Improving Crowded Object Detection via Copy-Paste

Jiangfan Deng, Dewen Fan, Xiaosong Qiu, Feng Zhou

Algorithm Research, Aibee Inc.
jfdeng100@foxmail.com, {dwfan,xsqiu,fzhou}@aibee.com

Abstract

Crowdedness caused by overlapping among similar objects is a ubiquitous challenge in the field of 2D visual object detection. In this paper, we first underline two main effects of the crowdedness issue: 1) IoU-confidence correlation disturbances (ICD) and 2) confused de-duplication (CDD). Then we explore a pathway of cracking these nuts from the perspective of data augmentation. Primarily, a particular copy-paste scheme is proposed towards making crowded scenes. Based on this operation, we first design a "consensus learning" strategy to further resist the ICD problem and then find out the pasting process naturally reveals a pseudo "depth" of object in the scene, which can be potentially used for alleviating CDD dilemma. Both methods are derived from magical using of the copy-pasting without extra cost for hand-labeling. Experiments show that our approach can easily improve the state-of-the-art detector in typical crowded detection task by more than 2% without any bells and whistles. Moreover, this work can outperform existing data augmentation strategies in crowded scenario.

Introduction

The task of object detection has been meticulously studied for quite a long time. In the deep learning era, in recent years, many well-designed methods (Liu et al. 2020a) have been proposed and raised the detection performance to a surprisingly high level. Nevertheless, there still exist many intrinsic problems that are not fundamentally solved. One of them is the "crowdedness issue", which usually denotes the phenomenon that objects belonging to the same category are highly overlapped together. In a geometrical manner, the basic difficulty stems from the semantical ambiguities of the 2D space. As shown in Fig. 1, in our 3D world, each voxel has its "unique semantics" and lies on a "certain object". However, after projecting to 2D plane, one pixel might fall on several collided objects. After evolving the concept from a "pixel" to a "box", the semantical ambiguity in crowded scenes leads to the notion of *overlap*.

To probe the effects of this problem, we now dive into the essence of the detection paradigm. Generally, an object detector reads in an image and outputs a set of bounding-boxes each associated with a confidence score. For an ideally-performed detector, the score value should convey how well the predicted box is overlapped with the ground-truth. In other words, the Intersection-over-Union (IoU) between these two boxes should be positively correlated with the confidence score. After visualizing the mean and standard deviation of scores with respect to IoU in Fig. 2, it turns out that even for the off-the-shelf detectors like (He et al. 2017), this positive correlation would be gradually disturbed by the increase of crowdedness degree[1]. This experimental study clearly indicates the struggle of current detection algorithms in facing the super-heavy overlaps. We embody this effect as IoU-confidence Correlation Disturbances (ICD). On the other hand, a typical detection pipeline often ends with a de-duplication module, for example, the widely adopted Non-Maximum Suppression (NMS). Due to the 2D semantical ambiguity mentioned previously, these modules are often confused by heavily overlapped predictions, which leads to severe missing in a crowd. We cast this type of effect as Confused De-Duplication (CDD).

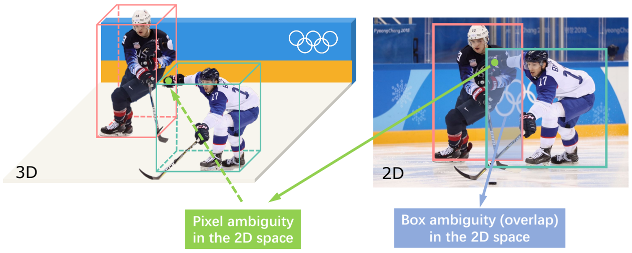

Figure 1: Semantic ambiguities in the 2D space. We exhibit the same scenario in the real 3D world (left) and the 2D space after photographing (right) respectively. The colored boxes represent two distinct objects (pucksters) while the *green* points denote a voxel in 3D space and its corresponding pixel in the 2D image. It is clearly illustrated that the 3D voxel lies on the body of a unique puckster while the 2D pixel lies on both of them. After evolving from a point to a bounding-box, the ambiguity arises in the form of overlap.

To overcome these two obstacles, we explore a pathway

[1] The crowdedness degree is indicated in terms of "occlusion ratio", *i.e.*, $1 - s_v/s_f$, where the s_v and s_f represent size of the visible box and full box of an object.

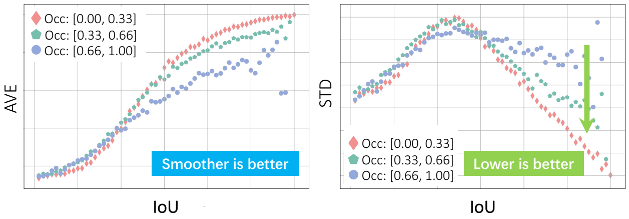

Figure 2: IoU-confidence correlation disturbances (ICD). We visualize the confidence score w.r.t the IoU between the predicted box from (He et al. 2017) and ground-truth in CrowdHuman (Shao et al. 2018). First, the IoU range of $[0, 1]$ are equally divided into 100 bins (each with the length of 0.01) as the horizontal axis. Then, *average value* (left) or *standard deviation* (right) of confidence scores are computed within each bin, generating a corresponding point in the coordinate plane. Marker shapes of diamond (*red*), pentagon (*green*) and circle (*blue*) refer to crowdedness degrees with the occlusion ratio on three ranges of $[0, 0.33]$, $[0.33, 0.66]$ and $[0.66, 1]$ respectively. On the left figure, the average score curve corresponding to the most crowded range (blue) are obviously more jittering than the other two curves; On the right figure, the heavier the crowdedness is, the larger the standard deviations are. Both figures suggest that the IoU-confidence correlation would become more uncertain when the crowdedness increases.

from the perspective of data augmentation. Referring to the preceding works (Ghiasi et al. 2021; Dwibedi, Misra, and Hebert 2017; Li et al. 2021; Dvornik, Mairal, and Schmid 2018; Fang et al. 2019), a simple copy-paste variant is proposed. Firstly, object segmentation patches are pasted to the training images following some specialized rules dedicated for making crowded scenes. Then, revolved from copy-pasting, we design a "consensus learning" approach to align confidence distributions of overlaid objects to their *identical but non-overlaid* counterparts, which further restrains the ICD problem. Moreover, thanks to the program-controlled pasting process, we can naturally get the extra order information of which one is in the front and which one is in the back when two (pasted) objects are overlapped. This cost-free knowledge provides cues on the additional *third dimension of depth* apart from x and y-axis spanning the image plane, which can be deemed as a breakthrough of the aforementioned 2D restrictions inducing the CDD dilemma. From this motivation, we propose a concept named "overlay depth" and semi-supervisely train the detector to predict this label. Then, an Overlay Depth-aware NMS (OD-NMS) is introduced to make use of the depth knowledge during de-duplication. Experiments show that this strategy can help distinguish boxes gathered in 2D space and further boost the detection results.

We evaluate our method from multiple aspects. As a data augmentation strategy, this work can outperform other counterparts in crowded scenes, no matter hand-craft methods or automated ones. As an approach of countering crowdedness issue, our method can stably improve the state-of-the-art detector by more than 2% without any bells and whistles. Moreover, since hand-labeling the crowded data is resource-consuming, this method provides a way of training on "sparse data" only and applying to crowded scenes via data augmentation.

To sum up, the major contributions of this work are two-fold: (1) We propose a crowdedness-oriented copy-paste scheme and introduce a consensus learning strategy, which effectively helps the detector resisting the ICD problem and bring improvements in crowded scenes. (2) We design a simple method to utilize the weak depth knowledge produced by the pasting process, which further optimize the detector.

Related Works

Crowded Object Detection. Detecting objects in crowded scenes has been a long-standing challenge (Liu et al. 2020a) and much effort has been spent on this topic. For example, (Wang et al. 2018) and (Zhang et al. 2018) propose specific loss functions to constrain proposals closer to the corresponding ground-truth and further away from the nearby objects, thereby enhancing discrimination between overlapped individuals. CaSe (Xie et al. 2020) uses a new branch to count pedestrian number in a region of interest (RoI) and generates similarity embeddings for each proposal. As a response to the CDD problem mentioned above, a group of works focuses on alleviating the deficiency of Non-Maximum Suppression (NMS). Adaptive-NMS (Liu, Huang, and Wang 2019) introduces an adaptation mechanism to dynamically adjust the threshold in NMS, leading to better recall in a crowd. In (Gählert et al. 2020) and (Huang et al. 2021), NMS leverages the less-occluded visible boxes to guide the selection of full boxes, whereas extra labeling (of the visible boxes) is required. CrowdDet (Chu et al. 2020) conducts one proposal to make multiple predictions and uses an artfully designed Set-NMS to solve heavily-overlapped cases. Some recent works explore other ways. (Zhang et al. 2021) models the pedestrian detection task as a variational inference problem. (Zheng et al. 2022) refines the end-to-end detector Sparse R-CNN (Sun et al. 2021) to adapt to the crowded detection scenario.

Data Augmentation in Object Detection. In the field of computer vision, data augmentation (Shorten and Khoshgoftaar 2019) has long been used to optimize the model training, which originates mainly from the image classification task (He et al. 2016; Tan and Le 2019). Early approaches usually include strategies such as color shifting (Szegedy et al. 2015) and random crop (Krizhevsky, Sutskever, and Hinton 2012; LeCun et al. 1998; Simonyan and Zisserman 2015; Szegedy et al. 2015). Naturally, the core ideas were transferred to the detection domain and some operations (*e.g.*, image flipping and scale jittering) have been widely adopted as a standard module (Liu et al. 2016; Redmon et al. 2016; Ren et al. 2015). Currently, methods with more concrete theoretical basis have emerged. These variants, ranging from hand-crafted Cutout (Devries and Taylor 2017), Mixup (Zhang et al. 2017) and CutMix (Yun et al. 2019) to learning based AutoAugment (Cubuk et al. 2018), Fast AutoAugment (Lim et al. 2019) and RandAugment (Cubuk et al. 2020), perform considerable effects on image classification and suggest huge potential in object detection. Meanwhile, there are also some works focusing on detection task. Stitcher (Chen et al. 2020) and YOLOv4 (Bochkovskiy, Wang, and Liao 2020) introduce mosaic inputs containing rescaled image patches to enhance robustness. (Zoph et al. 2020) and (Chen et al. 2021) re-design the AutoAugment scheme to adapt to object detection. In (Tang et al. 2021), researchers propose a method searching the policy of data augmentation and loss function jointly. In (Liu et al. 2020b), a novel APGAN is proposed to transfer pedestrians from other datasets in making augmentation.

Copy-Paste Augmentation. Copy-paste augmentation is first invented in (Dwibedi, Misra, and Hebert 2017). By cutting object patches from the source image and pasting to the target one, a combinatorial amount of synthetic training data can be easily acquired and improve the detection/segmentation performance significantly. This amazing magic power is then verified by subsequent works (Remez, Huang, and Brown 2018; Li et al. 2021; Fang et al. 2019; Dvornik, Mairal, and Schmid 2018; Ghiasi et al. 2021) and the method has been further polished by context adaptation (Fang et al. 2019; Remez, Huang, and Brown 2018; Dvornik, Mairal, and Schmid 2018). In (Ghiasi et al. 2021), the authors claim that simple copy-paste can bring considerable improvement as long as the training is sufficient enough. Their experiments further suggest the potential of this augmentation strategy on instance-level image understanding. It should be noted that the initial motivation of copy-paste is to diversify the sample space, especially for the rare categories (Ghiasi et al. 2021) or alleviating the complex mask labeling (Remez, Huang, and Brown 2018). However, in our work, we utilize this operation to precisely solve the crowdedness issue. Although there has been simple practice in previous works (Dwibedi, Misra, and Hebert 2017; Ghiasi et al. 2021), the actual effects of this strategy on dealing with crowdedness scenario has never been systematically designed and studied.

Resist the IoU-Confidence Disturbances

This part focuses on solving the Iou-Confidence Disturbances (ICD). We explore two consecutive ways in achieving this aim. First, doing copy-paste to make crowded scenes. Then, introducing consensus learning between overlaid objects and their non-overlaid counterparts, which relies on the copy-pasting.

Crowdedness Oriented Copy-Paste

Based on observations of Fig. 2, an intuitive idea is to make more crowded cases to dominate the training. To this end, we carefully re-design the copy-paste strategy. First, the conception of "group" is introduced. An image should include several groups and each group consists of multiple heavily overlapped objects. Following this logic scheme, we first generate the group centers on an image and then paste objects around them.

Formally, for every training image to be augmented, we initialize a set $\mathcal{C}$ of "group centers":

$$\mathcal{C} = \{(x_1, y_1, s_1), ..., (x_{|\mathcal{C}|}, y_{|\mathcal{C}|}, s_{|\mathcal{C}|})\},$$

where each tuple represents the object locating at center of the corresponding group (x_i, y_i and s_i denote the coordinates and normalized object size respectively). We obtain these group centers by sampling from original objects on the current image. The group number $|\mathcal{C}|$ is randomly chosen from an integral range of $[0, N]$, where *N* is a hyper parameter.

The second step is pasting objects around these group centers. For each $c_i \in \mathcal{C}$, we should generate a set $\hat{\mathcal{G}}_i$ of objects in the group i:

$$\hat{\mathcal{G}}_i = \{(\hat{x}_1^i, \hat{y}_1^i, \hat{s}_1^i), ..., (\hat{x}_{|\hat{\mathcal{G}}_i|}^i, \hat{y}_{|\hat{\mathcal{G}}_i|}^i, \hat{s}_{|\hat{\mathcal{G}}_i|}^i)\},$$

similarly, object number $|\hat{\mathcal{G}}_i|$ in the group comes from range $[0, M]$ where *M* is another hyper parameter. Since the nature of crowdedness is "overlapping", every $\hat{g}_j^i \in \hat{\mathcal{G}}_i$ is enforced to be overlapped with the group center object c_i. We manipulate the overlapping from three aspects of the *x*, *y* and *s* conditioning in a probabilistic sense.

First, objects in a group usually have similar sizes. Let $p(\hat{s}_j^i|s_i, I)$ be the probability density function of $\hat{s}_j^i$ on conditions of the center object size s_i in the image *I*. We choose $p(\cdot)$ to be a Gaussian as:

$$p(\hat{s}_j^i|s_i, I) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\hat{s}_j^i - s_i)^2}{2\sigma^2}), \tag{1}$$

where σ is the standard deviation which a constant value 0.2 is used in this paper. To guarantee overlapping, we adopt two independent uniform distributions in modeling the coordinate values $\hat{x}_j^i$ and $\hat{y}_j^i$:

$$\hat{x}_j^i \sim U(x_i - \frac{d_w}{\tau}, x_i + \frac{d_w}{\tau}), \tag{2}$$

$$\hat{y}_j^i \sim U(y_i - \frac{d_h}{\epsilon}, y_i + \frac{d_h}{\epsilon}), \tag{3}$$

where d_w and d_h are the maximum distances of $\hat{g}_j^i$ shifting from group center c_i with overlap. Coefficients $\tau > 1$ and $\epsilon > 1$ are used to adjust the crowdedness degree.

During training, for every image loaded, the set $\mathcal{C}$ and $\hat{\mathcal{G}}_i$-s are generated obeying rules above. Then object segmentation patches would be sampled, re-scaled and pasted to the image accordingly.

Consensus Learning

With the toolkit of copy-pasting, we augment detector training with a dedicated strategy for resisting the ICD issue. Given the observation shown in Fig. 2 that the instability of predicted scores derives from crowdedness, an emerging fix is to align the score of an object in crowded circumstances (overlaid by other objects) to that when it is not overlaid. Thanks to the copy-paste method, we can easily generate this type of object pairs in which two identical objects lie in different surroundings. Fig. 3 illustrates our idea. Following the previous data augmentation, we pick out a set $\mathcal{B}_{ovl}$ of objects which are overlaid by others. Then, the same object patches with those in $\mathcal{B}_{ovl}$ are re-pasted to the image without been overlaid, constructing another set $\mathcal{B}_{ovl}^*$. During training, we enforce the predicted score distributions of each object $b_i \in \mathcal{B}_{ovl}$ in an alignment with its counterpart $b_i^* \in \mathcal{B}_{ovl}^*$.

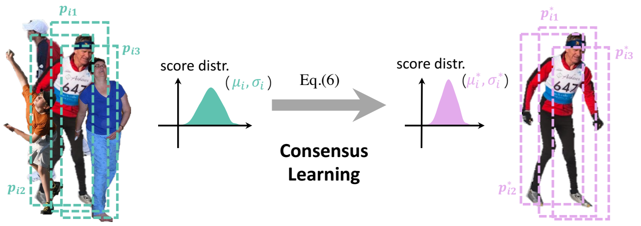

Figure 3: Consensus Learning. Learn to reach consensus between the overlaid object (the man in red on the left) and its identical but non-overlaid counterpart (right).

We term this process as *consensus learning* by drawing an analogy of "reaching consensus" within each pair. Specifically, let $\mathcal{P}_i$ be the set of proposals matched to b_i and $\mathcal{P}_i^*$ be the set of proposals matched to b_i^*. We first compute the mean μ and standard deviation σ of scores for each object:

$$\mu_i = \frac{1}{m} \sum_{p_{ij} \in \mathcal{P}_i} c(p_{ij}), \quad \sigma_i = \sqrt{\frac{1}{m} \sum_{p_{ij} \in \mathcal{P}_i} (c(p_{ij}) - \mu_i)^2}, \quad (4)$$

$$\mu_i^* = \frac{1}{m^*} \sum_{p_{ij}^* \in \mathcal{P}_i^*} c(p_{ij}^*), \;\; \sigma_i^* = \sqrt{\frac{1}{m^*} \sum_{p_{ij}^* \in \mathcal{P}_{ij}^*} (c(p_i^*) - \mu_i^*)^2}, \quad (5)$$

where m and m^* are the sizes of $\mathcal{P}_i$ and $\mathcal{P}_i^*$ respectively and $c(\cdot)$ denotes the predicted confidence score of a proposal. Then we pursue a pair of $\{\mu_i, \sigma_i\}$ approaching $\{\mu_i^*, \sigma_i^*\}$ through the mean squared error (MSE) loss:

$$L_{cl} = \frac{1}{|\mathcal{B}_{ovl}|} \sum_{b_i \in \mathcal{B}_{ovl}} (\mu_i - \mu_i^*)^2 + (\sigma_i - \sigma_i^*)^2. \quad (6)$$

It is worth to point that only the overlaid half $\{\mu_i, \sigma_i\}$ contributes to the gradient back-propagation while the non-overlaid half (marked by $*$) is treated as target.

Analyze the IoU-Confidence Disturbances

Now we analyze the effectiveness of our method on mitigating the aforementioned ICD issue. To revisit the original motivation raised from the right of Fig. 2, we plot the standard deviation (STD) of scores in Fig. 4. First, it is clearly demonstrated that score STDs of the model trained with our Crowdedness-oriented Copy-Paste (CCP) are obviously ***lower*** than those of the baseline model (BL) and the gap becomes larger by improving the crowdedness degree (from Fig. 4-(a) to (d)). Second, although the curves of CCP and CCP+CL seems with no clear distinction, after computing their average STDs (the four histograms in Fig. 4), we find the value of the latter is actually lower than that of the former. Moreover, we plot another model augmented with random copy-paste (RCP) without specially taking crowdedness into consideration. It is obvious that the decline of score STDs is with a much smaller margin. These observations convince that our method can significantly improve the detector's robustness in crowded scenes and therefore alleviate the ICD problem.

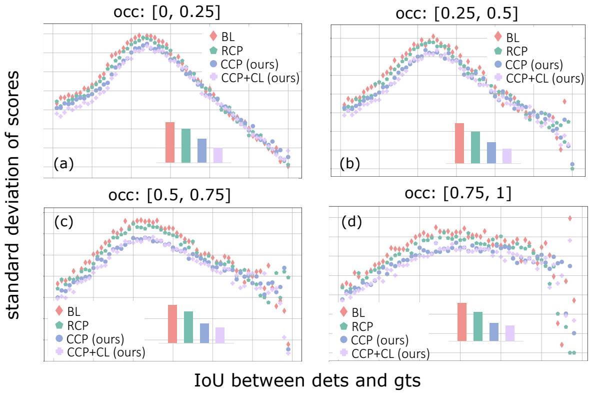

Figure 4: Effects of our method on the ICD issue, lower is better. We plot only the *standard deviation* of confidence scores w.r.t the IoU value on CrowdHuman. The crowdedness (occlusion ratio) gradually increases from (a) to (d).

Alleviate the Confused De-Duplications

Our augmentation strategy has a natural by-product: for these overlapped objects pasted, the relative "order of depth" is known a priori. In other words, we are aware of which one is in the front and which one is in the back. Now let us return to the semantical ambiguity described in our introduction. Basically, ambiguities in 2D space are caused by the absence of one dimension in the real (3D) world. From this point of view, the depth order can be viewed as some weak knowledge of the additional *third dimension*, which shed light on mitigating the vagueness. As a feasible practice, in this work, we utilize the depth order information to resolve the confused de-duplication (CDD) problem.

First, we introduce a variable named "overlay depth" (OD) that depicts the extent of how an object is visually overlaid by others. Fig. 5 demonstrates the process of calculating OD. We start by assuming that the overlay depth of an object equals to 1.0 if there are no other objects covering it. Let $ovl(b_1, b_2)$ be the region of object b_1 overlaid by object b_2 and $S(\cdot)$ denote the size of a region. For any object b_i in the image, there exists a set $\mathcal{O}_i$ of objects overlying b_i:

$$\mathcal{O}_i = \{b_j \in \mathcal{B} | b_j \neq b_i, S(ovl(b_i, b_j)) > 0\}, \quad (7)$$

where $\mathcal{B}$ is the set of all objects in current image. Then, the OD value of b_i can be clearly defined:

$$od_i = 1.0 + \frac{1}{S(b_i)} \sum_{b_j \in \mathcal{O}_i} S(ovl(b_i, b_j)). \quad (8)$$

Therefore, the severer an object is occluded by others (objects of the same category), the higher OD value it would be assigned (such as objects b_1 and b_2 in Fig. 5). Starting from this property, application of the overlay depth is based on a plausible observation: two heavily overlapped objects usually lie in different depth, or more specifically, hold distinct OD values. So by taking extra knowledge from the axis of depth, the OD value can be adopted during de-duplication in a confused 2D plane.

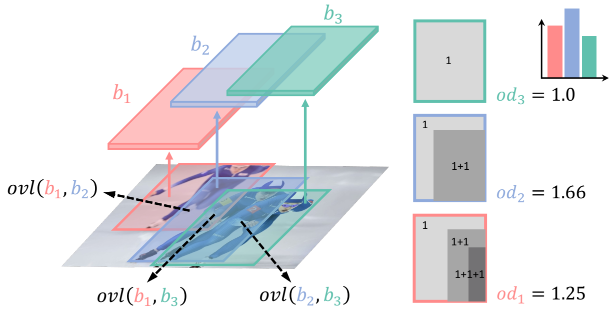

Figure 5: Definition of overlay depth (OD). Calculation process of the OD value as defined in Eq.(8). Boxes of b_1, b_2 and b_3 are three overlapped objects (skaters), in which b_2 is overlaid by b_3 only while b_1 is overlaid by both b_2 and b_3.

Now we enable the detector to predict the OD values. Generally, a detection model takes a branch to regress the coordinates of the bounding-box. Following this design, we add an extra predictor to the branch in taking responsibility for the OD regression. This modification incurs neglectable computing burden and can be easily implemented in both one-stage and two-stage structures (refer to the Appendix for details). During training, a common L2 loss is adopted. It should be emphasized that only the OD of pasted objects can be acquired due to the semi-supervised knowledge of the overlay depth. So we activate the OD regression loss only when the ground-truth is available. Formally, the whole loss can be written as below:

$$L_{det} = \begin{cases} \alpha \cdot L_{cls\_reg} + \gamma \cdot L_{cl} + \eta \cdot L_{od} & \text{if } od \text{ available} \\ \alpha \cdot L_{cls\_reg} + \gamma \cdot L_{cl} & \text{elsewise,} \end{cases} \tag{9}$$

where L_{cls_reg} is the conventional detection loss, L_{cl} is the consensus learning loss and L_{od} is OD regression loss respectively. We use $\alpha = \gamma = 1$ and $\eta = 0.1$ in this paper.

Algorithm 1: Overlay Depth-aware NMS

```
Input: $\mathcal{B} = \{b_1, ..., b_N\}$: All boxes; $\mathcal{S} = \{s_1, ..., s_N\}$: Scores; $th_{iou}$: IoU threshold.
$\mathcal{D} \leftarrow \varnothing$
while $\mathcal{B} \neq \varnothing$ do
    $m \leftarrow argmax\{\mathcal{S}\}$
    $\mathcal{M} \leftarrow b_m; \mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{M}; \mathcal{B} \leftarrow \mathcal{B} - \mathcal{M}$
    for $b_i$ in $\mathcal{B}$ do
        $th_{od} = \delta \cdot e^{\psi \cdot IoU(\mathcal{M}, b_i)}$
        if $IoU(\mathcal{M}, b_i) \geqslant th_{iou}$ and $|od_i - od_m| \leqslant th_{od}$ then
            $\mathcal{B} \leftarrow \mathcal{B} - b_i; \mathcal{S} \leftarrow \mathcal{S} - s_i$
        end if
    end for
end while
```

During inference, we invent a novel de-duplication strategy named Overlay Depth-aware NMS (OD-NMS). In the original NMS pipeline, boxes are recursively compared with each other and one of them would be suppressed in each step if the IoU exceeds a threshold th_{iou}. Following this scheme, objects might be de-duplicated by mistake in a crowded scenario. In our OD-NMS, for difficult scenario where IoU is higher than th_{iou}, we integrate the predicted OD value into a more comprehensive decision. If the two objects are in different depth, *i.e.*, the absolute difference of the two OD values is higher than a predefined threshold th_{od}, we can cancel the suppression in the current step. Empirically, ambiguous cases often raise in the range of large IoU: when two boxes are more heavily overlapped, we need stricter OD threshold to judge if they are distinct objects. So we design a dynamic threshold of OD with respect to the IoU value:

$$th_{od} = \delta \cdot e^{\psi \cdot IoU}, \tag{10}$$

where δ and ψ are constant coefficients.

Algorithm 1 summarizes the whole process. In this way, objects in a crowded scenario can be effectively recalled instead of being inappropriately de-duplicated. This strategy can be viewed as an evolvement of the original NMS with comparable time complexity.

Experiment

Datasets. Pedestrian detection is the most typical task burdened by the crowdedness problem, so our experiments are conducted mainly on two datasets: CrowdHuman (Shao et al. 2018) and CityPersons (Zhang, Benenson, and Schiele 2017). Annotations in these datasets consist of a full box and a visible box for each person, in which we only adopt the full ones to make the data crowded enough. Since both the training and validation data hold the same level of crowdedness, we prepare another "sparse training set" by re-labeling full body box of persons in COCO (Lin et al. 2014) to further evaluate the potential of our method. We name this train set as COCO-fullperson (we will release this dataset). Moreover, we use the category of "car" in KITTI (Geiger, Lenz, and Urtasun 2012) to further estimate the generality.

Augmentation Details. For pasting instance generation, we choose the open source Mask R-CNN (He et al. 2017) model adopting ResNet-50 (He et al. 2016) as backbone. We run this model on the train set and select 1000 instances with only three rough criteria: high confidence, relatively large size and not been occluded. A group of fixed hyper parameters are used in our experiments, where sample numbers $N = 3$ and $M = 5$, shifting coefficients $\tau = 4$, $\epsilon = 2$ and OD-NMS coefficient $\delta = 0.001$, $\psi = 10$. Copy-paste augmentation strategies are processed online within each training step, along with the generation of the semi-supervised OD ground-truths according to Eq.(8). We start consensus learning at the 10-th epoch during training.

Experimental Settings. We conduct experiments on both two-stage and one-stage detection frameworks. For two-stage structure, we use the standard Faster R-CNN (Ren et al. 2015) with FPN (Lin et al. 2017a). For one-stage structure, we choose RetinaNet (Lin et al. 2017b) as a representative. All those detectors use ResNet-50 as backbone. We train the networks on 8 Nvidia V100 GPUs with 2 images on each GPU. We also apply our method to the state-of-the-art

	MR^{-2}	AP@0.5	JI
Aug Method	on *Faster R-CNN*		
Baseline	50.42	84.95	-
Baseline^{+}	42.46	87.07	79.77
Mosaic	43.71	85.21	78.35
RandAug	42.17	87.48	80.40
SAutoAug	42.13	87.64	80.39
SimCP	41.88	87.36	79.53
CrowdAug (**Ours**)	**40.21**	**88.61**	**81.41**
Aug Method	on *RetinaNet*		
Baseline	63.33	80.83	-
Baseline^{+}	50.65	83.80	76.40
Mosaic	52.53	82.95	75.60
RandAug	50.25	83.94	76.58
SAutoAug	50.21	84.02	76.80
SimCP	50.01	84.12	77.02
CrowdAug (**Ours**)	**47.35**	**85.29**	**77.79**
on *SOTA* pedestrian detectors			
CrowdDet	41.35	90.06	82.07
ProgS-RCNN	41.45	92.15	83.13
CrowdDet + AutoPed	40.58	-	-
CrowdDet + **Ours**	**38.98**	91.50	**83.89**
ProgS-RCNN + **Ours**	40.12	**92.31**	83.35

Table 1: Results on CrowdHuman val set. The Baseline^{+} denotes newly trained strong baselines. Results are in percentage (%).

pedestrian detectors CrowdDet (Chu et al. 2020) and ProgS-RCNN (Zheng et al. 2022). Other training details will be reported in the following subsections.

Results on CrowdHuman

Three metrics are used to evaluate results on CrowdHuman: the *log-average miss rate on False Positive Per Image* (FPPI) in the range of $[10^{-2}, 10^{0}]$ (shortened as MR^{-2}, lower is better), the *Average Precision* (AP@0.5, higher is better) and the *Jaccard Index* (JI, higher is better), among which the MR^{-2} is the main indicator. To make our experiments convincing enough, we use very strong baselines (the Baseline^{+}s in Table 1), which are 8%-12% superior than those in the CrowdHuman paper (Shao et al. 2018). During training, the short side of each image is resized to 800 and the long side is limited within 1400. Models are trained for 60k iterations starting from an initial learning rate of 0.02 (Faster R-CNN) or 0.01 (RetinaNet) and is reduced by 0.1 on 30k and 40k iters respectively. Table 1 compares results of our method (CrowdAug) with other approaches. First, the widely used Mosaic augmentation (Bochkovskiy, Wang, and Liao 2020) leads to a decline. This phenomenon is mainly attributed to the fact that in CrowdHuman, many boxes extend across image boundary. After the mosaic operation, these near-boundary boxes are truncated at the joints of image patches, losing original characteristics. We also make trials of two automated strategies: the Random-Augmentation (RandAug) (Cubuk et al. 2020) and the Scale-Aware Auto-Augmentation (SAutoAug) (Chen et al. 2021). It needs to be noted that in these works, the search space does not include policies in dealing with crowded scene, which we hypothesize is the main reason of their marginal effects.

	MR^{-2}	AP@0.5	JI
Faster R-CNN	53.51	85.30	77.21
Faster R-CNN + **Ours**	**50.12**	**86.40**	**78.50**
RetinaNet	59.45	80.86	74.22
RetinaNet + **Ours**	**56.80**	**81.42**	**75.30**

Table 2: Results of model trained on COCO-fullperson and evaluated on CrowdHuman val set. We list results on Faster R-CNN and RetinaNet respectively.

CCP	CL	OD	MR^{-2}	AP@0.5
			42.46	87.07
(RCP)			42.01	87.10
√			41.11	87.75
√	√		40.80	88.02
√	√	√	40.21	88.61

Table 3: Ablation results on CrowdHuman val set. Experiments are conducted on Faster R-CNN.

The Simple Copy-Paste (Ghiasi et al. 2021) (SimCP in Table 1)improves the detector by nearly 0.6%. Instead, our CrowdAug can consistently improve the detection results by 2.2% and 3.3% for Faster R-CNN and RetinaNet respectively from the strong baselines. Moreover, the proposed method has exceptional performance on the state-of-the-art (SOTA) pedestrian detectors CrowdDet (Chu et al. 2020) and ProgS-RCNN (Zheng et al. 2022). As shown in the last two lines of Table 1. On CrowdDet, our method can achieve an improvement of 2.37% and reach a new SOTA of **38.98%** in MR^{-2}. On ProgS-RCNN (only the CCP is applied since the CL and OD-NMS is not needed for end-to-end detector), our method can bring an enhancement of 1.33%. The proposed CrowdAug can also outperform the previously SOTA augmentation strategy AutoPedestrian (Tang et al. 2021) by 1.6% in MR^{-2}. These experiments confirm that the CrowdAug can effectively optimize the crowded detection even on a supremely high base.

We also train the detector on the "sparse" dataset COCO-fullperson and report results on the "crowded" CrowdHuman val set in Table 2. Since training samples are generally not crowded, the CrowdAug can bring significant improvement (more than 3% in MR^{-2}). These results suggest that our method can largely help the detector to handle crowded scenes when there is limited or even no crowded data available for training.

Ablation Study

Crowdedness-oriented Design. The third line of Table 3 shows the contribution of our augmentation strategy (CCP). The CCP can improve the detection result by nearly 1.3%. For comparison, we try the random copy-paste (RCP) mentioned before. In this strategy, average number and size distribution of pasting objects are kept the same with those in our CCP while the positions to paste are randomly allocated rather than specially making crowded scenes. The 2nd line of Table 3 shows that the RCP improves the baseline by 0.45%, which is inferior to our CCP.

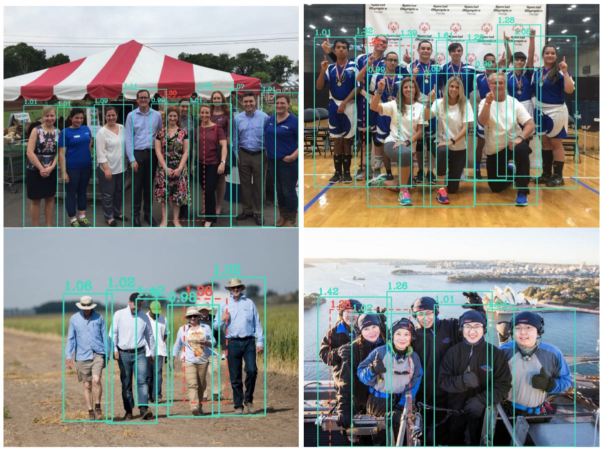

Figure 6: Visualization of the OD prediction. The value of predicted overlay depth (OD) is marked at the top left corner of each box. The *red* boxes denote the persons who are wrongly deleted by the original NMS are recalled.

Pasting Object Numbers	MR^{-2}	AP@0.5	JI
1000 (default)	40.21	88.61	81.41
3000	40.25	88.53	81.39
500	40.23	88.57	81.40
1000 sel	40.20	88.60	81.32
1000 sel+mask gt	40.21	88.62	81.42

Table 4: Robustness to Pasting Objects. The “sel” denotes manually selected high-quality objects and the “mask gt” means using segmentation annotations instead of those predicted by the Mask R-CNN model.

Consensus Learning. As shown in the 4-th line of Table 3, the proposed consensus learning (CL) strategy can further enhance the the Faster R-CNN by 0.3% from CCP baseline. This improvement becomes much larger (0.88%, not shown) when applying to RetinaNet. With qualitative analysis in the method part, we can make a conclusion that this module makes a step further in alleviating the ICD problem.

Overlay Depth. Comparing the last two lines of Table 3 can find out contribution of the overlay depth (OD). As a breakthrough of the 2D constraint, this weak depth knowledge brings a stable enhancement.We make visualizations of the OD prediction in Fig. 6. It can be seen that although the training process is semi-supervised, overlay depths learned by the detector are quite discriminative and can recall missing pedestrians (red dotted boxes in Fig. 6) of the baseline model. In the structure design, the simplicity of our OD predictor guarantees the ease of use during application.

Robustness to Pasting Objects. Our method is robust to the quantity and quality of pasting objects. Results in Table 4 show that variations of either quantity or quality of pasting objects will not essentially effect the final performance,

Method	MR^{-2}				AP@0.5
	Reasonable	Partial	Bare	Heavy	
FRCNN	11.20	11.55	6.62	52.05	82.95
Mosaic	11.05	11.42	6.77	51.62	83.01
RandAug	10.84	11.20	6.31	51.27	82.97
APGAN	11.9	11.9	6.8	49.6	-
AutoPed	10.3	-	-	49.4	-
Ours	**10.02**	**10.48**	**5.79**	**48.50**	**83.78**
RetinaNet	13.60	14.32	7.22	55.61	79.31
Mosaic	13.20	14.58	7.50	54.90	79.31
RandAug	13.23	13.96	7.02	54.61	79.77
Ours	**12.38**	**13.07**	**6.49**	**52.96**	**80.86**

Table 5: Results on CityPersons val set. We list the MR^{-2} on four crowdedness levels: *reasonable*, *partial*, *bare* and *heavy*. The metric of AP@0.5 is also reported.

	Easy	Moderate	Hard
	on *Faster R-CNN*		
Baseline	97.24	89.77	79.44
CrowdAug	**98.30**	**91.07**	**81.69**

Table 6: Results on KITTI val set. We use the category of “cars” in KITTI (Geiger, Lenz, and Urtasun 2012) dataset. AP@0.7 (%) of *easy*, *moderate* and *hard* objects are listed.

Results on CityPersons

On CityPersons, images are trained and evaluated with input scale of ×1.3. During training, we use an initial learning rate of 0.02 (Faster R-CNN) or 0.01 (RetinaNet) for the first 5k iterations and reduce it by 0.1 continuously on the next two groups of 2k iterations. Table 5 compares our CrowdAug with other methods. The results show that the CrowdAug can stably optimize the detector and once the crowdedness becomes heavier, the improvement becomes larger.

Results on KITTI

To estimate the generalization of our method to other crowded objects, we make experiments on the category of “cars” in KITTI (Geiger, Lenz, and Urtasun 2012). Table 6 shows the results. After applying the CrowdAug, Average Precision of cars get improvement if 1.05%, 1.20% and 2.25% for the objects of easy, moderate and hard respectively for the Faster R-CNN structure, which demonstrate the similar trend of its performance on pedestrian detection.

Conclusion

In this paper, we point out two main effects of crowdedness issue in the visual object detection task and propose a solution from the perspective of data augmentation. First, we invent a novel copy-paste strategy to improve crowdedness and design a consensus learning method. Then, we reasonably use the weak information of depth produced by the pasting process. Both contributions can help alleviating the ambiguities of crowded 2D object detection. We think this is a new pathway of solving the crowdedness issue with the advantages of significant effect and resource conservation.

References

Bochkovskiy, A.; Wang, C.; and Liao, H. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR*, abs/2004.10934.

Chen, Y.; Li, Y.; Kong, T.; Qi, L.; Chu, R.; Li, L.; and Jia, J. 2021. Scale-aware automatic augmentation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9563–9572.

Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Meng, G.; Xiang, S.; Sun, J.; and Jia, J. 2020. Stitcher: Feedback-driven Data Provider for Object Detection. *CoRR*, abs/2004.12432.

Chu, X.; Zheng, A.; Zhang, X.; and Sun, J. 2020. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12214–12223.

Cubuk, E.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2018. AutoAugment: Learning Augmentation Policies from Data. *arXiv preprint arXiv:1805.09501*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Devries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *CoRR*, abs/1708.04552.

Dvornik, N.; Mairal, J.; and Schmid, C. 2018. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 364–380.

Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 1310–1319. IEEE Computer Society.

Fang, H.; Sun, J.; Wang, R.; Gou, M.; Li, Y.; and Lu, C. 2019. InstaBoost: Boosting Instance Segmentation via Probability Map Guided Copy-Pasting. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 682–691. IEEE.

Gählert, N.; Hanselmann, N.; Franke, U.; and Denzler, J. 2020. Visibility guided nms: Efficient boosting of amodal object detection in crowded traffic scenes. *arXiv preprint arXiv:2006.08547*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.; Cubuk, E. D.; Le, Q. V.; and Zoph, B. 2021. Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2918–2928. Computer Vision Foundation / IEEE.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Z.; Yue, K.; Deng, J.; and Zhou, F. 2021. Visible feature guidance for crowd pedestrian detection. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, 277–290. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 1106–1114.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11): 2278–2324.

Li, C.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 9664–9674. Computer Vision Foundation / IEEE.

Lim, S.; Kim, I.; Kim, T.; Kim, C.; and Kim, S. 2019. Fast AutoAugment. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 6662–6672.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T. Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017b. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP(99): 2999–3007.

Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. 8693: 740–755.

Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P. W.; Chen, J.; Liu, X.; and Pietikäinen, M. 2020a. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.*, 128(2): 261–318.

Liu, S.; Guo, H.; Hu, J.-G.; Zhao, X.; Zhao, C.; Wang, T.; Zhu, Y.; Wang, J.; and Tang, M. 2020b. A novel data augmentation scheme for pedestrian detection with attribute preserving GAN. *Neurocomputing*, 401: 123–132.

Liu, S.; Huang, D.; and Wang, Y. 2019. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6459–6468.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Remez, T.; Huang, J.; and Brown, M. 2018. Learning to Segment via Cut-and-Paste. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, 39–54. Springer.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, 91–99.

Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; and Sun, J. 2018. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *arXiv preprint arXiv:1805.00123*.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on Image Data Augmentation for Deep Learning. *J. Big Data*, 6: 60.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14454–14463.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. 1–9.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tang, Y.; Li, B.; Liu, M.; Chen, B.; Wang, Y.; and Ouyang, W. 2021. Autopedestrian: an automatic data augmentation and loss function search scheme for pedestrian detection. *IEEE transactions on image processing*, 30: 8483–8496.

Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; and Shen, C. 2018. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7774–7783.

Xie, J.; Cholakkal, H.; Anwer, R. M.; Khan, F. S.; Pang, Y.; Shao, L.; and Shah, M. 2020. Count-and similarity-aware r-cnn for pedestrian detection. In *European Conference on Computer Vision*, 88–104. Springer.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3221.

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. 2018. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 637–653.

Zhang, Y.; He, H.; Li, J.; Li, Y.; See, J.; and Lin, W. 2021. Variational pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11622–11631.

Zheng, A.; Zhang, Y.; Zhang, X.; Qi, X.; and 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*, 857–866.

Zoph, B.; Cubuk, E. D.; Ghiasi, G.; Lin, T.; Shlens, J.; and Le, Q. V. 2020. Learning Data Augmentation Strategies for Object Detection. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVII*, 566–583. Springer.