KITTI	79.2	73.1	67.1
Dynamic Iterative strategy	Caltech	KITTI	79.2	73.1	67.2
Multi-experts strategy	Caltech	KITTI	79.4	73.8	68.0

Table 20
Results of adaptation from Caltech to INRIA. Numbers refer to MR^{-2}. $\Downarrow$ means lower is better.

Method	Source domain	Target domain	**All**$\Downarrow$
Lower bound	Caltech	INRIA	18.0
Upper bound	Caltech	INRIA	4.4
Baseline	Caltech	INRIA	9.7
Dynamic Iterative strategy	Caltech	INRIA	7.1
Multi-experts strategy	Caltech	INRIA	6.7

and a Multi-experts strategy are proposed to optimize domain adaptation process. Unsupervised domain adaptation experiments show the effectiveness of our proposed methods. Although our proposed domain adaptation strategies are evaluated on PRNet and PRNet++, the key ideas are not restricted to detection models. Limited by the single-stage detector design of PRNet and PRNet++, we cannot obtain cropped local features of each detected instance and further rectify false pseudo annotations based on feature discrepancy. For other detectors, *e.g.*, Faster R-CNN based detectors [27] [29], we can construct different experts based on their architectures and make some efforts on ROI features, which we will leave for future.

CRediT authorship contribution statement

Xiaolin Song: Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Binghui Chen:** Methodology, Validation, Writing - review & editing. **Pengyu Li:** Writing - review & editing. **Biao Wang:** Writing - review & editing. **Honggang Zhang:** Supervision, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

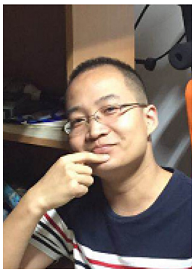
This work was funded by National Natural Science Foundation of China under Grant No.62076034, No.62076036. We hereby give specifical thanks to Alibaba Group for their contribution to this paper.

References

[1] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: CVPR, 2012, pp. 3354–3361. .
[2] Efficient deep network for vision-based object detection in robotic applications, Neurocomputing 245 (2017) 31–45. .
[3] H. Dhia, R.F. Ghani, A proposed method for scale drawing calculating depending on line detector and length detector, Iraqi Journal For Computer Science and Mathematics 2 (2) (2021) 6–17.
[4] Z.R. Mohsin, Investigating the use of an adaptive neuro-fuzzy inference system in software development effort estimation, Iraqi Journal For Computer Science and Mathematics 2 (2) (2021) 18–24.
[5] J. Nascimento, J. Marques, Performance evaluation of object detection algorithms for video surveillance, IEEE Transactions on Multimedia 8 (4) (2006) 761–774, https://doi.org/10.1109/TMM.2006.876287.
[6] Pedestrian search in surveillance videos by learning discriminative deep features, Neurocomputing 283 (2018) 120–128. .
[7] Hogg: Gabor and hog-based human detection for surveillance in non-controlled environments, Neurocomputing 100 (2013) 19–30, special issue: Behaviours in video. .
[8] , Neurocomputing 449 (2021) 229–244.
[9] J. Noh, S. Lee, B. Kim, G. Kim, Improving occlusion and hard negative handling for single-stage pedestrian detectors, in: CVPR, 2018, pp. 966–974. .
[10] T. Song, L. Sun, D. Xie, H. Sun, S. Pu, Small-scale pedestrian detection based on topological line localization and temporal feature aggregation, in: ECCV, 2018, pp. 536–551. .
[11] W. Liu, S. Liao, W. Hu, X. Liang, X. Chen, Learning efficient single-stage pedestrian detectors by asymptotic localization fitting, in: ECCV, 2018, pp. 618–634. .
[12] C. Zhou, M. Yang, J. Yuan, Discriminative feature transformation for occluded pedestrian detection, in: ICCV, 2019, pp. 9557–9566..
[13] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection?, in: CVPR, 2016, pp. 1259–1267. .
[14] W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: A new perspective for pedestrian detection, in: CVPR, 2019, pp. 5187–5196..
[15] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, C. Shen, Repulsion loss: Detecting pedestrians in a crowd, in: CVPR, 2018, pp. 7774–7783. .
[16] S. Zhang, L. Wen, X. Bian, Z. Lei, S.Z. Li, Occlusion-aware R-CNN: Detecting pedestrians in a crowd, in: ECCV, 2018, pp. 637–653. .
[17] S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, in: CVPR, 2017, pp. 3213–3221..
[18] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2011) 743–761.
[19] V.D. Shet, J. Neumann, V. Ramesh, L.S. Davis, Bilattice-based logical reasoning for human detection, in: CVPR, 2007, pp. 1–8. .
[20] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors, in: ICCV, 2005, pp. 90–97. .
[21] M. Enzweiler, A. Eigenstetter, B. Schiele, D.M. Gavrila, Multi-cue pedestrian classification with partial occlusion handling, in: CVPR, 2010, pp. 990–997. .
[22] G. Duan, H. Ai, S. Lao, A structural filter approach to human detection, in: ECCV, 2010, pp. 238–251..
[23] W. Ouyang, X. Wang, A discriminative deep model for pedestrian detection with occlusion handling, in: CVPR, 2012, pp. 3258–3265..
[24] M. Mathias, R. Benenson, R. Timofte, L.V. Gool, Handling occlusions with franken-classifiers, in: ICCV, 2013, pp. 1505–1512. .
[25] W. Ouyang, X. Zeng, X. Wang, Modeling mutual visibility relationship in pedestrian detection, in: CVPR, 2013, pp. 3222–3229. .
[26] Y. Tian, P. Luo, X. Wang, X. Tang, Deep learning strong parts for pedestrian detection, in: ICCV, 2015, pp. 1904–1912..
[27] Y. Pang, J. Xie, M.H. Khan, R.M. Anwer, F.S. Khan, L. Shao, Mask-guided attention network for occluded pedestrian detection, in: ICCV, 2019, pp. 4967–4975. .
[28] S. Zhang, J. Yang, B. Schiele, Occluded pedestrian detection through guided attention in CNNs, in: CVPR, 2018, pp. 6995–7003. .
[29] C. Zhou, J. Yuan, Bi-box regression for pedestrian detection and occlusion estimation, in: ECCV, 2018, pp. 135–151. .
[30] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: NIPS, 2015, pp. 91–99. .
[31] C. Lin, J. Lu, G. Wang, J. Zhou, Graininess-aware deep feature learning for pedestrian detection, in: ECCV, 2018, pp. 732–747..
[32] M. Braun, S. Krebs, F.B. Flohr, D.M. Gavrila, Eurocity persons: A novel benchmark for person detection in traffic scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019), 1–1.
[33] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR. .
[34] X. Song, K. Zhao, W.-S. Chu, H. Zhang, J. Guo, Progressive refinement network for occluded pedestrian detection, in: ECCV, 2020, pp. 32–48..
[35] Recent advances in deep learning for object detection, Neurocomputing 396 (2020) 39–64. .
[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, in: ECCV, 2016, pp. 21–37..
[37] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection?, in: ECCV, 2016, pp. 443–457. .
[38] G. Brazil, X. Yin, X. Liu, Illuminating pedestrians via simultaneous detection & segmentation, in: ICCV, 2017, pp. 4950–4959. .
[39] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-cnn for pedestrian detection, IEEE Transactions on Multimedia 20 (4) (2017) 985–996.
[40] G. Brazil, X. Liu, Pedestrian detection with autoregressive network phases, in: CVPR, 2019, pp. 7231–7240. .
[41] R. Girshick, Fast R-CNN, in: ICCV, 2015, pp. 1440–1448. .

[42] W. Ouyang, X. Wang, Joint deep learning for pedestrian detection, in: ICCV, 2013, pp. 2056–2063..
[43] C. Zhou, J. Yuan, Multi-label learning of part detectors for heavily occluded pedestrian detection, in: ICCV, 2017, pp. 3486–3495. .
[44] W. Ouyang, X. Wang, Single-pedestrian detection aided by multi-pedestrian detection, in: CVPR, 2013, pp. 3198–3205. .
[45] S. Tang, M. Andriluka, B. Schiele, Detection and tracking of occluded people, International Journal of Computer Vision 110 (1) (2014) 58–69.
[46] B. Pepikj, M. Stark, P. Gehler, B. Schiele, Occlusion patterns for object class detection, in: CVPR, 2013, pp. 3286–3293. .
[47] S. Liu, D. Huang, Y. Wang, Adaptive nms: Refining pedestrian detection in a crowd, in: CVPR, 2019, pp. 6459–6468. .
[48] X. Huang, Z. Ge, Z. Jie, O. Yoshie, Nms by representative region: Towards crowded pedestrian detection by proposal pairing, in: CVPR, 2020, pp. 10750–10759. .
[49] Mapd: An improved multi-attribute pedestrian detection in a crowd, Neurocomputing 432 (2021) 101–110. .
[50] Lla: Loss-aware label assignment for dense pedestrian detection, Neurocomputing 462 (2021) 272–281. .
[51] Semantic head enhanced pedestrian detection in a crowd, Neurocomputing 400 (2020) 343–351. .
[52] Y. Luo, C. Zhang, M. Zhao, H. Zhou, J. Sun, Where, what, whether: Multi-modal learning meets pedestrian detection, in: CVPR, 2020, pp. 14065–14073..
[53] J. Wu, C. Zhou, M. Yang, Q. Zhang, Y. Li, J. Yuan, Temporal-context enhanced detection of heavily occluded pedestrians, in: CVPR, 2020, pp. 13430–13439. .
[54] Y. Zhang, H. He, J. Li, Y. Li, J. See, W. Lin, Variational pedestrian detection, in: CVPR, 2021, pp. 11622–11631. .
[55] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153.
[56] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: CVPR, 2017, pp. 7167–7176. .
[57] S. Sankaranarayanan, Y. Balaji, A. Jain, S.N. Lim, R. Chellappa, Learning from synthetic data: Addressing domain shift for semantic segmentation, in: CVPR, 2018, pp. 3752–3761. .
[58] W. Zhang, W. Ouyang, W. Li, D. Xu, Collaborative and adversarial network for unsupervised domain adaptation, in: CVPR, 2018, pp. 3801–3809..
[59] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: CVPR, 2018, pp. 3723–3732. .
[60] Z. Lu, Y. Yang, X. Zhu, C. Liu, Y.-Z. Song, T. Xiang, Stochastic classifiers for unsupervised domain adaptation, in: CVPR, 2020, pp. 9111–9120. .
[61] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: ICML, 2015, pp. 1180–1189. .
[62] R. Volpi, P. Morerio, S. Savarese, V. Murino, Adversarial feature augmentation for unsupervised domain adaptation, in: CVPR, 2018, pp. 5495–5504. .
[63] Unsupervised domain adaptation via representation learning and adaptive classifier learning, Neurocomputing 165 (2015) 300–311. .
[64] A novel class restriction loss for unsupervised domain adaptation, Neurocomputing 461 (2021) 254–265. .
[65] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, Cycada: Cycle-consistent adversarial domain adaptation, in: ICML, 2018, pp. 1989–1998. .
[66] W. Hong, Z. Wang, M. Yang, J. Yuan, Conditional generative adversarial network for structured domain adaptation, in: CVPR, 2018, pp. 1335–1344. .
[67] Y. Zou, Z. Yu, B. Kumar, J. Wang, Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: ECCV, 2018, pp. 289–305. .
[68] Y. Chen, W. Li, C. Sakaridis, D. Dai, L. Van Gool, Domain adaptive faster r-cnn for object detection in the wild, in: CVPR, 2018..
[69] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, M.-H. Yang, Progressive domain adaptation for object detection, in: WACV, 2020, pp. 749–757..
[70] K. Saito, Y. Ushiku, T. Harada, K. Saenko, Strong-weak distribution alignment for adaptive object detection, in: CVPR, 2019, pp. 6956–6965..
[71] X. Zhu, J. Pang, C. Yang, J. Shi, D. Lin, Adapting object detectors via selective cross-domain alignment, in: CVPR, 2019, pp. 687–696. .
[72] C.-D. Xu, X.-R. Zhao, X. Jin, X.-S. Wei, Exploring categorical regularization for domain adaptive object detection, in: CVPR, 2020, pp. 11724–11733. .
[73] Z. He, L. Zhang, Domain adaptive object detection via asymmetric tri-way faster-rcnn, in: ECCV, 2020, pp. 309–324. .
[74] T. Kim, M. Jeong, S. Kim, S. Choi, C. Kim, Diversify and match: A domain adaptive representation learning paradigm for object detection, in: CVPR, 2019, pp. 12456–12465. .
[75] S. Kim, J. Choi, T. Kim, C. Kim, Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection, in: ICCV, 2019, pp. 6092–6101. .
[76] I. Hasan, S. Liao, J. Li, S.U. Akram, L. Shao, Generalizable pedestrian detection: The elephant in the room, in: CVPR, 2021, pp. 11328–11337..
[77] L. Liu, W. Lin, L. Wu, Y. Yu, M.Y. Yang, Unsupervised deep domain adaptation for pedestrian detection, in: ECCV, 2016, pp. 676–691..
[78] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778. .
[79] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: ICCV, 2017, pp. 2980–2988..
[80] C. Chi, S. Zhang, J. Xing, Z. Lei, S.Z. Li, X. Zou, Selective refinement network for high performance face detection, in: AAAI, 2019, pp. 8231–8238..
[81] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps. .
[82] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2017. .
[83] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, R. Urtasun, 3d object proposals for accurate object class detection, in: NIPS, 2015, pp. 424–432. .
[84] e. Chollet, François, Keras, url:https://github.com/fchollet/keras (2015). .
[85] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255. .
[86] Z. Cai, M. Saberian, N. Vasconcelos, Learning complexity-aware cascades for deep pedestrian detection, in: ICCV, 2015, pp. 3361–3369..
[87] J. Cao, Y. Pang, X. Li, Learning multilayer channel features for pedestrian detection, IEEE Transactions on Image Processing 26 (7) (2017) 3210–3220.
[88] Z. Cai, Q. Fan, R. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: ECCV, 2016, pp. 354–370. .
[89] N. Inoue, R. Furuta, T. Yamasaki, K. Aizawa, Cross-domain weakly-supervised object detection through progressive domain adaptation, in: CVPR, 2018, pp. 5001–5009. .

Xiaolin Song is currently pursuing the Ph.D. degree in the School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), China. Her research interests include computer vision and deep learning.



Binghui Chen received the B.E. degree in telecommunication engineering and the Ph.D degree in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT) in 2015, and 2020, respectively. His research interests include computer vision, deep learning, face recognition, deep embedding learning and machine learning.

Pengyu Li received the Master's degree from the Beijing University of the Posts and Telecommunications, Beijing, China, in 2015. His research interests include face recognition, crowd counting, object detection, knowledge distilling, unsupervised/semi-supervised learning, and deep learning with limited computational resources.

Biao Wang received the Ph.D. degree in the Department of Electronic Engineering from Tsinghua University, Beijing, China, in 2013. He has published over 20 papers in top tier academic conferences and journals. His current research interest include image classification, object detection, action recognition and unsupervised learning.

Honggang Zhang (Senior Member, IEEE) received the B. S. degree from the Department of Electrical Engineering, Shandong University, in 1996, and the master's and Ph. D. degrees from the School of Information Engineering, Beijing University of Posts and Telecommunications (BUPT), in 1999 and 2003, respectively. He worked as a Visiting Scholar with the School of Computer Science, Carnegie Mellon University (CMU), from 2007 to 2008. He is currently an Associate Professor and the Director of the Web Search Center, BUPT. He has published more than 130 articles on IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), Science, CVPR, ICCV, and NeurIPS. His research interests include image retrieval, computer vision, and pattern recognition.