

# Detection in Crowded Scenes: One Proposal, Multiple Predictions

Xuangeng Chu<sup>1\*</sup>, Anlin Zheng<sup>2\*</sup>, Xiangyu Zhang<sup>2\*†</sup>, Jian Sun<sup>2</sup>

<sup>1</sup>Peking University <sup>2</sup>MEGVII Technology

xg\_chu@pku.edu.cn, {zhenganlin, zhangxiangyu, sunjian}@megvii.com

## Abstract

We propose a simple yet effective proposal-based object detector, aiming at detecting highly-overlapped instances in crowded scenes. The key of our approach is to let each proposal predict a set of correlated instances rather than a single one in previous proposal-based frameworks. Equipped with new techniques such as EMD Loss and Set NMS, our detector can effectively handle the difficulty of detecting highly overlapped objects. On a FPN-Res50 baseline, our detector can obtain 4.9% AP gains on challenging CrowdHuman dataset and 1.0%  $MR^{-2}$  improvements on CityPersons dataset, without bells and whistles. Moreover, on less crowded datasets like COCO, our approach can still achieve moderate improvement, suggesting the proposed method is robust to crowdedness. Code and pre-trained models will be released at <https://github.com/megvii-model/CrowdDetection>.

## 1. Introduction

Proposal-based framework has been widely used in modern object detection systems [29, 11, 12, 20, 21, 24, 14, 4, 5, 2, 13, 28, 34, 33], both for one-stage [24, 28, 21, 10] and two/multi-stage [29, 20, 4, 5, 13, 2] methods. The paradigm in general has a two-step pipeline: first, generating *over-complete* object proposals in handcraft (e.g. predefined anchors [28, 24, 21]) or learnable (e.g. RPNs [29, 20, 13]) manner; then, predicting a *single* instance corresponding to each proposal box with a confidence score and a refined location. To remove duplicate predictions, methods such as *Non-maximum Suppression (NMS)* are usually required for post-processing.

Although proposal-based approaches have achieved state-of-the-art performances [13, 2, 20, 21] in popular

\* Equal contribution. The work is done when Xiangyu Zhang was an intern in MEGVII Technology.

† Corresponding author.

This work is supported by The National Key Research and Development Program of China (2018YFC0831700) and Beijing Academy of Artificial Intelligence (BAAI).

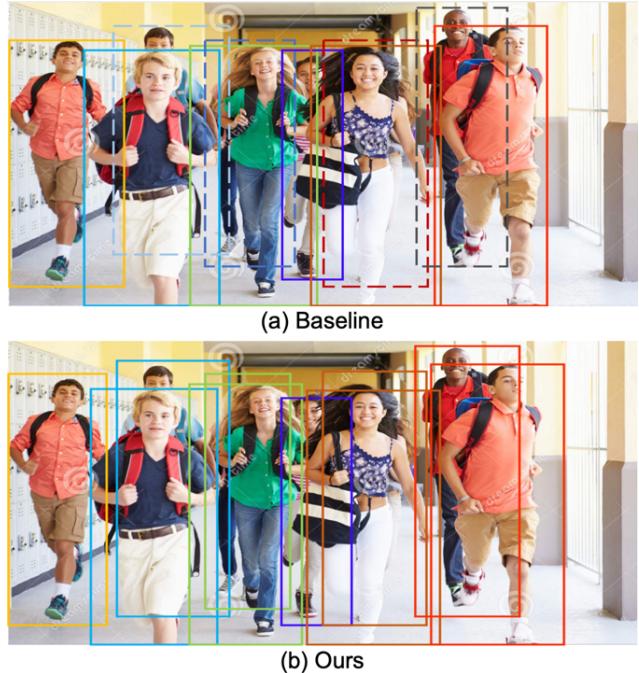


Figure 1. Human detection in crowds. (a) Results predicted by FPN [20] baseline. The dashed boxes indicate the missed detections (suppressed by NMS mistakenly). (b) Results of our method applied to FPN. All instances are correctly predicted. Boxes of the same color stem from the identical proposal (best viewed in color).

datasets such as *COCO* [22] and *PASCAL VOC* [8], it is still very challenging for crowded detection in practice. Fig. 1 (a) shows a common failure case: the detector fails to predict instances heavily overlapped with others (indicated in dashed box).

This kind of typical failure in crowded scenes is mainly ascribed to two reasons. First, highly overlapped instances (as well as their associated proposals) are likely to have very similar features. As a result, it is difficult for a detector to generate distinguishing prediction for each proposal respectively (illustration is shown in Fig. 2 for a concrete example). Second, since instances may heavily overlap each other, the

predictions are very likely to be mistakenly suppressed by *NMS*, as depicted in Fig. 1 (a).

Previous works have tried to address this issue from different perspectives, such as sophisticated NMS [1, 16, 23, 18, 25, 17], new loss functions [44, 40], re-scoring [19], part-based detectors [44, 38, 46, 3]. However, as we will analyze later (Sec. 2), current works are either too complex or less effective for handing highly-overlapped cases, or degrading the performance of less-overlapped cases.

In this paper, a new scheme is introduced to handle this difficulty: for each proposal box, instead of predicting a single instance, as usual, we suggest predicting *a set of instances* that might be highly overlapped, as described in Fig. 2. With this scheme, the predictions of nearby proposals are expected to infer the *same set of instances*, rather than *distinguishing individuals*, which is much easier to be learned. We also introduce several techniques in the new scheme. Firstly, a *EMD loss* is proposed to supervise the learning of instance set prediction. Secondly, a new post-processing method named *Set NMS* is introduced to suppress the duplicates from different proposals, which aims at overcoming the drawbacks of naïve NMS in crowd scenes. Lastly, an optional *refinement module* (RM) is designed to handle the potential false positives.

Our method is simple and almost cost-free. It is applicable to all the proposal-based detectors such as [29, 20, 21, 13]. The major modification is adding a prediction branch, which only brings negligible extra cost. But the improvement is significant: on *CrowdHuman* [32] dataset, our method boosts the AP score by **4.5%** (without refinement module) or **4.9%** (with refinement module); in addition, the recall on crowded instances improves by **8.9%**. More importantly, fewer false positive appears, suggested by the slightly improved  $MR^{-2}$  index even without refinement module. Besides, on less crowded datasets, our method can still obtain moderate gains. For example, on *CityPersons* we achieve 0.9% and 1.0% improvements in AP and  $MR^{-2}$  over the baseline; and on *COCO*, [22] it obtains 1.0% higher AP score. All experiments conducted on different datasets illustrate that our method can handle all scenes gracefully, regardless of the crowdedness.

## 2. Background

As mentioned in the introduction, the paradigm of proposal-based object detectors mainly involves two steps: the first step is *proposal box generation*, which could be achieved by *Selective Search* [12, 11], predefined/learnable anchors [29, 28, 24, 21, 41, 39, 45] or *Region Proposal Networks (RPNs)* [29, 20, 13, 4, 2], etc. The second step is *instance prediction*, i.e. predicting the refined detection results corresponding to each proposal box. We primarily focus on the second step in this paper.

For *instance prediction*, current state-of-the-art object

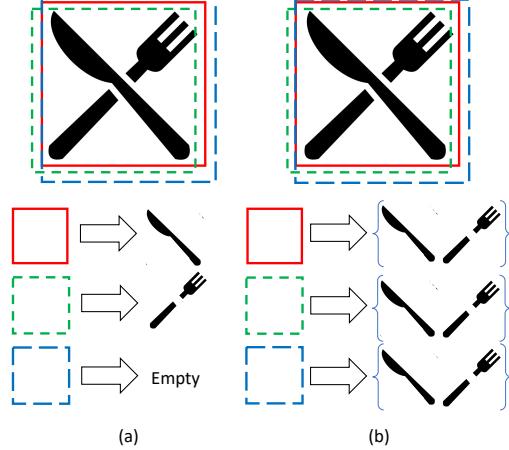


Figure 2. A typical case in crowded detection. A knife and a fork share almost the same bounding boxes. Three proposal boxes (red, green and blue, best viewed in color) are heavily overlapped. (a) Single predication paradigm (see Sec. 2). Each proposal box is expected to predict a *single instance* [29, 20, 21, 24, 28, 13] (may be empty), which is intrinsically difficult as the proposals share very similar feature. Moreover, after NMS, it is very likely that only one prediction survives. (b) In our approach, each proposal predicts *a set of instances*. Our *Set NMS* can easily remove the duplicate prediction sets together (not illustrated in the figure).

detection frameworks [29, 20, 21, 24, 28] usually attach a *detection function* to each proposal box, which is used to determine whether the proposal is associated to some ground truth instance; if true, further to predict the corresponding class label and the refined bounding box for the object. This mechanism implies that each proposal box corresponds to **single** ground truth (usually the one with the largest overlap to the proposal box). Therefore, the proposal boxes have to be over-completed to ensure every instance has a chance to be detected, which introduces many duplicates to the predictions. As a result, duplicate removal methods such as *Non-Maximum Suppression (NMS)* are necessary for those frameworks to filter out the duplicate results.

Although the above paradigm seems to obtain outstanding results on some benchmarks such as *COCO* [22] and *PASCAL VOC* [8]. It suffers from missing detection in crowded scenarios due to post-processing methods, e.g. NMS. Fig. 1 (a) shows an example: people in the dashed boxes are suppressed by the nearby boxes mistakenly. Thus, several approaches or workarounds have been proposed to address this limitation, which can be categorized as follows:

**Advanced NMS.** The effectiveness of naïve NMS is based on the assumption that multiple instances rarely occur at the same location, which is no longer satisfied in the crowded scenes. Several improved NMS approaches have been proposed. For example, *Soft-NMS* [1] and *Softer-NMS*

[16] suggest decaying the confidence score of the neighboring predictions for suppression rather than directly discard them. [30] employs Quadratic Binary Optimization to predict instances, taking advantage of the prior distribution of ground truth’s sizes. However, such heuristic variants of NMS are not always valid under different circumstances. Thus, more complex mechanisms may be introduced, for example, [18, 25] uses a neural network for more sophisticated and data-dependent duplicate removal. Although these methods raise the upper bound of naïve NMS, the pipeline becomes much more complex and costly in computation. Other works such as [23, 17] propose to predict different NMS thresholds for different bounding boxes. As the major drawback, they need an extra structure for IoU/density estimation, which introduces more hyper-parameters. Besides, it is still difficult to distinguish heavily overlapped boxes as in Fig 2 (a).

**Loss functions for crowded detection.** A few previous works propose new loss functions to address the problem of crowded detection. For example, [43] proposes *Aggregation Loss* to enforce proposals to be close and locate compactly to the corresponding ground truth. [40] proposes *Repulsion Loss*, which introduces extra penalty to proposals intertwine with multiple ground truths. The quality of detections in crowded scenes is improved with the help of these loss functions. However, since traditional NMS is still needed in the frameworks, it is still difficult to recall the overlapped instances illustrated in Fig 2 (a).

**Re-scoring.** In many detection frameworks [29, 24, 20, 21] a proposal box is bound to a ground truth as long as the overlap is larger than a given threshold, which usually leads to a *many-to-one* relation between proposals and ground-truth instances thus NMS is required to remove the duplicate proposals. Instead, if we redesign the loss function to encourage *one-to-one* relation, the NMS procedure may be eliminated to avoid miss detection. We name the method *re-scoring*. Some of the previous works follow the idea. For example, in [26, 27], each ground-truth instance is associated strictly to one proposal box during training. However, in the architectures of [26, 27], due to lack of connections between proposals, the prediction may be ambiguous, because it is not sure for one proposal to determine whether the related instance has been predicted by another proposal. Actually, in [26, 27] NMS is still involved. *RelationNet* [19], instead, explicitly models the relations between proposals, which is supposed to overcome the limitations of [26, 27]. Using re-scoring, RelationNet obtains outstanding performance on COCO [22] dataset even without NMS. However, in a more crowded dataset *CrowdHuman* [32], we find RelationNet with re-scoring performs relatively poor (see Sec. 4 for details). It may be because on CrowdHuman

dataset, proposals have to be much denser than those on COCO. As a result, the re-scoring network needs to generate different predictions from very close proposals (so their features and relations are also very similar, as shown in Fig. 2 (a)), which is infeasible for neural networks.

There are other approaches on crowded detection, for example, *part-based detectors* [44, 38, 46, 3], which is mainly used in detecting special instances such as pedestrian. The discussions are omitted in this paper.

In conclusion, based on the above analyses, we argue that object detection in the crowded scene may be *fundamentally* difficult, or at least *nontrivial* and complex for the mentioned existing proposal-based frameworks. The key issue lies in the basic paradigm of predicting only one instance for each proposal box. It inspires us to explore new *instance prediction* schemes, i.e. **multiple instance prediction** for each proposal.

### 3. Our Approach: Multiple Instance Prediction

Our approach is motivated by the observation: consider there are multiple objects heavily overlapped with each other, like the case in Fig. 2; if one proposal corresponds to any of the objects, it is very likely to overlap all the other objects. So, for such a proposal box, rather than predict a single object, why not predict them *all*? Formally, for each proposal box  $b_i$ , the new scheme suggests predicting the correlated **set** of ground-truth instances  $G(b_i)$  instead of a single object:

$$G(b_i) = \{g_j \in \mathcal{G} | \text{IoU}(b_i, g_j) \geq \theta\}, \quad (1)$$

where  $\mathcal{G}$  is the set of all the ground truth boxes and  $\theta$  is a given threshold of *intersection-over-union* (IoU) ratio. Fig. 2 (b) visualizes the concept. Compared with the previous *single-instance-prediction* framework, we find our new scheme may greatly ease the learning in crowded scenes. As shown in Fig. 2 (b), all of the three proposal boxes are assigned to the same set of ground-truth instance – it is a feasible behavior since the three proposals actually share almost the same feature. While for the previous single-instance-prediction paradigm (Fig. 2 (a)), each proposal has to produce distinguishing predictions, which might be intrinsically difficult.

We introduce the details of our approach as follows:

**Instance set prediction.** For each proposal box  $b_i$ , most of the modern proposal-based detection frameworks [29, 20, 21, 24, 13] employ a *detection function* to predict a pair  $(c_i, l_i)$  to represent the associated instance, where  $c_i$  is the class label with confidence and  $l_i$  is the relative coordinates. In our approach, to predict a set of instances, we

introduce a simple extension – just by introducing  $K$  detection functions to generate a set of predictions  $P(b_i)$ :

$$P(b_i) = \left\{ (\mathbf{c}_i^{(1)}, \mathbf{l}_i^{(1)}), (\mathbf{c}_i^{(2)}, \mathbf{l}_i^{(2)}), \dots, (\mathbf{c}_i^{(K)}, \mathbf{l}_i^{(K)}) \right\}, \quad (2)$$

where  $K$  is a given constant standing for the maximum cardinality of  $G(b_i)$  in the dataset (see Eq. 1).  $P(b_i)$  can be simply implemented by introducing extra prediction branches in most of the existing detection frameworks [29, 20, 21, 13, 24], which is shown in Fig. 3 (a). Note that even though  $K$  is fixed for all proposals, the network could still predict some  $\mathbf{c}_i^{(k)}$  to background class, representing that the  $k$ -th detection function does not predict instance for the proposal  $b_i$ .

**EMD loss.** We aim to design a loss  $\mathcal{L}(b_i)$  to minimize the gap between predictions  $P(b_i)$  and ground-truth instances  $G(b_i)$  corresponding to the proposal  $b_i$ , which can be cataloged into the problem of *set distance measurement*. Similar problems have been discussed in some early object detection papers, such as [36, 7, 35]. Inspired by them, we design the following *EMD loss* to minimize the *Earth Mover’s Distance* between the two sets:

$$\mathcal{L}(b_i) = \min_{\pi \in \Pi} \sum_{k=1}^K \left[ \mathcal{L}_{cls}(\mathbf{c}_i^{(k)}, g_{\pi_k}) + \mathcal{L}_{reg}(\mathbf{l}_i^{(k)}, g_{\pi_k}) \right] \quad (3)$$

where  $\pi$  represents a certain permutation of  $(1, 2, \dots, K)$  whose  $k$ -th item is  $\pi_k$ ;  $g_{\pi_k} \in G(b_i)$  is the  $\pi_k$ -th ground-truth box;  $\mathcal{L}_{cls}(\cdot)$  and  $\mathcal{L}_{reg}(\cdot)$  are classification loss and box regression loss respectively, following the common definitions as [29, 24, 20, 21, 13]. Note that in Eq. 3, we assume  $|G(b_i)| = K$ ; if not, we add some “dummy” boxes (whose class label is regarded as background and without regression loss) to  $G(b_i)$  until it is satisfied. Intuitively, the formulation in Eq. 3 implies to explore all possible one-to-one matches between predictions and ground truths, thus finding the “best match” with the smallest loss. It is also worth noting that if  $K = 1$ , Eq. 3 becomes equivalent to the loss in traditional *single-instance-prediction* frameworks, implying that our EMD loss is a natural generalization to the commonly-used detection loss [29, 24, 20, 13].

**Set NMS.** In our approach, although each proposal is able to predict multiple associated instances, if naïve NMS is still involved for post-processing it is impossible to detect objects effectively in crowded scenes. Fortunately, because of the *EMD loss*, instances predicted by one proposal are expected to be *unique* by definition. In other words, duplicates exist only between the predictions from different proposals, as illustrated in Fig. 2 (b). With this prior, we introduce a simple patch to naïve NMS pipeline, named *Set NMS* – each time before one box suppressing another

one in the NMS algorithm, we insert an additional test to check whether the two box come from the same proposal; if yes, we skip the suppression. Experiments in Sec. 4 also suggest that only when *multiple-instance-prediction* and *Set NMS* are used together can our approach achieve significant improvement in crowded detection.

**Refinement module.** In our approach, each proposal is expected to generate a set of instances rather than a single one, which may suffer from increase in false positives since more predictions are generated.

Although the failure cases are rarely observed in our experiments on real images, we introduce an *optional* refinement module in case of the risk, as shown in Fig. 3 (b). The module simply takes the predictions as input, combining them with proposal feature, then performs a second round of predicting. We expect the refinement module to correct the possible false predictions.

**Discussion: relation to previous methods.** Predicting multiple instance is not new. *Double-person detector* [37] models person pairs in the *DPM* [9] framework. In the deep learning era, some early detection systems [26, 27, 36, 7] also imply the high-level idea of *multiple-instance-prediction*, while the methods are **not** proposal-based. For example, *MultiBox* [7, 36] directly predicts all the instances in an image patch; *YOLO v1/v2* [26, 27] generates multiple predictions for each cell (i.e. predicting all instances centered at a certain location). Special loss functions are also proposed in [7, 36, 26, 27] for set prediction, whose design purposes are similar to our *EMD loss*.

The most related previous work to us is [35] which introduces *LSTM* to decode instance boxes in each grid of an image. Similar to our *EMD loss*, they use *Hungarian Loss* for multiple instance supervision. For post-processing, a box stitching method is employed to merge the predictions produced by adjacent grids. They mainly evaluated the method on head detection task, which shows some capability to predict crowded objects. However, since the method does not make use of proposals, it may have difficulty in detecting objects of various sizes/shapes, such as pedestrians or general objects. Moreover, the LSTM predictor is complex, which may be nontrivial to integrated in current state-of-the-art detection frameworks [20, 21, 13] efficiently.

### 3.1. Network Architecture

Theoretically, our approach can be applied to most of the state-of-the-art proposal-based detectors, no matter one-stage [24, 21, 28] or two-stage [29, 20, 13] frameworks. In this paper, we choose *FPN* [20] with *RoIAlign* [13] as a baseline detector to evaluate our method. In FPN, the *Region Proposal Network (RPN)* branch takes the responsibility for proposal generation, while the *RCNN* (or named

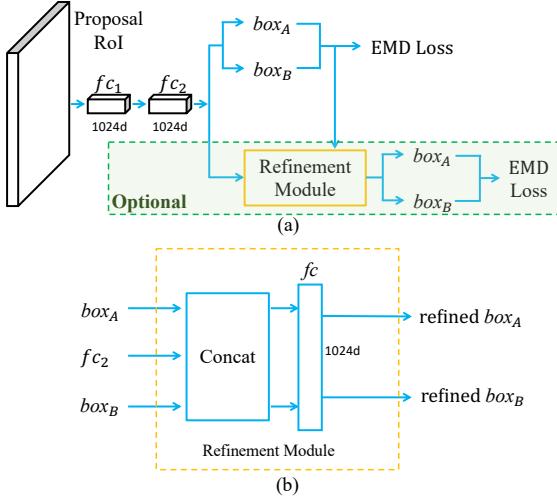


Figure 3. Overall architecture. (a)  $box_A$  and  $box_B$  are the two instances predicted by one proposal, using our EMD Loss. *Refinement module* is an optional step. (b) The refinement module concatenates feature and box information to optimize the results.

*RoI*) branch is used to predict the instance corresponding to the RoI proposal. So, our method is attached to the latter branch. From Sec. 3, easy to see that there is only one extra hyper-parameter in our approach –  $K$ , the maximum cardinality of  $G(\cdot)$  (refer to Eq. 2). In the rest of the paper, we let  $K = 2$ , which we find is satisfied for almost all the images and proposals in many detection datasets like *CrowdHuman* [32], *COCO* [22] and *CityPersons* [42].

Fig. 3 (a) illustrates the usage of our method in FPN (only RCNN branch is shown). Based on the original architecture, only slight modifications need to be made: just attach an additional instance prediction head to the tail. *EMD loss* is applied to the two predictions instead of the original loss. The *refinement module* is optional; if applied, we use the refined results as the final predictions.

## 4. Experiment

In this section, we evaluate our approach from different perspectives. Intuitively, a detection algorithm specially optimized for crowded scenes tends to recall more instances, however, often have the risk of **increasing false positive predictions**. Our benchmarks focus on both of the opposite aspects.

**Datasets.** An ideal object detector for crowded scenes should be robust to instance distributions, i.e. not only effective for crowded detections, but also stable to detect single/less-crowded objects. We adopt three datasets – *CrowdHuman* [32], *CityPersons* [42] and *COCO* [22] – for comprehensive evaluations on heavily, moderately and slightly overlapped situations respectively. Table 1 lists the

Dataset	# objects/img	# overlaps/img
CrowdHuman [32]	22.64	2.40
CityPersons [42]	6.47	0.32
COCO* [22]	9.34	0.015

Table 1. *Instance density* of each dataset. The threshold for overlap statistics is IoU > 0.5. \*Averaged by the number of classes.

“instance density” of each dataset. Since our proposed approach mainly aims to improve crowded detections. So, we perform most of the comparisons and ablations on *CrowdHuman*. Note that the experiment on uncrowded dataset like COCO is to verify whether our method does harm to isolated object detection, **not for significant performance improvements**.

**Evaluation metrics** We mainly take the following three criteria for different purposes:

- *Averaged Precision (AP)*, which is the most popular metric for detection. AP reflects both the precision and recall ratios of the detection results. In our experiment, we empirically find AP is more sensitive to the recall scores, especially on crowded dataset like *CrowdHuman*. Larger AP indicates better performance.
- $MR^{-2}$  [6], which is short for *log-average Miss Rate on False Positive Per Image (FPPI)* in  $[10^{-2}, 100]$ , is commonly used in pedestrian detection.  $MR^{-2}$  is very sensitive to false positives (FPs), especially FPs with high confidences will significantly harm the  $MR^{-2}$  ratio. Smaller  $MR^{-2}$  indicates better performance.
- *Jaccard Index (JI)* [24] is mainly used to evaluate the counting ability of a detector. Different from AP and  $MR^{-2}$  which are defined on the prediction *sequence* with decreasing confidences, JI evaluates how much the prediction *set* overlaps the ground truths. Usually, the prediction set can be generated by introducing a confidence score threshold. In this paper, for each evaluation entry, we report the best JI score by exploring all possible confidence thresholds. We use the official SDK of *CrowdHuman* [32] for JI calculation. Larger JI indicates better performance.

**Detailed Settings.** Unless otherwise specified, we use standard *ResNet-50* [15] pre-trained on *ImageNet* [31] as the backbone network for all the experiments. The baseline detection framework is *FPN* [20], while using *RoIPooling* [13] instead of original *RoIPooling*. As for anchor settings, we use the same anchor scales as [20], while the aspect ratios are set to  $H : W = \{1 : 1, 2 : 1, 3 : 1\}$  for *CrowdHuman* and *CityPersons*, and  $\{2 : 1, 1 : 1, 1 : 2\}$  for *COCO*. For training, we use the same protocol as in [20]. The batch

size is 16, split to 8 GPUs. Each training runs for 30 epochs. During training, the sampling ratio of positive to negative proposals for RoI branch is 1 : 1 for CrowdHuman and 1 : 3 for CityPersons and COCO. Multi-scale training and test are not applied; instead, the short edge of each image is resized to 800 pixels for both training and test. All box overlap IoU thresholds (e.g.  $\theta$  in Eq. 1, NMS thresholds, and those in calculating evaluation metrics) are set to 0.5 by default. For our method, we use  $K = 2$  (see Eq. 2). The *refinement module* in Fig. 3 is enabled by default.

#### 4.1. Experiment on CrowdHuman

*CrowdHuman* [32] contains 15,000, 4,370 and 5,000 images for training, validation and test respectively. For fair comparison, we retrain most of the involved models with our own implementation under the same settings. Results are mainly evaluated on the validation set, using full-body benchmark in [32].

**Main results and ablation study.** Table 2 shows the ablation experiments of the proposed methods in Sec. 3, including *multiple instance prediction* with *EMD loss*, *set NMS* and *refinement module*. The baseline is *FPN* [20] using NMS (IoU threshold is 0.5) for post-processing. It is clear that our methods consistently improve the performances in all criteria. Especially, even without *refinement module* our method still obtains **4.5%** improvements in AP and **2.2%** in JI, suggesting that more instances may correctly detected; more importantly, we find the  $MR^{-2}$  ratio also improves, indicating that our model does not introduce more false predictions. The refinement module affects little on AP and JI, while further boosting  $MR^{-2}$  by  $\sim 0.8\%$ , suggesting that the module mainly reduces false positives as we expected.

MIP	Set NMS	RM	AP/%	$MR^{-2}/\%$	JI/%
			85.8	42.9	79.8
✓			87.4	42.8	80.8
✓	✓		90.3	42.2	82.0
✓	✓	✓	<b>90.7</b>	<b>41.4</b>	<b>82.3</b>

Table 2. Ablation experiments evaluated on *CrowdHuman* validation set. The baseline model (the first line) is our reimplemented *FPN* [20] with *ResNet-50* [15] backbone. *MIP* – multiple instance prediction with *EMD loss*. *RM* – refinement module.

**Comparisons with various NMS strategies.** In Fig. 1, since some instances are mistakenly suppressed by *NMS*, one possible hypothesis is that the predictions may be improved by using different NMS strategies. Table 3 explores some variants. For naïve NMS, compared with the default setting (0.5), slightly enlarging the IoU threshold (from 0.5 to 0.6) may help to recall more instances, so AP increases; however, the  $MR^{-2}$  index becomes much worse (from 42%

Method	IoU*	AP/%	$MR^{-2}/\%$	JI/%
NMS	0.3	72.3	48.5	69.6
	0.4	80.7	44.6	76.3
	0.5	85.8	<b>42.9</b>	<b>79.8</b>
	0.6	<u>88.1</u>	45.4	79.4
	0.7	87.1	56.5	74.4
	0.8	82.8	68.5	62.6
Soft-NMS [1]	0.5	88.2	42.9	79.8
Ours (w/o RM)	0.5	90.3	42.2	82.0
Ours (with RM)	0.5	<b>90.7</b>	<b>41.4</b>	<b>82.3</b>

Table 3. Comparisons of different NMS strategies on *CrowdHuman* validation set. The baseline model is *FPN* [20]. \*IoU threshold for post-processing. *RM* – refinement module.

to 45.4%), indicating that more false positives are introduced. *Soft-NMS* [1] can boost AP score, but no improvements are obtained in  $MR^{-2}$  and JI. In contrast, our method achieves the best scores in all the three metrics even without *refinement module*.

**Comparisons with previous works.** To our knowledge, very few previous works on crowded detection report their results on *CrowdHuman*. To compare, we benchmark two methods – *GossipNet* [18] and *RelationNet* [19] – which are representative works categorized into *advanced NMS* and *re-scoring* approaches respectively (see Sec. 2 for the analyses). For *GossipNet*, we use the open-source implementation to benchmark<sup>1</sup>. And for *RelationNet*, we re-implement the *re-scoring* version<sup>2</sup>. All methods use *FPN* [20] as the base detector with the same training settings.

Table 4 lists the comparison results. Surprisingly, both *RelationNet* and *GossipNet* suffer from significant drop in AP and  $MR^{-2}$ . Further analyses indicate that the two methods have better recall ratio than baseline NMS for crowded objects (see Table 5), however, tend to introduce too many false positive predictions. Though it is still too early to claim [19, 18] do not work on *CrowdHuman* (we have not fully explored the hyper-parameters), at least the two methods are nontrivial for tuning. In contrast, our method is not only effective, but also very simple, as it has almost no additional hyper-parameters.

Table 4 also compares a recent work *AdaptiveNMS* [23], which is an enhanced NMS strategy for crowded detection. In [23], *CrowdHuman* results based on *FPN* are reported. Note that since the baseline are not aligned, we cannot make the direct comparison with our results. From the numbers, we find that our method can achieve significant improvement from a stronger baseline (especially in AP), in addi-

<sup>1</sup><https://github.com/hosang/gossipnet>

<sup>2</sup>We use the *re-scoring* version rather than *NMS*, as NMS is clearly not suitable for crowded detection. We have checked *COCO* [22] scores to ensure correct re-implementation.

tion, the pipeline is much simpler.

Table 4 also evaluate our method on the Cascade R-CNN [2] framework. We add the EMD loss and Set NMS into the last stage of Cascade R-CNN. The results show our method can still boost the performance of Cascade R-CNN significantly on crowded datasets like CrowdHuman.

Method	AP/%	$MR^{-2}/\%$	JI/%
FPN baseline	85.8	42.9	79.8
FPN + Soft-NMS [1]	88.2	42.9	79.8
RelationNet [19] (our impl.)	81.6	48.2	74.6
GossipNet [18] (our impl.)	80.4	49.4	81.6
<b>Ours</b>	<b>90.7</b>	<b>41.4</b>	<b>82.3</b>
FPN baseline (impl. by [23])	83.1	52.4	—
AdaptiveNMS (impl. by [23])	84.7	49.7	—
CascadeR-CNN [2] (our impl.)	86.2	40.2	80.4
CascadeR-CNN + Ours	<b>90.6</b>	<b>38.7</b>	<b>83.9</b>

Table 4. Comparisons of various crowded detection methods on *CrowdHuman* validation set. All methods are based on *FPN* detector [20]. Higher values of AP and JI indicate better performance, which is in contrast to the  $MR^{-2}$ . We use only one more stage in both of our Cascade R-CNN implementation instead of two for better performance in CrowdHuman.

**Analysis on recalls.** To further understand the effectiveness of our method on crowded objects, we compare the recalls of different approaches for both crowded and uncrowded instances respectively. Results are shown in Table. 5. Note that recall relates to the confidence score threshold. For fair comparison, we use the thresholds corresponding to the best JI index for each entry respectively. From the table we find that for *FPN* baseline/*Soft-NMS*, recall of crowded objects is much lower than that of uncrowded objects, implying the difficulty of crowded detection. In contrast, our method greatly improves the recall ratio of the crowded instances (from 54.4% to 63.3%, by **8.9%**), in addition, uncrowded recall is also slightly improved.

## 4.2. Experiments on CityPersons

CityPersons [42] is one of the widely used benchmarks for pedestrian detection. The dataset contains 5, 000 images (2, 975 for training, 500 for validation, and 1, 525 for testing, respectively). Each image is with a size of  $1024 \times 2048$ . Following the common practice in previous works, all of the object detectors are trained on the training (*reasonable*) subset and tested on the validation (*reasonable*) subset with the enlarged resolution by  $1.3\times$  compared to the original one, which is slightly different from the settings we used for *CrowdHuman* [32] and *COCO* [22]. To obtain a better baseline, we follow the strategy proposed

Method	Conf.	Total	Sparse	Crowd
Ground truth	-	99481	78665	20816
FPN +	0.7	80385	66880	13505
NMS		(80.8%)	(85.0%)	(64.9%)
FPN +	0.7	80385	66880	13505
Soft-NMS [1]		(80.8%)	(85.0%)	(64.9%)
FPN +	0.0	80196	65083	15113
GossipNet [18]		(80.6%)	(82.7%)	(72.6%)
FPN +	0.5	75178	60083	15095
RelationNet [19]		(75.6%)	(76.4%)	(72.5%)
<b>Ours</b>	0.7	<b>83246</b>	<b>67716</b>	<b>15530</b>
		( <b>83.7%</b> )	( <b>86.1%</b> )	( <b>74.6%</b> )

Table 5. Detection boxes recalled on *CrowdHuman* validation set. Only boxes with confidence score higher than a certain threshold are taken into account. The confidence thresholds are subject to the best JI scores respectively and noted in the “Conf” column. Numbers in the last three columns indicate the number of recalled boxes. “Crowd” means the corresponding ground-truth box overlaps with some other ground truth with  $IoU > 0.5$ , otherwise marked “Sparse”. Note that recalls of *Soft-NMS* [1] are the same with the NMS baseline, which is because the confidence threshold is relatively high (0.7), thus NMS is roughly equivalent to Soft-NMS.

in [3], namely evolving ground-truths into proposals by jittering. Other hyper-parameters remains the same as that in Sec. 4.

Method	Conf.	Total	Sparse	Crowd
Ground truth	-	1579	1471	108
FPN	0.6	1430	1366	64
+NMS		(90.6%)	(92.9%)	(59.3%)
FPN	0.6	1430	1366	64
+ Soft-NMS [1]		(90.6%)	(92.9%)	(59.3%)
<b>Ours</b>	0.6	<b>1476</b>	<b>1380</b>	<b>96</b>
		( <b>93.5%</b> )	( <b>93.8%</b> )	( <b>88.9%</b> )

Table 6. Detection recalls on *CityPersons* validation set. Please refer to the caption in Table 5 for the descriptions. Note that recalls of *Soft-NMS* [1] are the same as the NMS baseline, which is ascribed to the confidence threshold is relatively high (0.6), thus NMS is roughly equivalent to Soft-NMS.

**Qualitative results.** Table 7 compares our method with *FPN* baselines with naïve NMS and *Soft-NMS* respectively. Our approach improves AP and  $MR^{-2}$  by 0.9% and 1.0% respectively over the NMS baseline, indicating the effectiveness of our method. Table 7 also lists some other state-of-the-art results on *CityPersons*. Though it may be unfair for direct comparisons due to different hyper-parameter settings, however, at least it implies our method achieves significant gains over a relatively strong baseline.

Table 6 further analyzes the recalls of different methods.

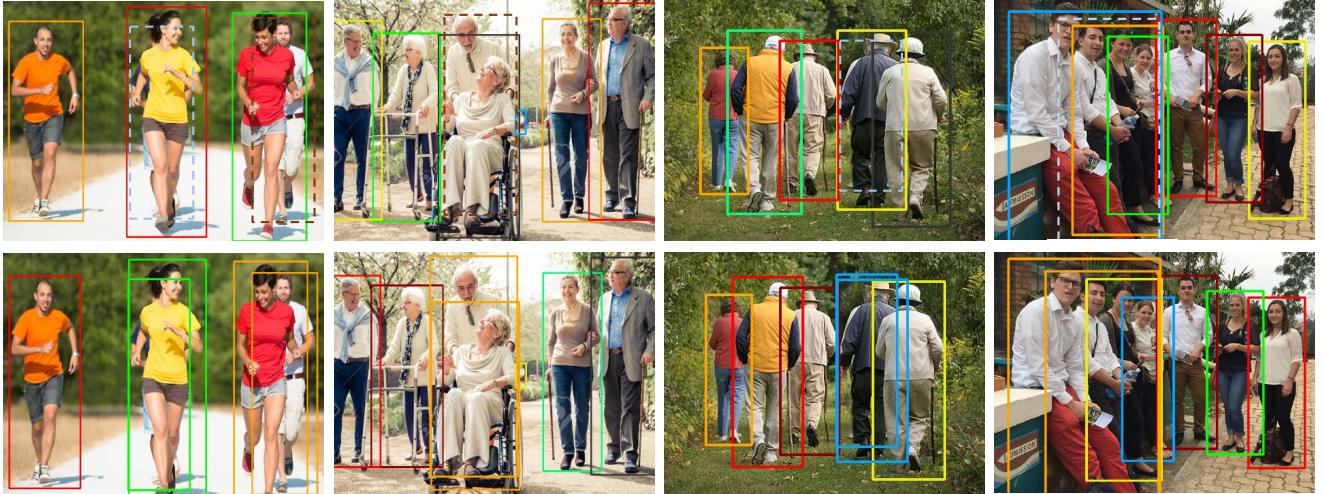


Figure 4. Visual comparison of the baseline and our approach. The first row are the results produced by FPN with NMS. The last row are the results of our approach. The scores threshold for visualization is 0.3. Boxes with the same color stem from the identical proposal. The dashed boxes are the missed detection ones.

Similar to those in *CrowdHuman* (refer to Table 6), our method mainly significantly boosts the recall on crowded objects – from 64 increased to **96** out of a total of 108 instances in the validation set. The comparison further indicates our approach is very effective to deal with crowded scenes again.

Method	Backbone	$MR^{-2}$	AP
FPN + NMS	Res-50	11.7%	95.2%
FPN + Soft-NMS [1]		11.8%	95.3%
<b>Ours</b>	Res-50	<b>10.7%</b>	<b>96.1%</b>
AF-RCNN [42]		12.8%	—
OR-CNN [43]	VGG-16	11.0%	—
Adaptive-NMS [23]		10.8%	—
FRCNN (our impl.) [29]	Res-50	11.6%	95.0%
Repulsion Loss [40]		11.6%	—

Table 7. Comparisons of different methods on *CityPersons* validation set. All models are evaluated with enlarged resolution of  $1.3 \times$  compared to the original size. Models in the upper part are trained by ours with the same *FPN* base detector. Models in the lower part are trained with other protocols.

### 4.3. Experiments on COCO

According to Table 1, the *crowdedness* of *COCO* [22] is very low, which is out of our design purpose. So, we **do not** expect a significant performance gain on *COCO*. Instead, the purpose of introducing *COCO* is to verify: 1) whether our method generalizes well to multi-class detection problems; 2) whether the proposed approach is robust to different crowdedness, especially to *isolated* instances.

Following the common practice of [20, 21], we use a

Method	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>
FPN	37.5	59.6	40.4	23.0	41.2	48.6
Soft-NMS	38.0	59.4	41.5	23.3	41.8	49.0
<b>Ours</b>	<b>38.5</b>	60.5	41.5	23.0	41.8	50.1

Table 8. Comparisons on *COCO* [22] *minival* set. All models are based on *FPN* detector. Results are evaluated in all the 80 classes.

subset of 5000 images in the original validation set (named *minival*) for validation, while using the remaining images in the original training and validation set for training. Table 8 shows the comparisons with *FPN* and *FPN+Soft-NMS* baselines. Moderate improvements are obtained, e.g. 1.0% better than naïve NMS and 0.5% better than Soft-NMS in AP. Interestingly, large objects achieve the most significant improvement (see AP<sub>L</sub> in Table 8), which may be because larger objects are more likely to overlap. The experiment suggests our method is not only very effective on crowded scenes, but also able to deal with multiple classes and isolated instances without performance drop.

## 5. Conclusion

In this paper, we propose a very simple yet effective proposal-based object detector, specially designed for crowded instance detection. The method makes use of the concept of *multiple instance prediction*, introducing new techniques such as *EMD loss*, *Set NMS* and *refinement module*. Our approach is not only effective, but also flexible to cooperate with most state-of-the-art proposal-based detection frameworks; in addition, also generalizes well to less crowded scenarios.

## References

- [1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS – Improving Object Detection With One Line of Code. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *arXiv preprint arXiv:1906.09756*, 2019.
- [3] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. *arXiv preprint arXiv:1909.06826*, 2019.
- [4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.
- [5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [6] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.
- [7] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.
- [8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015.
- [9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.
- [10] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. DSSD : Deconvolutional single shot detector. *CoRR*, abs/1701.06659, 2017.
- [11] Ross Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [13] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *european conference on computer vision*, pages 346–361. Springer, 2014.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [16] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2888–2897, 2019.
- [17] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. A convnet for non-maximum suppression. In *German Conference on Pattern Recognition*, pages 192–204. Springer, 2016.
- [18] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [19] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. *arXiv preprint arXiv:1711.11575*, 2017.
- [20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.
- [21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [23] Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.
- [25] Lu Qi, Shu Liu, Jianping Shi, and Jiaya Jia. Sequential context encoding for duplicate removal. *CoRR*, abs/1810.08770, 2018.
- [26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [27] Joseph Redmon and Ali Farhadi. Yolo9000: Better, Faster, Stronger. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [28] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.
- [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [30] Sitapa Rujikietgumjorn and Robert T. Collins. Optimized pedestrian detection for multiple and occluded people. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

- [31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [32] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.
- [33] Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection snip. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [34] Bharat Singh, Mahyar Najibi, and Larry S Davis. SNIPER: Efficient multi-scale training. *NIPS*, 2018.
- [35] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.
- [36] Christian Szegedy, Scott Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.
- [37] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele. Detection and tracking of occluded people. *International Journal of Computer Vision*, 110(1):58–69, Oct 2014.
- [38] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015.
- [39] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2974, 2019.
- [40] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [41] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. In *Advances in Neural Information Processing Systems*, pages 320–330, 2018.
- [42] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.
- [43] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.
- [44] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [45] Qiaoyong Zhong, Chao Li, Yingying Zhang, Di Xie, Shicai Yang, and Shiliang Pu. Cascade region proposal and global context for deep object detection. *Neurocomputing*, 2019.
- [46] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

## A. CrowdHuman Testing Benchmark

The *CrowdHuman*[32] testing subset has 5,000 images and the annotations of testing subset have not yet been released. To push the upper-bound of object detection research, the *Detection In the Wild Challenge* was held in CVPR 2019. The *CrowdHuman* testing subset is served as a benchmark in this challenge and this allows us to compare our approach to state-of-the-art methods on *CrowdHuman*.

To improve the performance of our approach, we replace the ResNet-50[15] with a larger model: SEResNeXt101[?, ?], the short edge of test images are resized to 1200 pixels and all other settings are the same as described in our paper. We then submit our result to the test server and find that our method outperforms all the results in this challenge. The results are shown in Table.9, and the full leaderboard is accessible on the official website of CrowdHuman Track Leaderboard<sup>3</sup>.

Rank	Team Name	Institution	mJI/%
1	zack0704	Tencent AI Lab	77.46
2	boke	Sun Yat-Sen University	75.25
3	ZNuanyang	Zhejiang University	74.46
Method		Backbone	mJI/%
Baseline		ResNet-50	72.20
Ours		ResNet-50	76.60
Ours		SEResNeXt101	<b>77.74</b>

Table 9. Part of the leaderboard and our results. The baseline model is our reimplemented FPN[20].

## B. Ablation on Number of Heads

For completeness, we further explore the only hyper-parameter of our method:  $K$  in this section. In our paper, we let  $K = 2$  because we find it is satisfied for almost all the images and proposals in *CrowdHuman*. If we make the  $K$  larger, the network will be able to detect instances in more crowded scenes. To explore the performance under the different values of the  $K$ , an experiment is conducted on the *CrowdHuman* dataset and all the settings remain the same as described in our paper except the value of  $K$  is changed. We show the results of different  $K$  values in the Table.10.

## C. More Results of Our Method

In this section we will show more results of our method on a video from YouTube and the *CrowdHuman* validation dataset. The visualization thresh is set to 0.7 to remove the redundant boxes in the results.

The video is in the attached file, and the results on the *CrowdHuman* validation dataset is shown in the Figure.5.

	AP/%	MR <sup>-2</sup> /%	JI/%
$K = 1$	85.8	42.9	79.8
$K = 2$	<b>90.7</b>	<b>41.4</b>	<b>82.3</b>
$K = 3$	<b>90.7</b>	41.6	82.1

Table 10. Ablation experiments evaluated on the *CrowdHuman* validation set. It worth noting that if  $K = 1$ , the architecture is the same as the *single-instance-prediction* baseline.

Method	AP/%	MR <sup>-2</sup> /%	JI/%
RetinaNet Baseline	82.0	56.3	73.2
<b>Ours</b>	<b>82.7</b>	<b>54.7</b>	<b>74.0</b>

Table 11. Experiment on RetinaNet[21]. All hyperparameters between the RetinaNet baseline and our method remain the same. The results show that one-stage detectors can also benefit from our method.

## D. Our Method in One-Stage Detector

To verify the effectiveness of our method in one-stage detectors, we conducted experiments on RetinaNet[21]. We adopt standard *ResNet-50* pre-trained on *ImageNet* as the backbone network. The anchor ratios are set to  $H : W = \{1 : 1, 2 : 1, 3 : 1\}$  because of the shape of human instances. We use a batch size of 16 pictures over 8 GPUs, and train the network with a learning rate of 0.005 for 50 epochs. The other hyperparameters are kept the same as [21]. For our method, we predict two predictions based on each anchor. And the Cross-Entropy Loss in EMD Loss is replaced by Focal Loss[21]. The results are shown in Tabel11, which indicates that our method can also have gain in one-stage detectors.

<sup>3</sup>[https://www.objects365.org/crowd\\_human\\_track.html](https://www.objects365.org/crowd_human_track.html)



Figure 5. Visual comparison of the baseline, GossipNet, RelationNet, and our approach. The blue boxes are the detection results, the white boxes are the missed instances, and the orange boxes are redundant boxes. The green boxes in our method are multiple predictions form one proposal.