Visual Intelligence

RESEARCH Open Access

Towards more reliable evaluation in pedestrian detection by rethinking "ignore regions"

Gang Li[1,2], Xiang Li[3], Shanshan Zhang[1,2*] and Jian Yang[1,2]

Abstract

It remains a challenging task to detect pedestrians in crowds and it needs more efforts to understand why the detectors fail. When we perform an error analysis based on the traditional evaluation strategy, we find that it produces many misleading false positives, which in fact cover occluded pedestrians. The reason for this is that we usually have two kinds of annotations in the dataset: regular pedestrians (detection targets) labeled by full-body boxes and ignored pedestrians (NOT detection targets) labeled by visible boxes. Ignored pedestrians are labeled as an additional category termed the "ignore region". Nevertheless, our detectors always predict a full-body box for each pedestrian. This gap results in the following case: when a detector successfully predicts a full-body box for those ignored pedestrians, a false positive is triggered due to the low overlap between the predicted full-body box and the labeled visible box for the ignored pedestrian. This becomes even more harmful as the detector improves and becomes more capable of locating occluded pedestrians. To alleviate this issue, we devise a new pedestrian detection pipeline, which considers the additional visible box at both the detection and evaluation stages. During detection, we predict an extra visible box apart from the full-body box for every instance; during evaluation, we employ visible boxes instead of full-body boxes to match the "ignore region". We apply the new pipeline to dozens of detection methods and validate the effectiveness of our pipeline in reducing the over-reporting of false positives and providing more reliable evaluation results.

Keywords: Pedestrian detection, Object detection, Occlusion handling, Evaluation metric

1 Introduction

Pedestrain detection has attracted much attention in the computer vision field, as it has various real-world applications such as autonomous driving, robotics and video surveillance. In addition, pedestrian detection also serves as a critical component for downstream tasks (e.g., pose estimation [1, 2], person re-identification [3, 4] and human parsing [5]), which take detected person boxes as inputs. Most recent works focus on detecting persons in complex scenes (e.g., crowded scenes with heavy occlusion), as detecting pedestrians has achieved satisfactory performance in simple scenes. When we perform error analysis on complex environments to understand why the detector fails, we find the existing evaluation strategy produces many misleading false positives, and hardly provides reliable evaluation results. Some examples are presented in Fig. 1, which are predicted by faster region-based convolutional neural networks (Faster R-CNNs [6]) on CrowdHuman [7] validation set. These false positives in fact successfully detect occluded persons. However, these persons are labeled as ignored in datasets due to heavy occlusion, and the existing evaluation process fails to match the detection

*Correspondence: shanshan.zhang@njust.edu.cn
[1] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2] Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
Full list of author information is available at the end of the article

Springer

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

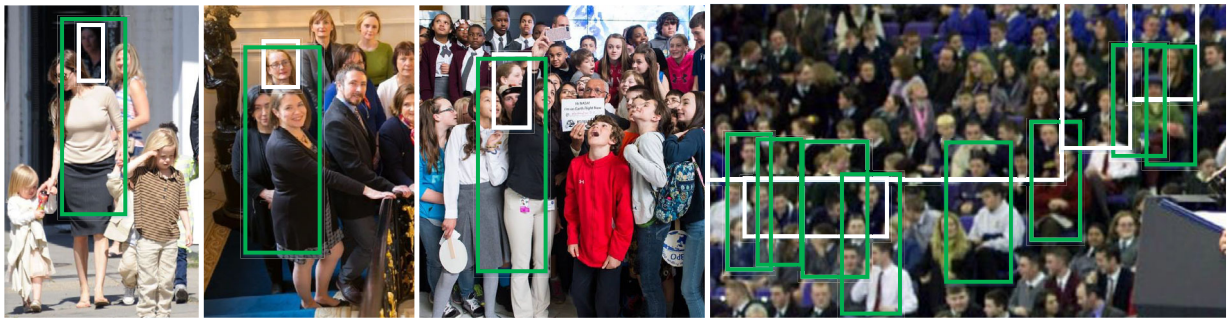


Figure 1 Visualization of full-body pedestrian detection results. White boxes refer to ignore regions, and green boxes refer to predicted full-body bounding boxes. The conventional benchmarking process usually treats these predictions as false positives since they fail to have high overlap ratios with the ground-truth or ignore regions. However, these predictions are considerably reasonable as they may perfectly match the underlying occluded persons. As a result, the existing benchmark pipeline provides imprecise evaluations of detector performance, which poses an obstacle to analyzing detection errors effectively (e.g., the above false positives are essentially not errors). This paper attempts to alleviate this problem. The presented examples are taken from the CrowdHuman dataset [7]

boxes with the ignored persons. In the following text, we will discuss this issue in detail.

On the popular pedestrian detection benchmarks (e.g., CityPersons [8] and CrowdHuman [7]), apart from the regular person category, there is an additional category termed the "ignore region", which is designed to alleviate the noisy annotation issue, incurred by the corner cases. Exactly, the real-world pedestrian detection system inevitably encounters some corner cases, i.e., some hard person examples, which are under the extreme conditions (e.g., occlusion, shadow and blur) with their boundaries unidentified. Given that labeling full-body boxes for these corner cases is usually infeasible and will introduce much uncertainty into the annotations, a common practice is to simply annotate "ignore regions" that can basically cover their visible regions, instead of labelling the invisible full-body boxes inaccurately. It is worth noting that there is no need for the pedestrian detection system to detect the ignore regions; therefore, detection boxes targeting ignore regions should be ignored, other than counted as the false/true positives during evaluation [7–9]. We visualize some ignore region examples in Fig. 1. During evaluation, each detection box is first matched with ground-truth boxes, and if matched, these detection boxes are counted as true positives. Then, the left unmatched boxes are further matched with ignore regions, and will be ignored if they have high overlaps with any ignore region. Only boxes with low overlaps with both ground-truth and ignore regions are counted as false positives [9].

Nevertheless, there is an inherent conflict between detection boxes and ignore regions. Specifically, detection boxes are always predicted to cover the full body; however, ignore regions are labeled to cover only visible parts. This gap between detection outputs (full-body boxes) and ignore regions (visible boxes) results in many misleading false positives as we have a large number of ignore regions in datasets. We take CrowdHuman dataset [7] as an example. There are in total 438, 822 box annotations, among which 99, 227 are ignore regions. These account for a non-trivial proportion (22.6%). Although these false positives actually cover the full body of ignored persons, they are hard to match the ignore regions during evaluation on account of low overlaps. Misleading false positives degrade evaluation results on one hand, and on the other hand, these false failures make up a substantial portion of the visualized errors during the error analysis process, while some true failures are obscured, leading to an ineffective error analysis.

To investigate how the unreasonable matching strategy between detection boxes and ignore regions quantitatively affects the evaluation performance, we randomly select 200 images from the CrowdHuman validation set, and manually correct evaluation errors on these 200 images. Specifically, we revisit every false positive box, and manually check whether the box overlaps with ignore regions. We remove those false positives that truly cover a person and fix the evaluation results. The average number of irrational matching in these images is 0.42 per image, and the revised evaluation results are reported in Table 1. We can draw some intriguing conclusions: 1) The aforementioned problem is severe and greatly affects the

Table 1 Evaluation results before and after manual revision. The numbers represent miss rates, and lower values indicate better performance. The bold numbers in brackets denote the improvements resulting from manual revision

Methods	Original	Revised	Δ*Rank*
ATSS	46.24	40.77 (**+5.47**)	1 ↑
GFLv1	45.55	41.71 (**+3.84**)	1 ↓
Faster R-CNN	42.58	39.02 (**+3.56**)	0

evaluation results. After revision, all methods achieve dramatic improvements, at least more than 3% gains in miss rate. 2) The traditional evaluation results fail to properly rank different methods. Original performance of ATSS [10] is weaker than that of GFLv1 [11]. However, after revision, ATSS outperforms GFLv1 by 0.94%. Moreover, the performance gap between one-stage detectors (ATSS and GLFv1) and the two-stage detector Faster R-CNN [6] is also narrowed from 3.66% to 2.69%.

To address this dilemma, we devise a new pedestrian detection pipeline for more reliable and accurate evaluation, which can be applied to any detector with simple modifications. Here, we define the pipeline as the combination of detection and evaluation. Considering the unreasonable matching strategy between detection boxes and ignore regions, we propose a method to predict an extra visible box for every full-body box during detection. Then, we use the visible box to match the ignore regions during evaluation. Extensive experiments are conducted to validate that the new pedestrian detection pipeline can alleviate unreasonable matching and produce more reliable evaluation results, thus benefiting further analysis.

As the number of ignore region increases in complex conditions, irrational matching between detection boxes and ignore regions happens more frequently. Irrational matching is barely considered, although detecting pedestrian in complex conditions is widely investigated in previous works [12–17]. By combining visual information with additional pose information, PEN [12] alleviates occlusion and false positive problems. MDFL [13] introduces a multi-scale network and fine-grained attention mechanism to address occlusion and small-scale issues. Wang et.al. [16] observed that aggregating the original RGB images with segmentation masks is beneficial for hard examples. These works aim to improve the detection performance for complex conditions. Different from them, we observe that the irrational matching between detection boxes and ignore regions leads to unreliable evaluation. To address this problem, we introduce a new pedestrian detection paradigm.

We summarize our contributions as follows:

1) To our best knowledge, we are the first to rethink the rationality of evaluation strategy for pedestrian detection, and find that the irrational matching between detection boxes and ignore regions results in many mistaken judgements and produces misleading evaluation results.
2) To address this problem, we devise a new pedestrian detection pipeline, where an extra branch for visible box prediction is added to the detector and during evaluation the predicted visible box, instead of full-body box, is used to match the ignore regions.
3) We demonstrate that the new pipeline can be implemented on any existing detection algorithm including one-stage and multi-stage detectors, and validate that the new pipeline provides more reasonable evaluation results and makes the error analysis more informative.

2 Related work

2.1 Object detection and pedestrian detection

CNN-based object detection can be roughly divided into two categories: one-stage and two-stage detectors. The former one [10, 11, 18, 19] directly makes dense predictions on feature maps, while the latter one [6, 20–22] is equipped with a region proposal network to generate sparse proposals, which are then passed into the second stage (i.e., R-CNN stage) for further refinement. With the object detection method as the basic detection framework, pedestrian detection methods [12–14, 16, 23–26] usually design dedicated modules to address some specific detection issues (e.g., object occlusion, small-scale detection, and night-time detection). For example, CrowdDet [17] performs multiple predictions on each proposal to improve the recall in crowd situations. Wang et al. [16] introduced a discriminant surface in the score scale space to balance performance over various scales. Detection transformers [27–35] have recently achieved significant progress. Iter-DETR [36] applies detection transformers to pedestrian detection with progressive query refinement, which boosts detection performance in crowded scenes. As the scenarios with pedestrian detection deployed are usually complicated, many ignore regions are annotated to exclude extremely hard cases from the detection and evaluation processing. However, previous works rarely consider the irrational matching between detection boxes and ignore regions, which is the focus of this paper.

2.2 Visible box prediction

Predicting visible boxes has been widely studied in previous works, but they are only designed for improving performance under traditional evaluation pipeline. Bi-box [24] designs two-stream branches in the R-CNN stage. Apart from full-body estimation branch, a visible part estimation branch is employed to regress visible regions. It also designs a novel label assignment, considering both full-body and visible boxes. With the help of visible regions, Bi-box achieves promising performance on occluded pedestrians. Paired box model (PBM) [37] first predicts paired full-body and visible boxes at the region proposal network (RPN) stage, and then features from paired boxes are aggregated into R-CNN. At last, the PBM uses visible boxes to perform non-maximum suppression (NMS) and shows robustness for crowd scenes. Mask-guided attention [25] considers the visible box as the coarse mask supervision, and thus it performs segmentation on every proposal to generate spatial attention for visible regions. With spatial attention, noise from occluded regions can be effectively suppressed. Beta R-CNN

[26] models the visible box as a 2D beta distribution, and only with several predicted distribution statistics, precise masks for visible regions can be obtained. These methods estimate visible boxes either to suppress background noise, or to perform NMS with visible boxes. Their goal is to obtain stronger detection performance under the traditional evaluation pipeline. By comparison, our purpose of predicting visible boxes is to build a novel pedestrian detection pipeline and acquire reliable evaluation results.

3 Method

In this section, we first revisit the current pedestrian detection pipeline, and then we elaborate on the irrational evaluation process. At last, we propose the new pedestrian detection pipeline for more accurate evaluation.

3.1 Revisiting pedestrian detection pipeline

The pedestrian detection pipeline consists of two components: detection and evaluation. The detector takes images as input, and predicts sets of bounding boxes with classification scores to cover persons. In evaluation process, predictions can be divided into three categories through comparing predicted boxes with annotations: true positives, false positives and ignored boxes. According to different metric protocols (e.g., average precision and miss rate), the final metric numbers are obtained. The pipeline is illustrated in Fig. 2. We use the intersection over foreground (*IoF*) to measure the overlaps between the detection boxes and ignore regions in the evaluation mechanism, which is formulated as $IoF \doteq \frac{\text{area}(Box_{\text{dt}} \cap Box_{\text{ig}})}{\text{area}(Box_{\text{dt}})}$.

Detection Pedestrian detection takes generic object detection algorithms as the base, and achieves significant progress with the development of convolutional neural networks [38]. Different from generic object detection, pedestrian detection only targets at the person category, and predicts full-body boxes for every person regardless whether the person is occluded or non-occluded.

Evaluation Compared with generic object detection, an extra step is required in the evaluation procedure of pedestrian detection. Specifically, the detection boxes need to match against the ignore regions. Ignore regions only serve as auxiliary information for evaluation, instead of detection targets. They mainly cover some extremely hard cases, which are expected to be excluded from the evaluation. Thus, any detection box that matches the ignore region should be ignored. The conventional process of evaluation [9] can be divided into two stages: For each detection box $b_i \in B$, its IoU_{ij} with all ground-truth boxes ($\forall g_j \in GT$) are calculated ($IoU_{ij} = \frac{\text{area}(b_i \cap g_j)}{\text{area}(b_i \cup g_j)}$). If the maximum *IoU* satisfies

$$\max_j IoU_{ij} > \tau_1, \tag{1}$$

b_i will be counted as the true positive; otherwise, it will be passed into the next stage. The unmatched boxes are noted as the set B_-. At the second stage, we check whether unmatched boxes B_- detect ignored persons. Concretely, for each detection box $b_i \in B_-$, we calculate its *IoF* scores w.r.t all ignore regions: $\{IoF_{ij}\}_{j=1}^{N}$, where N refers to the number of ignore regions. IoF_{ij} can be calculated as

$$IoF_{ij} = \frac{\text{area}(b_i \cap r_j)}{\text{area}(b_i)}, \tag{2}$$

where r denotes the ignore region box. If the below condition satisfies

$$\max_{1 \le j \le N} IoF_{ij} > \tau_2, \tag{3}$$

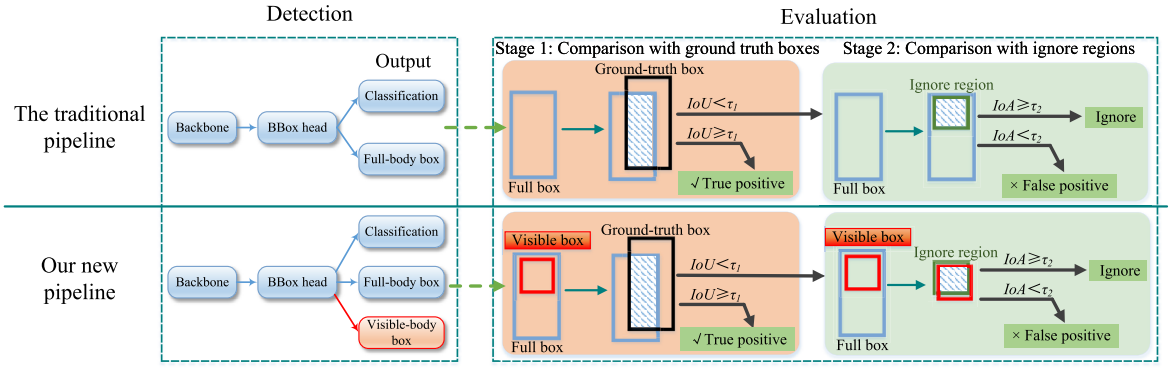


Figure 2 The overall pipeline of pedestrian detection including "detection" and "evaluation". The evaluation process takes the detection results as input. In detection, the new pipeline inserts a visible box prediction head in parallel to the full-body box prediction, so as to acquire two paired boxes for every instance. In evaluation, the new pipeline takes the paired box as input, and uses the visible box instead of the full-body box to match against the ignore regions. *IoU* refers to intersection over union

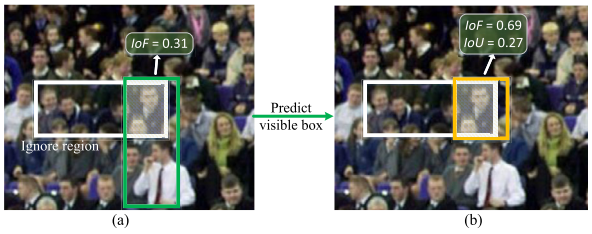


Figure 3 One specific example to indicate that our solution can handle the crowd scene through the additional visible box and *IoF*. White boxes refer to the ignore region, which contains a crowd; green (left) and orange (right) boxes refer to full-body and visible boxes, respectively

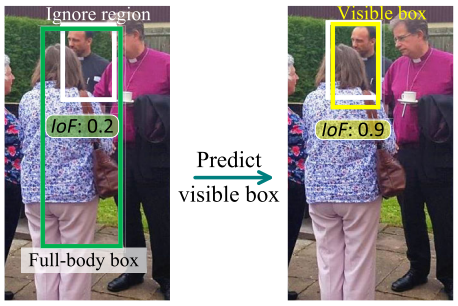


Figure 4 Due to a low *IoF* score (0.2), the detection box (green) fails to match the ignore region (white), while the predicted visible box (yellow) can perfectly match the ignore region

then b_i will be counted as ignored. Otherwise, it is the false positive. Therefore, it is only when detection boxes fail to match both ground truths and ignore regions, it is a false positive. The τ_1 and τ_2 are set to 0.5 by default.

Considering that ignore regions sometimes cover a crowd instead of a single person (e.g., the 4th figure in Fig. 1), *IoF* is intrinsically superior to *IoU* on matching detection boxes with ignore regions. One specific example is depicted in Fig. 3.

3.2 Problem in the current evaluation process

In pedestrian detection datasets, there are two kinds of annotations: regular persons and ignored persons. Because ignored persons lack enough information to obtain full-body boxes, they are labeled to only cover visible regions to reduce uncertainties. According to the existing evaluation method, detection boxes, which faithfully detect the ignored persons, are counted as ignored, rather than false positives.

However, we observe that an inherent contradiction between detection boxes and ignore regions is neglected in the literature. Specifically, detection boxes are predicted to cover the full-body, while ignore regions only cover visible parts. This contradiction results in their low *IoF* value and impedes their matching. As a result, though some prediction boxes successfully detect ignored persons, they are mistakenly counted as false positives due to their low *IoF* score.

We elaborate this problem with an example (Fig. 4). The green detection box is from the predictions of Faster R-CNN [6], and it almost matches the underlying occluded person. The white box is the ignore region, covering visible parts of the person. The *IoF* between two boxes is only 0.2, much smaller than the threshold τ_2 (i.e., 0.5), so the detection box is counted as false positive finally. This problem is severe in crowded scenes, and causes over-reporting of false positives. Furthermore, the miss rate is the most important evaluation metric in pedestrian detection, which is directly related to false positives. Over-reported false positives will unexpectedly damage the miss rate metric. In summary, results obtained from the current evaluation process hardly reflect the actual performance of detectors, which hinders the development of pedestrian detection and the effective analysis of errors.

3.3 Solution

Instead of modifying ignore region annotations to cover the full body, which will inevitably introduce uncertainties, we make a simple but effective modification on the detection framework to address the problem. As illustrated in Fig. 2, we also predict an extra visible box for every instance, apart from full-body boxes. Then, at the second stage of evaluation process, we calculate the intersection over foreground score between ignore regions with the predicted visible boxes, instead of full-body boxes. In Fig. 4, the visible box (in yellow) can precisely match the ignore region, so the corresponding detection box is successfully ignored. Fortunately, visible box annotations are commonly provided in pedestrian detection datasets (e.g. CrowdHuman [7] and CityPersons [8]).

Next, we will introduce how to predict visible boxes. The bounding box (BBox) head of conventional detectors can be formulated as

$$class = \mathcal{P}(F_{\text{cls}}), \qquad box = \mathcal{B}(F_{\text{reg}}), \tag{4}$$

where F_{cls} and F_{reg} indicate input features for classification and regression, respectively; $\mathcal{P}(\cdot)$ and $\mathcal{B}(\cdot)$ indicate the classifier and box regressor, respectively. We design an extra visible box regressor, consisting of only one fully-connected layer. We place it in parallel to the full-body box regressor, as depicted in Fig. 5. The new bbox head is formulated as

$$\begin{aligned} &class = \mathcal{P}(F_{\text{cls}}), \qquad box = \mathcal{B}(F_{\text{reg}}), \\ &box_{\text{vis}} = \mathcal{VB}(F_{\text{reg}}). \end{aligned} \tag{5}$$

Here, $\mathcal{VB}$ indicates the visible box regressor. The visible and full-body box prediction takes the same features and

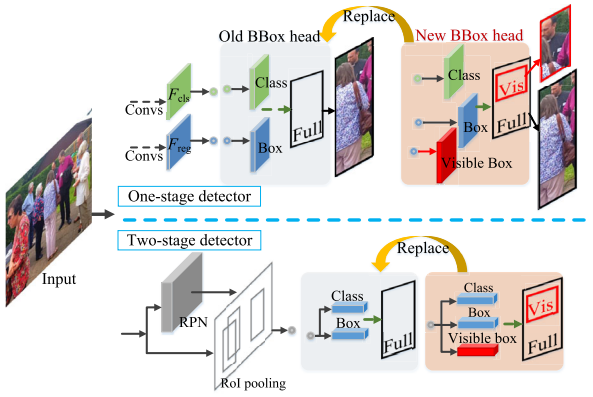


Figure 5 Comparison of detection structure in the old and new pipeline. We insert visible box prediction to BBox Head based on the original detection framework. For two-stage detectors, prediction is performed based on flattened RoI features and proposals. For one-stage detectors, we make prediction on densely pre-set anchors

reference boxes as input. The new BBox head can predict paired boxes (including full-body and visible boxes) for every instance. The proposed visible box prediction can be adapted into all existing detection frameworks (including one-stage and multi-stage) with simple modifications and negligible computational costs.

Discussion In pedestrian detection, it is expected to evaluate detection performance on regular persons and ignore detection boxes for ignore regions. For the paired boxes used in the new pipeline, the pair is considered a match for ignore regions and should be discarded as long as the visible box successfully matches the ignore regions.

4 Experiments

In this section, we conduct comprehensive experiments on CityPersons [8] and CrowdHuman [7] to validate that over-reported false positives widely appear in pedestrian detection benchmarks, and our new pipeline can effectively alleviate this issue. Moreover, we also make detailed error analyses on the old and new pipelines.

4.1 Datasets

On pedestrian detection benchmarks, ignore regions are elaborately annotated to alleviate issues incurred by the corner cases. Table 2 lists the number of persons and ignore regions in two datasets, where ignore regions take up a non-trivial proportion (i.e., 22.6% and 26.0%) in all ground truths for CrowdHuman and CityPersons, respectively. Therefore, matching detection boxes with ignore regions in a reasonable and credible manner is very important. We will provide more details about ignore regions in two datasets.

Table 2 Numbers of persons and ignore regions in different pedestrian detection datasets (training subset)

	CrowdHuman	CityPersons
#Images	15, 000	2, 975
#Persons	339, 565	19, 238
#Ignore regions	99, 227	6, 768

CityPersons CityPersons [8] groups all humans into four categories: pedestrian, rider, sitting person, and other persons (with unusual postures, e.g. stretching), based on postures. Except for the pedestrian category, the other three categories are treated as ignore regions. Different labeling principles are adopted for the regular pedestrian and ignore region. Specifically, the regular pedestrian is annotated by drawing a line from the top of the head to the middle of two feet, and then the bounding box is generated using a fixed aspect ratio (0.41). In contrast, ignore regions are labeled to cover visible parts of persons. [8].

CrowdHuman CrowdHuman [7] is collected in more complicated real-world scenarios with a focus on the crowd occlusion issue. Under such crowd situations, extremely hard cases appear more frequently. As these cases are usually heavily occluded, lacking necessary visual cues makes labeling the entire human body impossible. Most of these cases are treated as ignore regions with only visible parts labeled in CrowdHuman. Please refer to Fig. 1 for some visualized examples.

4.2 Evaluation metrics

We use two metrics to provide a comprehensive evaluation.

MR The log-average miss rate (MR) is computed by averaging miss rates over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$. It is sensitive to false positives, which makes it suitable to evaluate algorithms applied in real-world applications [9]. Particularly, false positives with high confidence will significantly damage the miss rate score.

AP Average precision (AP) averages the precision over different recall ratios. It is widely used in generic object detection tasks [39] for performance evaluation.

4.3 Implementation details

Our implementations are based on MMDetection [50]. For those without open-source codes, we re-implement them under MMDetection framework. Unless otherwise stated, we use ResNet-50 [51] with the feature pyramid network (FPN) [6] as the backbone network for all methods on CrowdHuman, while we use VGG-16 [52] as the backbone network on CityPersons. The training batch size is set to 4

on 2 GPUs. We train models for 20 epochs on CrowdHuman, with learning rate decay of 0.1 after epoch 16. Each training lasts for 12 epochs on CityPersons, with learning rate decay of 0.1 after epoch 9. The input images keep the original size [2048, 1024] on CityPersons for both training and inference. For input images of CrowdHuman, we resize them so that the short edge is 800 pixels while the long edge is smaller than 1400 pixels.

4.4 Comparisons of old and new pipelines

We propose a new pedestrian detection pipeline, which considers additional visible boxes at both detection and evaluation stages. We first run a dozen of detection algorithms on two datasets [7, 8] to obtain conventional evaluation results, and then we adopt our new pipeline to run them again to obtain new evaluation results. Experimental results are summarized in Table 3. Miss rate is considered a more critical metric than average precision for pedestrian detection, as pedestrian detection is widely used in autonomous driving and is highly sensitive to false positives. Therefore, we rank all methods based on miss rate. Rankings under the old and new pipelines (i.e., Rank (old), Rank (new)) are listed, respectively. $\Delta Rank$ refers to ranking changes for each method from the old pipeline to the new one. Comparisons between two datasets are described as follows.

CrowdHuman Compared with the old pipeline, all methods in our new pipeline achieve significant improvements on both average precision and miss rate. Especially on miss rate, the average improvement is 2.91%. These improvements validate that the new pipeline can effectively reduce over-reported false positives. We also observe that the new pipeline brings obvious changes to method rankings, and most methods have changed their ranking. For example, the ranking of Cascade R-CNN [20] improves from the 5th position to the 2nd position, surpassing Faster R-CNN [6]. With multi-stage refinements, Cascade R-CNN is supposed to have more accurate predictions than Faster R-CNN, which was validated on the COCO [39] benchmark. However, under the old pipeline, Cascade R-CNN is even three places behind Faster R-CNN. The change in method rankings indicates that the old pipeline fails to reflect the actual detection performance, thus designing a reasonable and credible pipeline is urgent. CrowdDet [17] and Iter-DETR [36] are two strong methods to demonstrate the superiority of our pipeline. It is interesting to find that the strong detectors tend to obtain larger performance changes under the new pipeline. Specifically, CrowdDet

Table 3 Detection results on the CrowdHuman and CityPersons validation sets. We compare the results in the old and new pipelines for every detector, and the rank is counted based on miss rate. Codes and trained models will be released

	Methods	Old pipeline		New pipeline		ΔMR	Rank (old)	Rank (new)	$\Delta Rank$
		Average precision	Miss rate	Average precision	Miss rate				
Crowd Human	RetinaNet [18]	86.47	52.95	87.31	51.88	+ 1.07	19	19	0
	FCOS [40]	86.67	52.15	87.52	49.78	+ 2.37	18	18	0
	PAA [41]	87.94	49.24	88.97	45.84	+ 3.40	17	16	1 ↑
	RepPoints [42]	87.95	48.26	88.69	46.97	+ 1.29	16	17	1 ↓
	Dynamic R-CNN [43]	84.72	47.71	85.42	44.09	+ 3.62	15	14	1 ↑
	VFNet [44]	88.39	47.56	89.23	44.66	+ 2.90	14	15	1 ↓
	Cascade RPN [21]	88.67	46.76	89.55	43.57	+ 3.19	13	12	1 ↑
	ATSS [10]	88.95	46.50	89.96	43.28	+ 3.22	12	11	1 ↑
	GFLv1 [11]	88.68	46.24	89.60	43.70	+ 2.54	11	13	**2** ↓
	GFLv2 [19]	89.01	45.88	89.94	42.79	+ 3.09	10	10	0
	R^2NMS [37]	89.43	45.14	90.27	42.20	+ 2.94	9	8	1 ↑
	Libra R-CNN [45]	87.37	45.07	88.07	42.55	+ 2.52	8	9	1 ↓
	Guided Anchoring [46]	88.97	44.79	89.83	41.41	+ 3.38	7	6	1 ↑
	Cascade R-CNN [20]	85.96	44.73	86.71	41.03	+ 3.70	6	4	**2** ↑
	Double Head [47]	86.18	44.50	86.92	41.50	+ 3.00	5	7	**2** ↓
	PISA [48]	87.52	44.12	88.30	40.76	+ 3.36	4	3	1 ↑
	Faster R-CNN [6]	88.20	44.01	88.95	41.07	+ 2.94	3	5	**2** ↓
	CrowdDet [17]	90.60	41.89	91.39	38.09	+ 3.80	2	2	0
	Iter-DETR [36]	92.14	41.52	92.61	37.58	+ 3.94	1	1	0
City Persons	R^2NMS [37]	74.94	41.58	75.54	40.04	+ 1.54	8	8	0
	A-FRCNN [8]	75.28	40.79	75.95	38.64	+ 2.15	7	6	1 ↑
	MGAN [25]	76.02	40.53	76.67	38.88	+ 1.65	6	7	1 ↓
	Cascade R-CNN [20]	74.03	40.48	74.61	37.72	+ 2.76	5	2	**3** ↑
	HTC [22]	76.88	39.97	77.58	37.75	+ 2.22	4	3	1 ↑
	Bi-box [24]	75.15	39.95	75.77	37.79	+ 2.16	3	4	1 ↓
	HGPD [23]	77.36	39.69	78.03	37.94	+ 1.75	2	5	**3** ↓
	CSP [49]	80.63	38.29	81.48	34.89	+ 3.40	1	1	0

and Iter-DETR achieve +3.80% miss rate and +3.94% miss rate, respectively, demonstrating that a reasonable pipeline helps to release the full capacity of detectors. It should be noted that, although the changes in the ranking are not so dramatic, our new pipeline provides fairer comparisons for different methods, and also offers more reasonable error analysis.

CityPersons For all experiments on CityPersons [8], we take the *All* subset (height>20 pixels, visibility>20%) as the training setting. Previous works [17, 26, 37] aim to address some specific detection problems (e.g., small object detection and object occlusion), and thus they adopt the corresponding subsets (*Small Scale* or *Heavy Occlusion*) to enlarge the problem. Different from them, we prefer to comprehensively evaluate two pipelines on *All* subset with more instances. From Table 3, the changes in the ranking are also obvious. Specifically, with the bottom-up detection paradigm, CSP [49] achieves state-of-the-art performance on CityPersons under the old pipeline. Our new pipeline further releases its capacity and obtains the greatest improvements among all methods (i.e., +3.40% miss rate). Moreover, multi-stage structure methods (i.e., Cascade R-CNN and HTC [22]) return to the top of the list, demonstrating that the new pipeline succeeds in providing more reasonable method rankings.

4.5 Results of visible box detection

Results of visible box detection are reported in Table 4. Apart from the miss rate of visible boxes (denoted as MR_{vis}), the miss rate of full-body boxes (denoted as MR_{full}) and the performance changes brought by the new pipeline (ΔMR) are also listed. From the results of Table 4, it is very clear that detectors with better MR_{full} can produce more precise visible box detection. Full-body and visible box detectors share the same backbone and most of network parameters, and thus their performance is strongly correlated. This conclusion also holds for two-stage detectors. For example, CrowdDet [17] surpasses the Faster R-CNN [6] by 2.12% in terms of MR_{full}, while a 3.38% performance gap is also found for MR_{vis}. The scatter diagram between the MR_{full} and MR_{vis} is plotted in Fig. 6 to better demonstrate their relations. Furthermore, we find that better MR_{vis} score contributes to larger ΔMR as presented in Fig. 7. This is easy to understand. Precise visible boxes can easily match the ignore regions. As a result, stronger detectors will obtain larger improvements in the new pipeline, which alleviates the underestimation of strong detectors in the old pipeline.

On the other hand, through comparing MR_{vis} and MR_{full}, we observe that MR_{vis} consistently lags behind MR_{full} by a large margin, indicating visible boxes are harder to detect than full-body boxes. The precision of visible box detection limits the effectiveness of our new pipeline, as some coarse visible boxes cannot match the ignore regions. However, visible box detection only adopts a very simple network in this paper (i.e., a single fully-connected layer), and it can be improved by designing delicate network structures and advanced training strategies. Once the

Table 4 Results of visible box detection. (denoted as MR_{vis}). Performance differences between the old and new pipelines (i.e., ΔMR) are also listed

	Methods	MR_{full}	MR_{vis}	ΔMR
One-Stage	RetinaNet [18]	52.95	70.71	+ 1.07
	FCOS [40]	52.15	70.34	+ 2.37
	RepPoints [42]	48.26	64.32	+ 1.29
	PAA [41]	49.24	61.86	+ 3.40
	VFNet [44]	47.56	61.71	+ 2.90
	ATSS [10]	46.50	60.30	+ 3.22
	GFLv1 [11]	46.24	60.15	+ 2.54
	GFLv2 [19]	45.88	59.87	+ 3.09
Two-Stage	Libra R-CNN [45]	45.07	60.92	+ 2.52
	Dynamic R-CNN [43]	47.71	59.11	+ 3.62
	Double Head [47]	44.50	59.07	+ 3.00
	Faster R-CNN [6]	44.01	58.96	+ 2.94
	PISA [48]	44.12	58.41	+ 3.36
	Guided Anchoring [46]	44.79	57.99	+ 3.38
	Cascade R-CNN [20]	47.71	56.99	+ 3.70
	CrowdDet [17]	41.89	55.58	+ 3.80

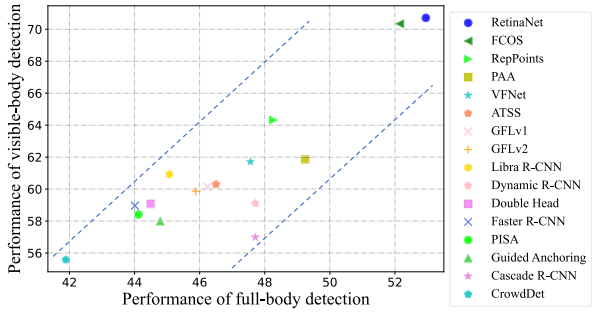


Figure 6 The scatter diagram of relations between MR_{full} and MR_{vis}. It shares the same legend with Fig. 7

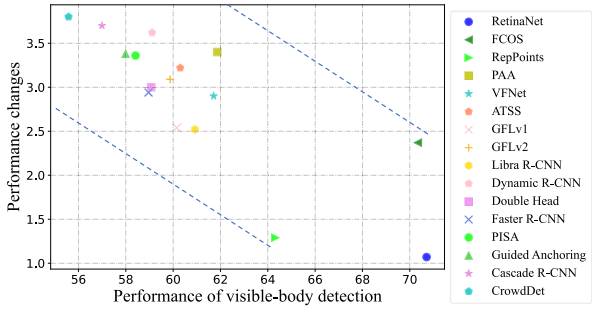


Figure 7 The scatter diagram of relations between MR_{vis} and ΔMR

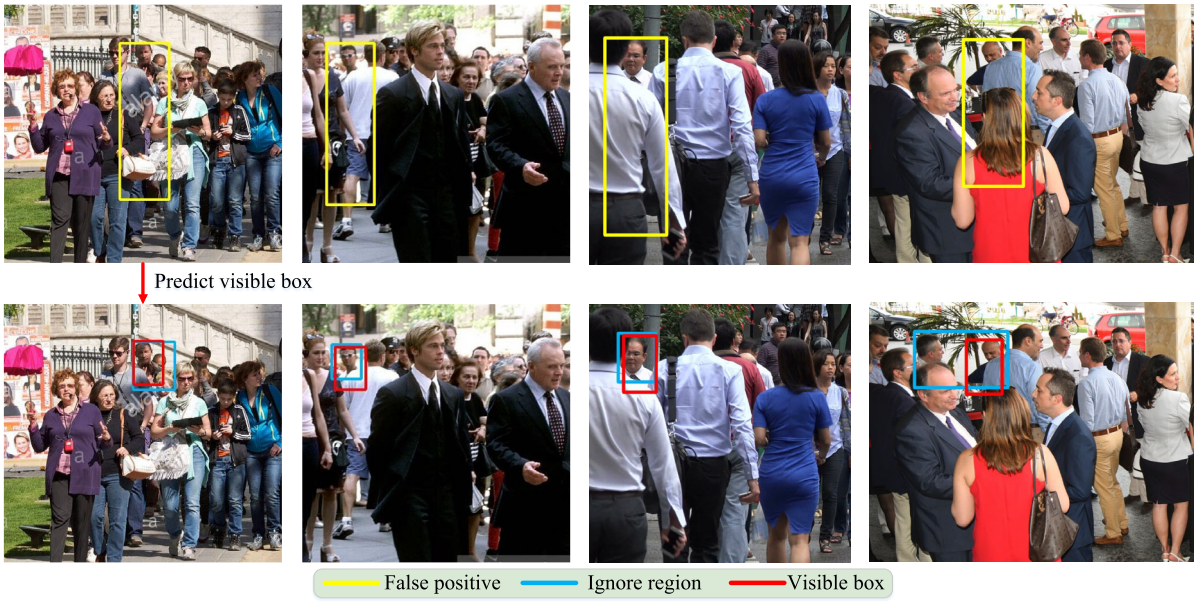


Figure 8 Visualization of predicted visible boxes. The 1st row: false positives in the old pipeline. The 2nd row: red boxes refer to visible boxes, which are predicted from the false positives in the 1st row; blue boxes refer to ignore regions. Images are collected from the CrowdHuman dataset [7]

precision of visible box detection achieves improvements, we believe that the effectiveness of our new pipeline will become more significant.

Some qualitative results of visible boxes are provided in Fig. 8. The predicted full-body boxes and visible boxes are displayed in the 1st and 2nd rows, respectively. The accurate predictions for these hard examples in crowds indicate the strong robustness of visible box detection. In addition, the visualized examples show that full-body boxes have difficulty matching with ignore regions, confirming the motivation of this paper. In contrast, visible boxes are a good match for the ignore regions.

4.6 Error analysis

Error analysis plays an important role in advancing deep learning methods. By visualizing false positives with high confidence, flaws and limitations of the detector can be observed, which will inspire further improvements for the detector. Here, we conduct error analysis under the old and new pipelines, respectively, and visualize false positives in Fig. 9. Under the old pipeline, some false positives are not necessarily serious (Fig. 9(a)), since they can faithfully detect occluded persons underneath. These detection boxes should be ignored, rather than treated as false positives. As a result, these *fake failures* overwhelm the error analysis of the old pipeline, which severely prevents the true errors from being observed. In contrast, the issue can be effectively alleviated in our new pipeline, where the number of over-reported false positives is significantly reduced.

In Fig. 9(b), some failure cases under the new pipeline are presented, which are randomly selected from the validation set. It is obvious that the emerging false positives are true failures of the detector. They can provide valuable error information and enable effective error analysis.

On the other hand, quantitative error analysis is also conducted (see Table 5). Based on the intersection over union score between false positives and ground truth boxes, we divide false positives into three categories: (a) $IoU > 0.5$: duplicate detections; (b) $0.3 < IoU \leq 0.5$: localization errors; (c) $IoU \leq 0.3$: classification errors. We depict the distribution over three kinds of errors in Table 5. We speculate that most false positives caused by ignore regions are counted as classification errors, since they detect ignore regions instead of ground truth boxes, resulting in low overlaps with ground truth boxes. The quantitative results support our assumption. Compared with the old pipeline, the proportion of classification errors in the new pipeline drops dramatically from 39.2% to 31.6%. Meanwhile, the proportion of the other two types of errors rises. The significant changes in error distributions validate that our new pipeline can offer more informative bad cases.

4.7 Ablation study

Study on visible box detection We add an extra prediction branch to perform visible box detection. For one-stage detectors, the visible box branch consists of one convolution layer with 9 k parameters; for two-stage detectors, one fully-connected layer with 4 k parameters is included.

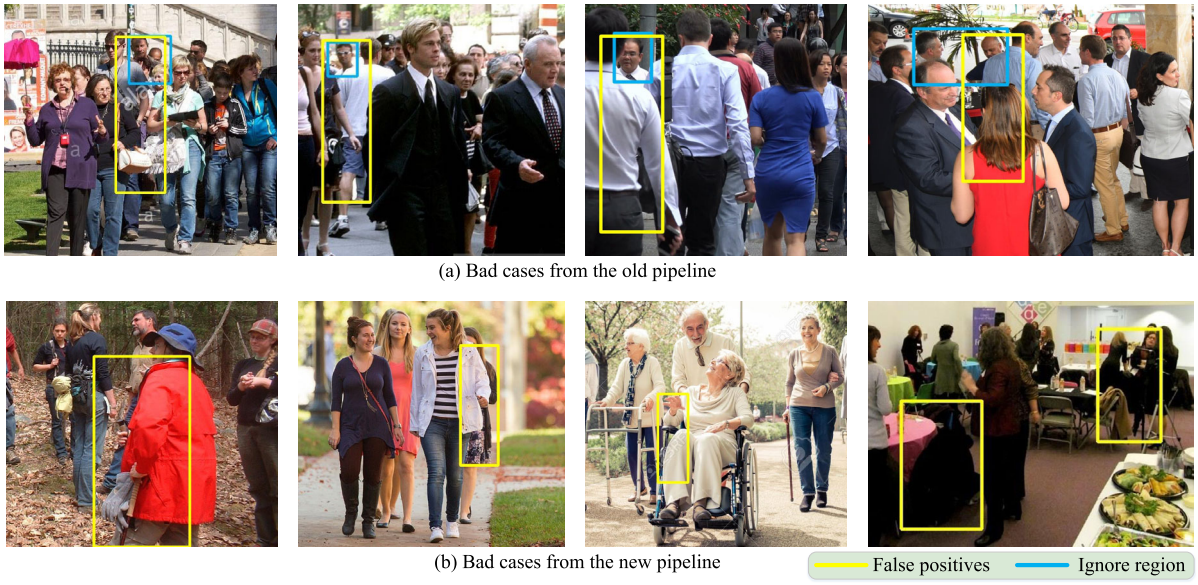


Figure 9 Qualitative error analysis in CrowdHuman validation. (a) Bad cases from the old pipeline are not necessarily serious. (b) The emerging false positives in the new pipeline reflect the true errors in the detectors, which are concealed in the old pipeline. All detection results are collected from Cascade R-CNN with the same false positive score per image

Table 5 The error distribution of Cascade R-CNN. Loc refers to localization and Cls refers to classification. The error distribution varies significantly between the old and the new pipeline. For example, the gap between the Loc error and the Cls error increases from 11.4% to 22.4%

Pipeline	Error distribution (%)		
	Duplicate	Loc error	Cls error
Old pipeline	10.2	50.6	39.2
New pipeline	14.4	54.0	31.6

Table 6 Effects of visible box prediction. "+ Visible box" denotes adding extra visible box prediction to the original detector. Lower miss rate score indicates better performance. Best results are marked in bold. Speed is measured in milliseconds to process each image

Methods	Miss rate (old)	Miss rate (new)	Speed
ATSS	47.02	–	70.3
+ Visible box	**46.50**	43.28	75.7 (+5.4)
Faster R-CNN	44.62	–	137.5
+ Visible box	**44.01**	41.07	139.4 (+1.9)

In Table 6, experiments are conducted to compare the detection results with and without visible box detection. We take ATSS [10] and Faster R-CNN [6] to represent one-stage and two-stage detectors, respectively. The results indicate visible box detection can provide a consistent gain of ∼0.5% miss rate score for both detectors. The improvements may benefit from auxiliary information provided by visible box annotations, which contributes to robust detection for occluded cases. Moreover, visible box detection only causes negligible time costs. Concretely, it slows down the inference speed by less than 7%. We also provide qualitative comparisons in Fig. 10, where we randomly sample detection examples from CrowdHuman *Val* set. It is clear that the detection results from the detectors with/without visible box detection are very consistent, indicating that visible box detection hardly changes the original detector. Therefore, adding visible box detection affects neither the performance nor inference speed of the original detector, which can be seamlessly integrated into most detectors.

Comparison between methods for visible box detection As discussed in Sect. 2.2, some previous methods also design dedicated modules to predict visible boxes. Bi-Box [24] is the first to employ visible box annotations as auxiliary information. It designs two-stream branches in R-CNN stage for visible and full-body boxes, respectively. A new label assignment strategy is also designed to consider visible boxes as well. PBM [37] generates visible boxes in RPN, and then it further aggregates features from full-body and visible boxes in R-CNN. At last, predicted visible boxes are employed in NMS. Our method is lighter and makes smaller modifications to the original detector, compared with these works. Comparison results are listed in Table 7. Because previous methods adopt Faster R-CNN [6] as the default detector, we also implement our method on Faster R-CNN. Our method not only obtains the best performance under both old and new pipelines but also achieves



(a) Original Faster R-CNN(the 1st row) *vs*. Faster R-CNN w/ visible box detection (the 2nd row)

(b) Original ATSS (the 1st row) *vs*. ATSS w/ visible box detection (the 2nd row)

Figure 10 Qualitative comparisons between the original detector and the visible box detector. The detectors with/without visible box detection show consistent detection biases, as indicated by the blue boxes

Table 7 Comparison of visible box detection methods. The best results are marked in bold. Speed is measured in milliseconds to process each image

Methods	Miss rate (old)	Miss rate (new)	Speed
PBM [37]	45.14	42.20	142.8
Bi-box [24]	44.63	41.39	140.8
Ours	**44.01**	**41.07**	**139.4**

the fastest inference speed thanks to the lightweight structure, as the results indicate.

Alignment between full-body and visible boxes For each proposal box, the new pipeline adds an extra prediction branch to predict visible boxes, instead of predicting one full-body detection, as done in previous work. Here we evaluate their alignment. We employ the intersection over union scores between full-body and visible boxes as the indicator to measure their alignment. For each visible box, its *IoU* scores with the corresponding full-body box should be larger than the *IoU* scores with other full-body boxes. Otherwise, it is a failed alignment. In Table 8, we report the ratio of successful alignment (RSA) over all visible box detections on CrowdHuman *Val* set. RSA is greater than 95% regardless of one-stage or two-stage detectors, even though full-body and visible boxes are predicted by two separate branches, indicating that the visible boxes can be aligned well with full-body boxes. It is noteworthy that RetinaNet is around 2% lower than Faster R-CNN in terms of RSA, which may account for limited performance improvements of RetinaNet (i.e., +1.07%) in the new pipeline. In future work, we will develop specific methods to improve their alignment.

Table 8 Alignment between the full-body and visible boxes. RSA denotes the ratio of successful alignment

Method	RSA	Method	RSA
RetinaNet [18]	95.38%	Faster R-CNN [6]	97.50%

4.8 Discussions

Visible boxes vary greatly in shape and appearance, which increases the detection difficulty and affects their performance. Even so, the new pipeline still brings considerable

performance changes to most detection methods and enables fairer comparisons for them. In the old pipeline, the actual detection performance is concealed. Consequently, some advanced detectors even lag behind some old ones, which seriously affects the development of pedestrian detection. In contrast, the new pipeline can unleash the capacity of strong detectors and provide credible method comparisons. Moreover, the new pipeline also enables more reasonable error analysis. By removing some misleading detection errors incurred by ignore regions, more informative errors are exposed, helping to discover the real bottleneck and pave the way for pedestrian detection research. Some advanced network structures with strong modelling capabilities (e.g., transformer) have recently been introduced, which can be employed for visible box detection replacing the vanilla fully-connected layer used in this paper. With the precision improvement of visible boxes, the effectiveness of our new pipeline can be further enhanced. We leave it for future work.

5 Conclusion

In this paper, we point out the conventional evaluation process in pedestrian detection has serious limitations, which leads to unreliable evaluation results and impedes error analysis. We devise a more reasonable pipeline, where an extra branch is added to the detectors to predict visible boxes for each person, and during evaluation, the predicted visible boxes are utilized to match the ignore regions. The new pipeline is applied to dozens of detection methods and comprehensive benchmark results are provided on two widely used datasets. We also demonstrate that our pipeline can benefit effective error analysis. We hope the proposed pipeline can serve as a new paradigm for pedestrian detection.

Funding

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62322602), Natural Science Foundation of Jiangsu Province, China (Grant No. BK20230033), National Natural Science Foundation of China (Grant No. 62172225), and CAAI-Huawei MindSpore Open Fund.

Abbreviations

FPPI, false positive per image; IoF, intersection over foreground; IoU, intersection over union; MR, log-average miss rate; RSA, the ratio of successful alignment.

Data availability

All data generated or analyzed during this study are included in published articles [7, 8].

Declarations

Competing interests

The authors declare no competing interests.

Author contributions

Material preparation, and data collection and analysis were performed by GL, XL and SZ. The first draft of the manuscript was written by GL and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Author details

[1]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. [2]Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. [3]VCIP, Nankai University, Tianjing, China.

Received: 2 July 2023 Revised: 16 January 2024 Accepted: 17 January 2024
Published online: 22 February 2024

References

1. Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5693–5703). Piscataway: IEEE.
2. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.-S., & Crowdpose, C. Lu. (2019). Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10863–10872). Piscataway: IEEE.
3. Shen, Y., Li, H., Yi, S., Chen, D., & Wang, X. (2018). Person re-identification with deep similarity-guided graph neural network. In V. Ferrari, M. Hebert, C. Sminchisescu, et al. (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 486–504). Cham: Springer.
4. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., et al. (2019). Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8514–8522). Piscataway: IEEE.
5. Gong, K., Liang, X., Zhang, D., Shen, X., & Lin, L. (2017). Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 932–940). Piscataway: IEEE.
6. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125). Piscataway: IEEE.
7. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., et al. (2018). Crowdhuman: a benchmark for detecting human in a crowd. arXiv preprint. arXiv:1805.00123.
8. Zhang, S., Benenson, R., & Citypersons, B. S. (2017). A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3221). Piscataway: IEEE.
9. Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: an evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.
10. Zhang, S., Chi, C., Yao, Y., Lei, Z., & Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9759–9768). Piscataway: IEEE.
11. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., et al. (2020). Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. In H. Larochelle, M. Ranzato, R. Hadsell, et al. (Eds.), *Proceedings of the 34th international conference on neural information processing systems* (pp. 21002–21012). Red Hook: Curran Associates.
12. Jiao, Y., Yao, H., & Xu, C. (2021). PEN: pose-embedding network for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(3), 1150–1162.
13. Lin, C., Lu, J., & Zhou, J. (2018). Multi-grained deep feature learning for robust pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(12), 3608–3621.
14. Wang, X., Liang, C., Chen, C., Chen, J., Wang, Z., Han, Z., et al. (2019). S^3d: scalable pedestrian detection via score scale surface discrimination. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(10), 3332–3344.
15. Li, G., Zhang, S., & Yang, J. (2021). Nighttime pedestrian detection based on feature attention and transformation. In *Proceedings of the 20th international conference on pattern recognition* (pp. 9180–9187). Piscataway: IEEE.
16. Wang, X., Shen, C., Li, H., & Xu, S. (2019). Human detection aided by deeply learned semantic masks. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(8), 2663–2673.

17. Chu, X., Zheng, A., Zhang, X., & Sun, J. (2020). Detection in crowded scenes: one proposal, multiple predictions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12214–12223). Piscataway: IEEE.
18. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988). Piscataway: IEEE.
19. Li, X., Wang, W., Hu, X., Li, J., Tang, J., & Yang, J. (2021). Generalized focal loss v2: learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11632–11641). Piscataway: IEEE.
20. Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: delving into high quality object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6154–6162). Piscataway: IEEE.
21. Vu, T., Jang, H., Pham, T. X., & Yoo, C. D. (2019). Cascade RPN: delving into high-quality region proposal network with adaptive convolution. arXiv preprint. arXiv:1909.06720.
22. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., et al. (2019). Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4974–4983). Piscataway: IEEE.
23. Li, G., Li, J., Zhang, S., & Yang, J. (2020). Learning hierarchical graph for occluded pedestrian detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1597–1605). New York: ACM.
24. Zhou, C., & Yuan, J. (2018). Bi-box regression for pedestrian detection and occlusion estimation. In V. Ferrari, M. Hebert, C. Sminchisescu, et al. (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 135–151). Cham: Springer.
25. Pang, Y., Xie, J., Khan, M.H., Anwer, R.M., Khan, F.S., & Shao, L. (2019). Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4967–4975). Piscataway: IEEE.
26. Xu, Z., Li, B., Yuan, Y., & Dang, A. (2020). Beta R-CNN: looking into pedestrian detection from another perspective. In H. Larochelle, M. Ranzato, R. Hadsell, et al. (Eds.), *Proceedings of the 34th international conference on neural information processing systems*. Red Hook: Curran Associates.
27. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, et al. (Eds.), *Proceedings of the 16th European conference on computer vision* (pp. 213–229). Cham: Springer.
28. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint. arXiv:2010.04159.
29. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., et al. (2022). Dino: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv preprint. arXiv:2203.03605.
30. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., et al. (2021). Conditional DETR for fast training convergence. In *Proceedings of the IEEE/CVF computer vision and pattern recognition* (pp. 3651–3660). Piscataway: IEEE.
31. Zhang, S., Wang, X., Wang, J., Pang, J., Lyu, C., Zhang, W., et al. (2023). Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7329–7338). Piscataway: IEEE.
32. Zhang, J., Lin, X., Zhang, W., Wang, K., Tan, X., Han, J., et al. (2023). Semi-DETR: semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23809–23818). Piscataway: IEEE.
33. Wang, Y., Zhang, X., Yang, T., & Sun, J. (2022). Anchor DETR: query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2567–2575). Palo Alto: AAAI Press.
34. Cao, X., Yuan, P., Feng, B., & Niu, K. (2022). CF-DETR: coarse-to-fine transformers for end-to-end object detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 185–193). Palo Alto: AAAI Press.
35. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M., & Zhang, L. (2022). DN-DETR: accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13619–13627). Piscataway: IEEE.
36. Zheng, A., Zhang, Y., Zhang, X., Qi, X., & Sun, J. (2022). Progressive end-to-end object detection in crowded scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 857–866). Piscataway: IEEE.
37. Huang, X., Ge, Z., Jie, Z., & Yoshie, O. (2020). NMS by representative region: towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10750–10759). Piscataway: IEEE.
38. Zhang, S., Benenson, R., Omran, M., Hosang, J., & Schiele, B. (2016). How far are we from solving pedestrian detection? In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1259–1267). Piscataway: IEEE.
39. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: common objects in context. In D. J. Fleet, T. Pajdla, B. Schiele, et al. (Eds.), *Proceedings of the 13th European conference on computer vision* (pp. 740–755). Cham: Springer.
40. Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627–9636). Piscataway: IEEE.
41. Kang, K., & Lee, H. S. (2020). Probabilistic anchor assignment with IoU prediction for object detection. arXiv preprint. arXiv:2007.08103.
42. Yang, Z., Liu, S., Hu, H., Wang, L., & Reppoints, S. L. (2019). Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9657–9666). Piscataway: IEEE.
43. Zhang, H., Chang, H., Ma, B., Wang, N., & Chen, X. (2020). Dynamic R-CNN: towards high quality object detection via dynamic training. In A. Vedaldi, H. Bischof, T. Brox, et al. (Eds.), *Proceedings of the 16th European conference on computer vision* (pp. 260–275). Cham: Springer.
44. Zhang, H., Wang, Y., Dayoub, F., & Varifocalnet, N. S. (2020). An iou-aware dense object detector. arXiv preprint. arXiv:2008.13367.
45. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019). Libra R-CNN: towards balanced learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 821–830). Piscataway: IEEE.
46. Wang, J., Chen, K., Yang, S., Loy, C. C., & Lin, D. (2019). Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2965–2974). Piscataway: IEEE.
47. Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., et al. (2020). Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10186–10195). Piscataway: IEEE.
48. Cao, Y., Chen, K., Loy, C. C., & Lin, D. (2020). Prime sample attention in object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11583–11591). Piscataway: IEEE.
49. Liu, W., Liao, S., Ren, W., Hu, W., & Yu, Y. (2019). High-level semantic feature detection: a new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5187–5196). Piscataway: IEEE.
50. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al. (2019). MMDetection: open MMLab detection toolbox and benchmark. arXiv preprint. arXiv:1906.07155.
51. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Piscataway: IEEE.
52. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.