Knowledge-Based Systems 293 (2024) 111685

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Overlapping object detection with adaptive Gaussian sample division and asymmetric weighted loss☆

Yao Xue [a], Yawei Zhang [a], Yuxiao Liu [b], Xueming Qian [a,*]

[a] School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[b] School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Keywords:
Overlapping object detection
Global location distribution head
Adaptive sample division
Asymmetric weighted loss

ABSTRACT

Existing deep learning based detectors are mostly designed for scenes with sparsely distributed objects. However, in certain scenarios such as dense crowds, objects often overlap severely. The dense anchor arrangement in anchor-based detectors is not quite suitable for the overlapping object detection. Anchor-free detectors have the potential to achieve high-performance in overlapping object detection, but troubled by the extreme imbalance of positive and negative samples. To this end, we propose an anchor-free overlapping object detector. Our adaptive Gaussian sample division (AGSD) can effectively allocate positive and negative samples with clear semantics to overlapping objects. Secondly, asymmetric weighted loss (AW Loss) adapts to continuous positive and negative sample values, thereby improving the classification ability of the detector. Lastly, our global location distribution head (GLD head) can introduce the supervision of overlapping object distributions. To verify the effectiveness of our method, we construct a large-scale high-quality overlapping object detection dataset containing 6173 images and 17,725 annotations. Compared with mainstream object detector, our method achieves the best performance of AP_{50} at 96.71%.

1. Introduction

Currently, deep learning based object detectors trained on the general object detection dataset COCO [1] and VOC [2] can handle most situations in real scenes well. However, there are many scenes with serious overlapping objects, such as densely crowded people, containers full of goods, and mobile phones during call behavior. The performance of classic detectors is not satisfactory in these scenes. Overlapping object detection is an important and challenging task.

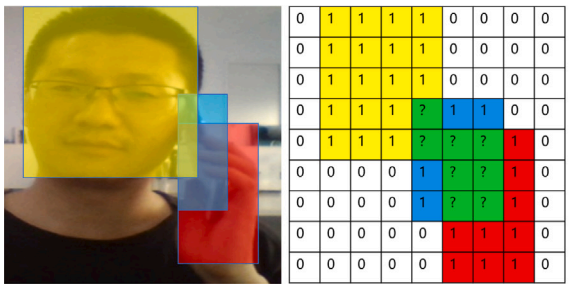
To cope with mutual occlusion of overlapping objects, repulsion loss [3] designs a new box regression loss function based on attraction by objects and the repulsion by other surrounding objects. It achieves good performance in pedestrian detection. NMS Loss [4] is end-to-end trainable by designing pull loss and push loss for the non-maximum suppression (NMS) process, so that the false negative is retained and the false positive is suppressed.

Even though these methods improve overlapping objects detection performance, they still have certain shortcomings. Tightly arranged anchors cannot cope well with scenes with severely overlapping objects. And the performance depends on boxes sizes, aspect ratios, etc. These hyperparameters need to be carefully adjusted. The aforementioned shortcomings motivate us to establish an accurate and effective anchor-free detector for overlapping object detection. The anchor-free detectors [5–8] remove the anchor mechanism to simplify the post-processing process. And it avoids the process of complicated IoU calculation for sample division. However, directly applying the anchor-free idea to overlapping object detection does not perform as expected. It is challenging due to the following difficulties.

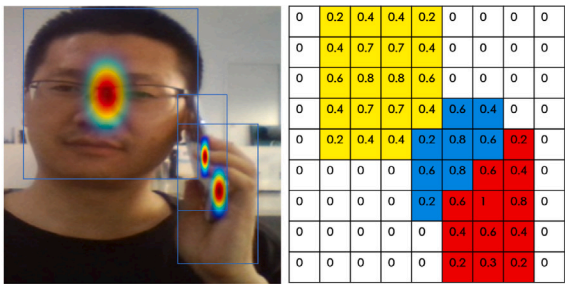
The first difficulty is that overlapping objects often have semantic conflict when dividing samples. The pixels in the overlapping area belong to multiple objects, but it is difficult to be reasonably classified as a positive sample of a suitable object. As shown in Fig. 1, the yellow, blue, and red boxes represent the division results of positive samples of face, mobile phone, and hand, respectively. In Fig. 1(a), all pixels in the boxes are divided into positive samples in FCOS [6], while the positive samples in the overlapping part (green regions) are difficult to be effectively divided into one of three classes due to semantic conflicts. The unreasonable division of samples does not make the ratio of positive and negative samples unbalanced. Facing the imbalance of

☆ This work was supported in part by NSFC, China under Grant 62103317, 61772407 and 61732008, in part by China Postdoctoral Science Foundation 2021M702600, by Shaanxi Key R&D 2022QFY01-17 and 2022FP-40.
* Corresponding author.
E-mail address: qianxm@mail.xjtu.edu.cn (X. Qian).

https://doi.org/10.1016/j.knosys.2024.111685
Received 31 December 2021; Received in revised form 26 December 2023; Accepted 20 March 2024
Available online 22 March 2024



(a) FCOS's sample division. Yellow, blue, and red box represents the division results of positive samples of face, mobile phone and hand. FCOS treats all pixels in the boxes as positive samples. The overlapping green pixels are difficult to be effectively divided into the three classes.



(b) Adaptive Gaussian sample division. AGSD performs Gaussian blur on ground-truths. The objectivity of positive samples is quantified according to the distance of the pixel to the object center. For the overlapping part, the class with the higher objectivity is regarded as the positive sample.

Fig. 1. Comparison of sample division methods. AGSD alleviates semantic conflicts of samples and quantifies objectivity.

positive and negative samples in the anchor-free detector, Tian et al. [6] alleviate the imbalance of positive and negative samples by specifying all pixels in the real boxes as positive samples. Zhang et al. [9] divide the positive and negative samples according to the statistical method by counting the IoU between the predicted boxes and the real ones. These works promote the study of positive and negative sample division methods in anchor-free detectors. However, the collision between overlapping objects requires more effective guidance for the division of positive and negative samples. In Fig. 1(b), AGSD can adaptively establish a Gaussian distribution according to the shape of the box, to divide the semantically overlapping area into a flexible and reasonable positive sample division. **We propose AGSD to avoid collisions of different classes and alleviate the imbalance problem of positive and negative samples.**

As for the second defeat in overlapping object detection, the weak classification ability of the anchor-free detector is not suitable for pixel-by-pixel classification. When the overlapping area between the objects belongs to two or more classes of objects at the same time, the pixels in the overlapping area are difficult to be judged as a certain class. Humans can classify overlapping pixels based on the edges of objects, but existing detectors are difficult to do the above. The semantic conflict caused by severe overlap poses a severe challenge to the pixel-by-pixel classification task of the anchor-free detector. In order to improve the classification ability of the detector, He et al. [10] made the detector pay more attention to the difficult samples by weighting the difficult and simple samples. Duan [5] et al. highlight the positive samples in the training process by suppressing the negative samples around them. The anchor mechanism and the suppression of negative samples around positive samples make the detector focus on positive samples in an image, thereby improving its classification ability. But in the anchor-free detector, improving the classification ability requires a more powerful classification loss to suppress the influence of pixels that are difficult to classify in the overlapping area. AGSD performs continuous processing ($y \in [0, 1]$) on discrete positive samples ($y = 1$) and negative samples ($y = 0$). The classic classification loss for example Fcoal loss [10] can only deal with the division of discrete positive and negative samples. On the basis of Gaussian partitioning, our asymmetric weighted loss **continuously processes the discretized positive and negative samples** with imbalanced weights for positive and negative samples.

The third difficulty is that after the detector performs feature extraction, the subsequent classification head and regression head cannot effectively constrain and predict the global location distribution of overlapping objects. In response to the down-sampling error caused by feature extraction, Duan et al. [5] introduces the offset head to predict the center point offset, which improves the anchor-free detector's ability to locate the object center point. Feng et al. [11] uses a new task alignment head to enhance the information interaction between the classification head and the regression head, thereby alleviating the mismatch between the classification score and the quality of the prediction box. The methods mentioned above mainly enhance the information interaction of the basic classification head and regression head to obtain better detection, but they cannot achieve effective predictions in the face of the location distribution between heavily overlapping objects. The direct regression to locations and classes of objects is very effective when objects are sparsely distributed. However, in the scene where objects overlap each other, the overlapping area troubles the object classification and box positioning. We introduce GLD head to introduce the ability for the detector to learn global location distribution of overlapping objects. This allows the detector to learn object distribution explicitly, which greatly improves the performance of the detector in the scenes with severe overlapping objects. **GLD head enhances classification features while explicitly learning the global location distribution of objects in overlapping scenes.** The main contributions of this paper are summarized as.

(1) AGSD deals with semantic conflict of positive and negative samples in overlapped areas. AGSD adaptively establishes a Gaussian distribution according to objects' shape, and obtains a clearly-attributed sample division result.

(2) We concatenate the mapping with 1 for positive samples and 0 for negative samples into the interval $[0, 1]$, and propose a dedicated AW Loss to improve the detector's ability to classify difficult samples.

(3) We propose GLD head to strengthen the constraint of the detector on the sample distribution of different objects during the training process. It can force the network to learn more clear classification heatmap results.

(4) To verify the effectiveness of our method, we construct a high-quality large-scale dataset containing 6173 images and 17,725 instances, with severe overlap between objects. Experiments confirm the effectiveness of our method.

2. Related works

2.1. Sample division in anchor-free detectors

It is annoying to delineate suitable positive and negative samples for pixels in the overlapping area. Unlike image classification task [12–16], we need to find and locate objects in the object detection task [17–19]. The object occupies a small number of pixels in the entire picture relative to the background. The problem of sample imbalance always bothers deep learning based detectors. The division of positive and negative samples in the overlapping area plagues object overlapping scenes such as crowds [20–24], object detection of commodities [25–28], and call behavior recognition [29].

According to whether it use anchor mechanism, detectors can be divided into anchor-based [10,30,31] and anchor-free [5,6,32]. The positive and negative sample division strategy of anchor mechanism is simple. Faster R-CNN [33] regards IoU greater than 0.7 as positive

samples, and IoU less than 0.3 as negative samples to alleviate the imbalance. Complex IoU and anchor regression increase the complexity and slow down the speed of anchor-based detectors. Tightly arranged anchors make it difficult to effectively and reasonably allocate samples for overlapping objects.

Anchor-free detectors alleviate the sample allocation problem between the prediction box and the overlapping objects. But since the anchor-free detector performs category prediction pixel by pixel, this mechanism leads to a sample imbalance phenomenon far beyond the anchor-based detectors. Duan et al. [5] divide the center point of the ground-truth box into positive samples, and other pixels in the image into negative samples. Furthermore, an improved Focal Loss [10] is used to suppress the difficult negative samples around the positive samples. This method improves the imbalance of positive and negative samples. However, it still cannot effectively solve the problem of sample division in the object overlapping area. YoloX [34] represents the current state-of-the-art model in anchor-free object detection. In terms of the label assignment strategy, YoloX introduces Similarity-based Optimal Transport Assignment (SimOTA), which analyzes label assignment from a global perspective rather than solely considering local information. In comparison, our AGSD dynamically generates Gaussian heatmaps, effectively addressing semantic conflicts during sample partitioning in the presence of overlapping targets.

In order to alleviate the semantic ambiguity faced by the sample division of the object overlapping area, we propose an adaptive method for dividing positive and negative samples based on Gaussian distribution. It can effectively respond to overlapping object sample division.

2.2. Classification loss for the sample imbalance

Classification loss is used to measure the degree of deviation between classification results and ground-truth. The imbalance between positive and negative samples lead to that the loss of easy samples is dominant in the total loss. Focal loss [10] en-balances the distribution between positive (easy) and negative (difficult) samples.

In particular, the anchor-free detector is troubled by the more serious imbalance of positive and negative samples than in the anchor-based detector. Duan et al. [5] propose to impose penalties on the difficult negative samples around the positive samples at the object center point to alleviate this problem. Furthermore, Qin et al. [35] assign weights to samples according to the quality of the detection boxes so that the network can pay more attention to the classification results of high-quality detection boxes. Liu et al. [36] design more reasonable weights for the classification loss of the positive samples and the classification loss of the negative samples according to the ratio between the positive and negative samples to balance the proportion of the respective loss functions of the positive and negative samples in the entire classification loss function.

The above method significantly alleviates the imbalance problem of positive and negative samples in the sparse object scene. In the face of overlapping objects, the sample definition and loss function design of overlapping area still puzzles researchers [3,4]. We make the positive samples of pixels in the overlapping area continuous based on the distance between the pixels and the object center. Furthermore, an asymmetric weighted loss is proposed to calculate the loss of positive samples in the overlapping area of objects.

2.3. Detection heads

The detector uses classification head and regression head to predict the object class and boundary information after extracting and fusing the features of the image. The feature map is usually 1/4 or 1/8 of the original image, so the error caused by downsampling is inevitably introduced to the center point of the object. Duan et al. [5] achieve more accurate center point positioning by adding an offset branch to the center point to predict the error caused by downsampling. Li et al. [37] propose a novel dynamic head framework to unify object detection head and attention. By coherently combining multiple self-attention mechanisms between the feature levels of scale perception, the spatial position of spatial perception, and the output channel of task perception, the proposed method significantly improves the representation ability of the object detection head. Kendall et al. [38] realize three tasks at the same time by adding heads of detection, segmentation, and depth estimation, and adopt a multi-task learning paradigm to improve the performance of the detector. Li et al. [39] use detection heads for the multi-level features of FPN to realize the hierarchical prediction of multi-scale faces, and construct a purely convolutional face detector. The distribution vector utilization method in this paper bear similarities to the technique in [40], which also employs a method of weighting features using distribution vectors. The difference lies in that our distribution vector focuses more in the domain of anchor-free object detection.

Improving or adding new detection heads for the detector can enhance the positioning ability and presentation ability. However, in the face of the complex location distribution between overlapping objects, the basic classification head and regression head cannot effectively learn them. Here GLD head enable the detector explicitly learn the global location distribution relationship of overlapping objects.

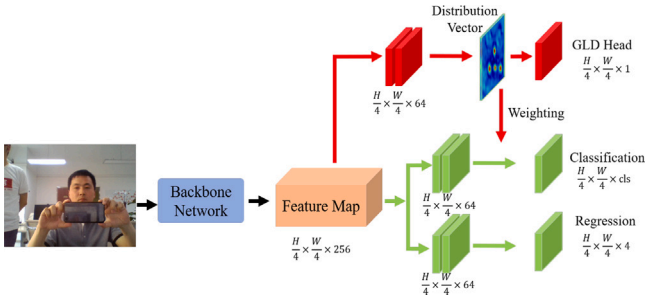


Fig. 2. Flowchart of network. Feature maps are generated by the backbone network. The classification head and the regression head respectively predict the center point and the shape of the object. The GLD head predicts the global object location distribution of the image and weights the feature map in the classification head pixel by pixel.

3. Proposed method

3.1. System overview

Object detection based on deep learning usually includes three processes: 1) Design the network structure to be responsible for feature extraction and result prediction. 2) According to the object classes and ground-truths, carry out suitable positive and negative samples division and box coding as the learning target of the prediction result. 3) Design loss function to quantitatively compare the difference between prediction and the target value, to supervise the network to update the parameters. Correspondingly, the network structure including GLD head will be introduced in the first part. AGSD is in the second part. AW Loss and the total loss function are given at the end.

3.2. Network structure

(1) Feature Generation

The pipeline of our network is shown in Fig. 2. ResNet50 [41] is used as a feature extractor. Suppose an image is expressed as $I \in R^{3\times W\times H}$. When features of resized images are extracted through ResNet50, we obtain the feature map of 1/4, 1/8, 1/16, and 1/32 down-sampled from the original image level by level. Feature Pyramid Network (FPN) is responsible for fusing extracted features. The output is $P2 \in R^{256\times W/4\times H/4}$. The features extracted from the backbone can

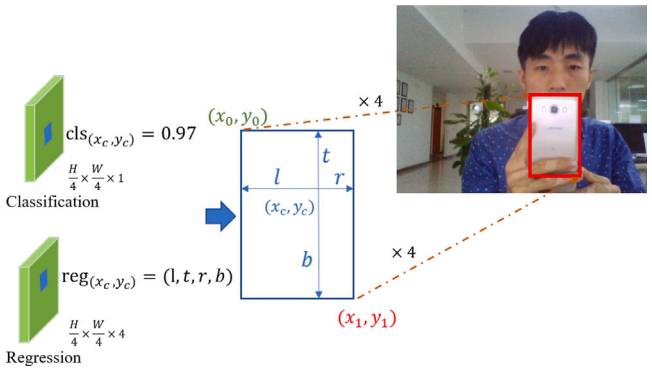


Fig. 3. Prediction box decoding. The classification head predicts the probability $cls_{(x_c,y_c)}$ that each pixel is an object center. The regression head predicts the distance $reg_{(x_c,y_c)}$ from this point to the four sides of a detection box. Then enlarge the decoded box four times to get the final results.

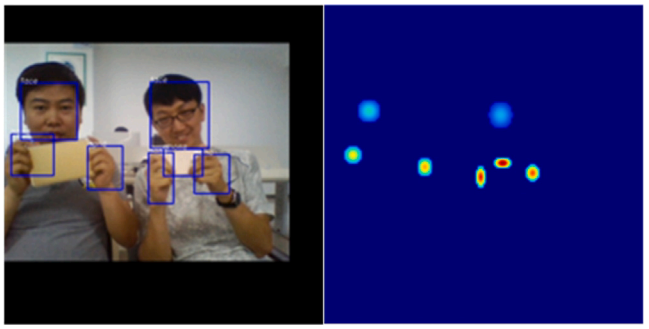


Fig. 4. Ground-truth generation of global location distribution. The first step is to perform data enhancement such as random cropping on the original data. The second step is to generate Gaussian blur based on the enhanced ground-truths and filter out positive samples with low objectivity.

obtain high-level semantic information and low-level detailed information at the same time after FPN fusion, which are more suitable for object detection.

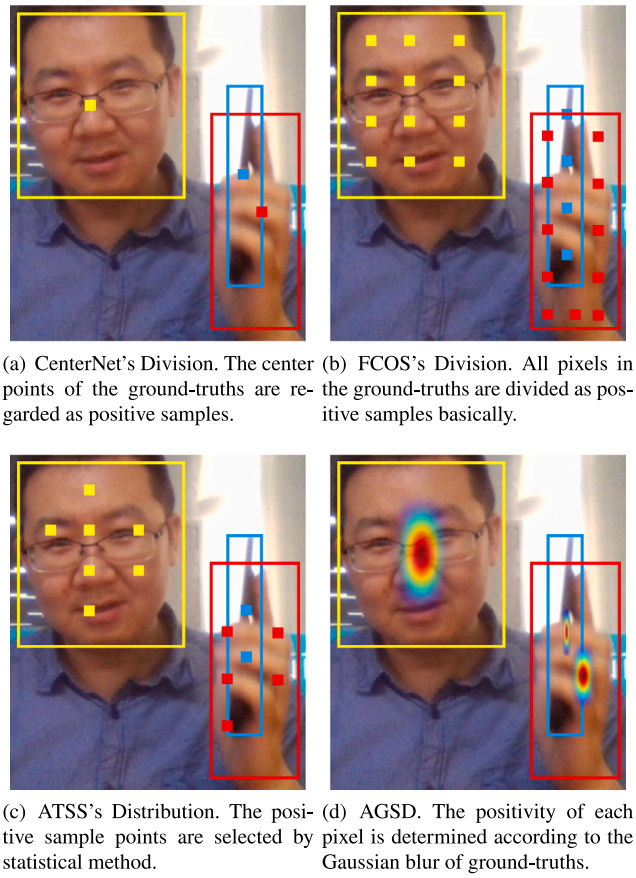
(2) Detection heads

The detection part of the detector is composed of three detection heads. The classification head performs binary classification at each pixel, predicting the possibility of object center. The regression head predicts a four-dimensional vector (l,t,r,b) for each pixel (x_c, y_c). The GLD head predicts the global position distribution of the targets in the overlapping scenes. We can decode the prediction box for each pixel with the results of classification head and regression head as shown in Fig. 3, for each pixel, predict the probability that it is an object center $cls_{(x_c,y_c)}$ (for example 0.97), and the distance between the point and the four sides of the detection box $reg_{(x_c,y_c)}$. Combine it into a detection box, and then enlarge its coordinates four times as the final detection result. (x_0, y_0) is the coordinate of the upper left corner of the prediction box, and (x_1, y_1) is the coordinate of the lower right corner of the prediction box.

(3) Global Location Distribution head (GLD head)

In order to enable the detector to learn the global location distribution between overlapping objects, we design the GLD head. The detector can obtain the prediction of the distribution relationship in the image by GLD head.

The ground-truth of global location distribution is generated as shown in Fig. 4. We firstly perform data augmentation such as random cropping and zooming on the image. Then the Gaussian pixel distribution of the object is generated according to the shape of the object box. In order to reduce the overlap of overlapping objects' distributions, we use 0.4 as a threshold to clarify the main distribution location of the object.



(a) CenterNet's Division. The center points of the ground-truths are regarded as positive samples.

(b) FCOS's Division. All pixels in the ground-truths are divided as positive samples basically.

(c) ATSS's Distribution. The positive sample points are selected by statistical method.

(d) AGSD. The positivity of each pixel is determined according to the Gaussian blur of ground-truths.

Fig. 5. Comparison of AGSD and sample division methods.

3.3. Adaptive Gaussian sample division (AGSD)

In this section, we will introduce AGSD. First of all, we improve the rough heatmap generation method in CenterNet [5]. The new method proposed can adaptively generate heatmaps according to the shape of the object boxes, instead of the circular heatmaps. Take the face heatmap as an example, assuming that the center point of face boxes on the original image is (x_{c0}, y_{c0}). As we all know, the expression of two-dimensional Gaussian function $G(x, y)$ is:

$$G(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} \exp\left(-\frac{(x-x_{c0})^2}{2\sigma_x^2} - \frac{(y-y_{c0})^2}{2\sigma_y^2}\right) \tag{1}$$

where σ_x and σ_y are the standard deviations along x and y directions. The standard deviation of Gaussian distribution is determined by the same calculation method as OpenCV:

$$\sigma_x = 0.3 \times ((w-1) \times 0.5 - 1) + 0.8 \tag{2a}$$

$$\sigma_y = 0.3 \times ((h-1) \times 0.5 - 1) + 0.8 \tag{2b}$$

where w and h are the width and height of box respectively.

The difference between AGSD and other sample divisions in the common detector is shown in Fig. 5. In CenterNet [5], only the center point of each box is regarded as a positive sample, and other points are regarded as negative samples. Since the number of positive samples is very small, the serious imbalance of positive and negative samples brought about affects the performance of CenterNet [5]. In FCOS [6], the samples in the ground-truth boxes are regarded as positive samples, and the points outside are regarded as negative samples basically. And the scale constraint in FPN, the strategy of preferentially dividing pixels in the overlapping area to small objects still cannot effectively deal

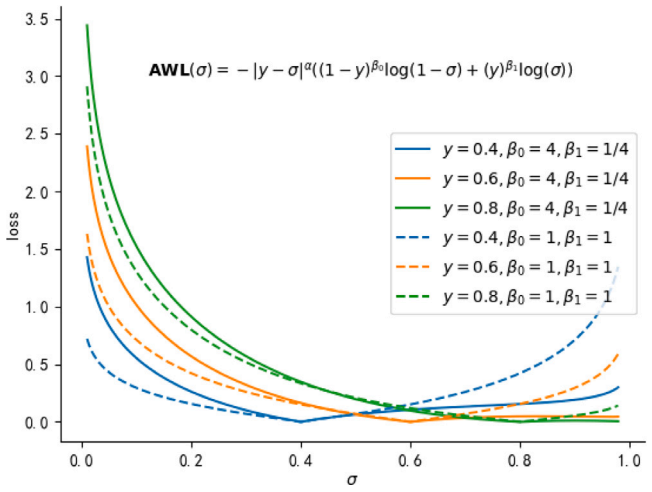


Fig. 6. The comparison of symmetric weighting and asymmetric weighting when $y = 0.4, 0.6, 0.8$. Asymmetric weighting has larger and smaller values when the difficult sample is misdetected ($\sigma < y$) and the simple sample is detected correctly ($\sigma > y$), so as to better deal with the imbalance of difficult and easy samples.

with the semantic conflicts in highly overlapping regions. ATSS [9] introduces statistical analysis into sample division, and selects pixels with higher IoU as the final positive sample. But like FCOS [6], the semantic conflicts of pixels in the overlapping area still cannot be handled properly.

As shown in Fig. 5(d), AGSD establishes a Gaussian distribution based on the shape of the object, thereby effectively avoiding sample conflicts in the overlapping regions.

The objectivity of an object usually decays from center to surroundings. Only a box center is divided into positive samples, it ignores the effect of other pixels around the center. At the same time, classification branch of the detector regards points around the center as negative samples. This increases the difficulty of classification branch training. And in the four corners of a ground-truth box, the object is basically close to the background. By establishing a Gaussian distribution according to shape of objects, AGSD can avoid semantic conflicts, and assign appropriate weights to each positive sample according to their Gaussian value.

3.4. Loss function

(1) Asymmetric Weighted Loss (AW Loss)

Focal loss [10] solves the problem of imbalance of positive and negative samples and the imbalance of difficult and easy samples in one-stage object detection. It reduces the weight of a large number of simple negative samples in training. Focal loss is defined as:

$$\mathbf{FL}(p) = \begin{cases} -(1-p)^{\alpha}\log(p) & \text{when } y = 1 \\ -(p)^{\alpha}\log(1-p) & \text{when } y = 0 \end{cases} \tag{3}$$

where p is the predicted probability of the classification branch that the pixel may be the object center. y indicates whether the pixel is the center of the object. α is a hyperparameter, generally set to 2.

With AGSD, pixels with high objectivity can be divided into positive samples and negative samples. And y value is continuously converted to decimals between 0 and 1 according to the objectivity. But positive pixels are treated equally during calculation of classification loss, the objectivity difference between positive samples is crucial but will be overwhelmed. As shown in Fig. 5, the objectivity of a center point is much higher than the positive samples at the heatmap edges. To use the objectivity difference among positive samples, we weight the samples according to the value of the heatmap, and propose AW Loss:

$$\mathbf{AWL}(\sigma) = -|y-\sigma|^{\alpha}((1-y)^{\beta}\log(1-\sigma) + (y)^{1/\beta}\log(\sigma)) \tag{4}$$

Table 1
The statistics of the overlapping object detection dataset.

	Small	Medium	Large	Total
Face	89	1661	6334	8084
Hand	25	3531	3514	7070
Phone	128	1535	908	2571
Total	242	6727	10 756	17 725

AW loss can calculate the classification loss for the object that is regarded as a Gaussian distribution. Where y is the ground-truth value of object Gaussian distribution. The closer the sample is to the object center, the stronger the objectivity of the samples, and the larger the corresponding y value is. σ is the predicted value for the object distribution. α is a hyperparameter with a value of 2, which has the same function as α in Focal loss [10]. β is a hyperparameter with a value of 4 used to asymmetrically weight the positive sample loss and the negative sample loss. For the imbalance of positive and negative samples, the weight of the positive samples is $(y)^{1/\beta}$, and the positive samples is mapped to get the higher weight. The weight of the negative sample is $(1-y)^{\beta}$, and the weight obtained by $(y)^{1/\beta}$ mapping is smaller to reduce the influence of negative samples. For the imbalance of difficult and easy samples, we compared the values of AWL in the asymmetric weighting and symmetric weighting at $y = 0.4, 0.5, 0.6$. As shown in Fig. 6, when the predicted value is lower than the y value (difficult sample is misdetected), the asymmetric weighted classification loss is greater. When the predicted value is higher than the y value (simple sample is detected correctly), the classification loss of asymmetric weighting is smaller than that of symmetric weighting. AWL can more effectively balance the imbalance of difficult and easy samples.

Through asymmetric weighting, AW loss can perform reasonable loss calculations on Gaussian distributed samples, and guide the detector to focus on the positive samples with high objectivity, thereby enhancing the detector's classification ability in the overlapping object detection tasks.

(2) Total Loss Function

We define total loss function L as follows:

$$L = L_{cls} + L_{reg} + L_{dis} \tag{5}$$

where L_{cls} is AW Loss, which is used to ease the imbalance of positive and negative samples of the dataset. L_{reg} uses L1 Loss, which is used to guide the detector's focus in different processes. For GLD head, we use Focal Loss [10] to measure the errors in distribution prediction.

4. Dataset

The dataset images are collected in real scenarios. There are a total of 6859 images with 17,725 annotations in our dataset, which share the same resolution of 640 × 480.

As shown in Fig. 7, the data of the handheld mobile phone and its corresponding deceptive actions are all included in the dataset. The three classes of objects such as mobile phones, hands, and faces often overlap, which poses a challenge to the pixel-by-pixel classification capability and sample division in the anchor-free detector. And the collected persons have various movements and the shape of the hands are varied. Medium and large goals occupy the majority. We place 6173 images (90%) into the training set and val set, and the remaining 686 images (10%) into the testing set. The statistics of the dataset is shown in Table 1.

The number of annotations for hands, faces and mobile phones is 8084, 7070, and 2571, respectively. Relatively speaking, there are fewer labels for hands, and there is a certain class imbalance. Furthermore, unlike the COCO dataset [1] that is dominated by small objects, the number of small, medium, and large objects in our dataset

Table 2
Performance comparison with mainstream detectors.

Model	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN [33]	ResNet50	65.7	95.5	74.5	17.7	59.6	72.2
SSD [31]	ResNet50	64.0	95.6	71.7	20.4	56.9	71.4
YOLO V3 [30]	DarkNet53	65.3	95.3	74.5	25.7	57.7	71.9
FASF [42]	ResNet50	65.40	95.1	73.5	14.4	58.6	72.1
free-anchor [43]	ResNet50	64.3	95.8	72.9	29.7	58.0	69.9
CenterNet [5]	ResNet50	64.7	96.2	72.3	28.8	56.8	73.2
Ours	ResNet50	66.09	96.71	75.13	21.14	58.53	72.91

[1] RED/BLUE indicate SOTA/the second best.

Fig. 7. Dataset contains overlapping object scenes and corresponding deception scenes with various person poses.

is 242, 6727, and 10,757, respectively. Medium objects, especially large ones, occupy the vast majority of the dataset. Most of the current detectors are evaluated and compared on the COCO dataset [1] with small objects, which means that some optimization methods are no longer suitable for the overlapping object detection task in this paper.

5. Experiments

To show the effectiveness of the proposed overlapping object detector, we make comprehensive comparisons between our approach with mainstream deep learning methods such as YOLOv3 [30], FASF [42], etc. Then we do ablation experiments to explore the effectiveness of AGSD, GLD head, and AW Loss for the detector.

5.1. Experiment setup

The ratio of training set: validation set: test set is 7:2:1. We use ResNet50 [41] pre-trained on ImageNet as the backbone of the detector for feature extraction. The model uses the Adam optimizer for training with 32 images per batch. The initial learning rate is 1.25×10^{-4}, and a total of 70 epochs are trained. At the 45th and 60th epoch, the learning rate is attenuated to the original 0.1. All experiments were performed on a device containing two 2080TI and CPU E5-2620 v4 @ 2.10 GHz. Similar to most publications, we use the average precision (AP) to evaluate the performance of our method. AP is the area of the precision/recall curve.

$$AP = \int_0^1 P(R)dP \tag{6}$$

where P is precision and R is recall. $P(R)$ is a function with R as its independent variable.

5.2. Comparison with mainstream detectors

We compare the proposed method with mainstream object detectors: (1) Anchor-based one-stage object detectors: YOLOv3 [30], SSD [31], freeanchor [43] and FASF [42]. (2) Anchor-based two-stage object detectors: Faster R-CNN [33]. (3) Anchor-free object detectors: CenterNet [5]. The feature extraction network adopts ResNet50 (the feature extraction network of YOLOv3 [30] is darknet53).

As shown in Table 2: Anchor-based object detectors such as SSD [31], Faster R-CNN [33] and YOLO v3 [30] achieve AP_{50} of 95.6%, 95.5% and 95.3%. Anchor-free object detectors like CenterNet [5] achieve AP_{50} of 96.2%. Our method reaches 96.7% AP_{50}. It is the best performance on severely overlapping object detection tasks.

The final detection result is shown in Fig. 8, the green boxes are the ground-truths, and the red ones are the predictions. It can be seen that the detector can accurately detect objects. In the first and second columns of Fig. 8, the location distribution of objects is relatively simple and scattered. objects like faces, hands, mobile phones all achieve high-confidence. In the third and fourth columns of Fig. 8, the distribution among objects is complex, and there is a serious overlap. Our method obtains effective sample division and object location distribution, and achieves high-performance detection for overlapping objects.

5.3. Ablation experiments

AGSD reasonably divides the samples into the overlapping area. GLD head learns the global location distribution of the objects, and AW Loss enhances the classification ability of the anchor-free detector. In order to analyze their influence on the detector, we conduct a series of ablation experiments under the same experimental settings.

All ablation experiments are briefly reported in Table 3. Since the number of small objects in the dataset is small, the performance mainly depends on the performance of the overlapping object detector on medium and large objects. Compared with the baseline, the performance of AGSD and AW loss has increased by 0.82% and 0.83% on medium and large objects (AP_M and AP_L), respectively. When the two are used at the same time, the AP increases from 64.72% to 65.55%. GLD head can give the detector an additional ability to learn the global location distribution of overlapping objects, which can further improve the performance of the detector from 65.55% to 66.09%.

(1) Ablation Study for GLD head

We compare the performance of detectors with and without GLD head. Experimental results are in Table 4. We see that GLD gains 0.54% and 0.66% improvement on AP and AP_{50}. The ability to learn the global location distribution among overlapping objects allows the detector to better align medium objects and large objects, finally improving 0.39% and 0.49% in AP_M and AP_L, respectively.

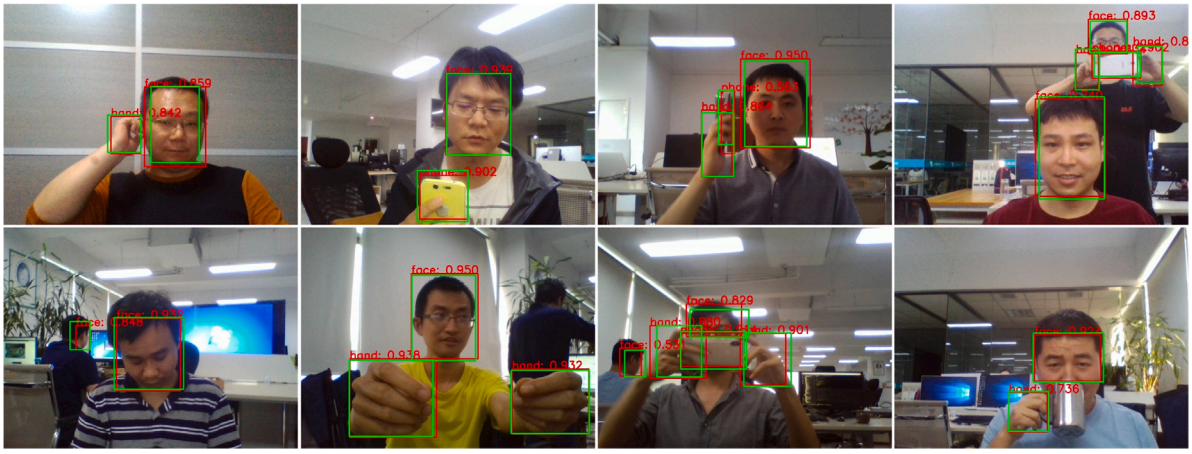


Fig. 8. The green boxes are ground-truths, and the red ones are the detection results. We realize the high-precision detection of discretely distributed (1st and 2nd column) and overlapping (3rd and 4th column) objects.

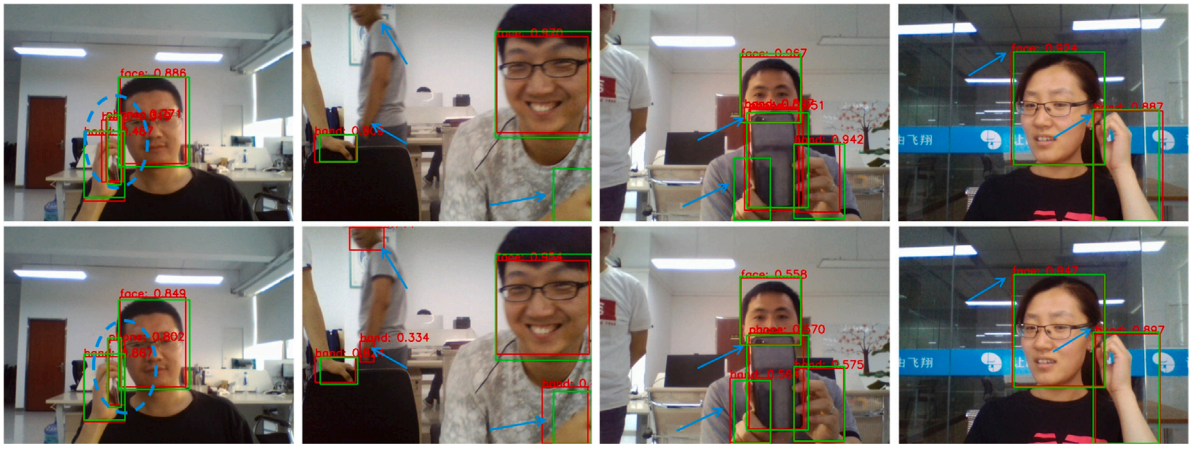


Fig. 9. Comparison of results with and without GLD. The first and second rows are the results without GLD head and with GLD head. The green boxes and red ones are the ground-truths and predictions.

Table 3
Ablation experiments.

AGSD	AW Loss	GLD	AP	AP_S	AP_M	AP_L
			64.72	25.81	56.81	73.21
✓			65.25	25.28	57.63	71.90
	✓		65.27	**25.82**	56.07	74.04
✓	✓		65.55	23.64	58.14	72.42
✓	✓	✓	**66.09**	21.14	**58.53**	72.91

Table 4
Ablation study for GLD head.

	AP	AP_{50}	AP_S	AP_M	AP_L
AGSD+AWL	65.55	96.05	**23.64**	58.14	72.42
AGSD+AWL+GLD	**66.09**	**96.71**	21.14	**58.53**	**72.91**

We also made a comparison before and after the addition of GLD head in Fig. 9. The first and second rows are the results without GLD head and with GLD head. The green and red boxes are the ground-truths and predictions. As shown in the first column of comparisons, the high overlap between the mobile phone and the hand makes the location distribution of the object complicated. GLD head can accurately separate the hand from the mobile phone and suppress the false detection box near the mobile phone detection box. Comparing the second and third groups of images, GLD head can perform a better overview of the overall global location distribution, so that the hands that were previously missed are detected. Comparing the last set of images, after adding GLD head, the detector can be more confident when predicting the object category. The confidence of the face and hand is compared with the previous 0.924 and 0.887 increased to 0.947 and 0.897. GLD head can increase the learning ability of the global location distribution among the objects for the detector, and also has a certain auxiliary improvement effect for the classification task of the detector.

(2) Influence of AGSD

To alleviate the problem of pixel semantic conflicts during sample division of overlapping objects, we design AGSD to adaptively generate Gaussian heatmaps for overlapping objects according to the shapes of their boxes to divide positive and negative samples. In the ablation experiment, we compare AGSD with the classic anchor-free detector FCOS [6] and CenterNet [5] sample division methods. FCOS [6] divides all pixels in the objects boxes into positive samples, and then uses the manually set scale information to filter out some low-quality samples. CenterNet [5] only regards the center point of the object as the positive sample, and the remaining pixels are regarded as negative samples. The results of the experiment are shown in the Table 5:

It can be seen that treating the pixels in the ground-truth box as a positive sample performs poorly in the overlapping object detection, with an AP of 56.69%. Treating the center point of the object as a positive sample can avoid semantic conflicts in overlapping areas, but

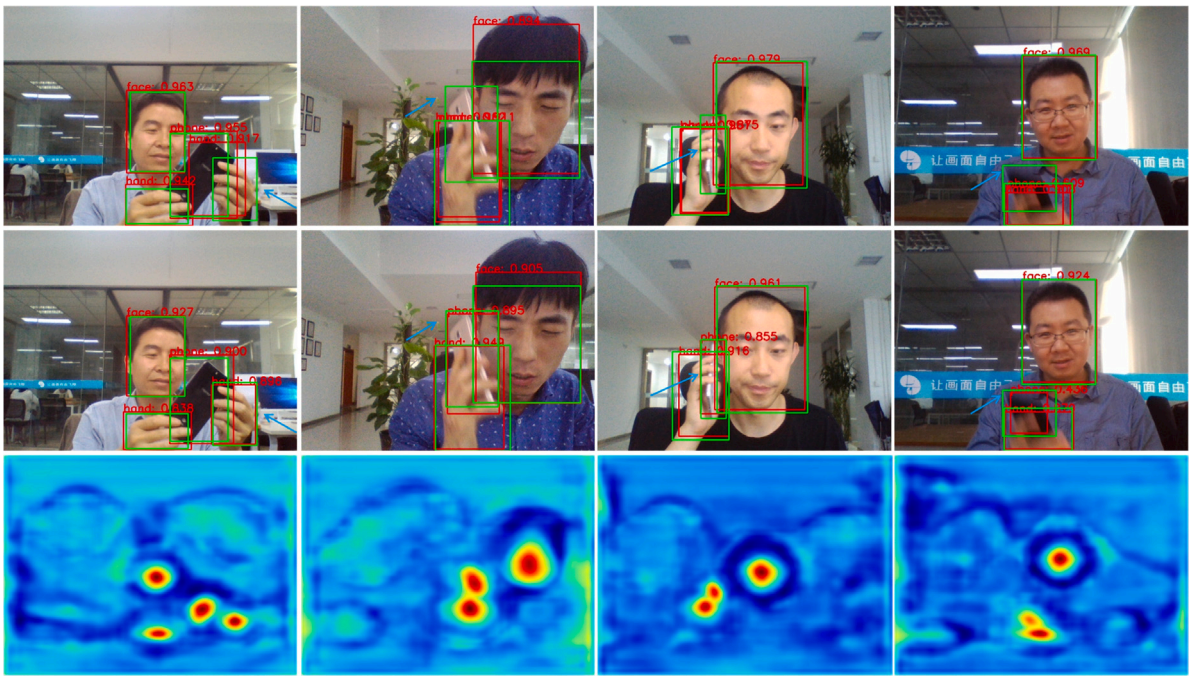


Fig. 10. Comparison of FCOS's division and AGSD detection results. The first and second rows are the results of FCOS's division and AGSD's division. The green boxes and red ones are the ground-truths and predictions. The third row is the visualization of AGSD Gaussian Distribution Prediction Results.

Table 5
Ablation study for AGSD.

	AP	AP_{50}	AP_S	AP_M	AP_L
FCOS [6]	56.69	84.44	25.38	47.74	64.72
CenterNet [5]	64.72	96.23	**25.82**	56.81	**73.21**
AGSD	**65.25**	**96.74**	25.28	**57.63**	71.9

Table 6
Ablation study for AW Loss.

	AP	AP_{50}	AP_S	AP_M	AP_L
Focal Loss	64.72	96.23	**28.83**	56.81	73.21
AWL	65.27	**96.83**	25.81	56.07	**74.04**
AGSD+AWL	**65.55**	96.05	23.64	**58.14**	72.42

the performance is limited by the more uneven ratio of positive and negative samples, the AP is 64.72%. While AGSD effectively alleviates the semantic conflict problem of the division of positive and negative samples, the ratio of positive and negative samples is also more balanced, with an AP of 65.25%. Compared with the sample division methods of FCOS [6] and CenterNet [5], the performance of AGSD exceeds 8.56% and 0.53%, respectively. In terms of small and large objects, AGSD achieves similar performance to CenterNet [5]. In terms of the medium object that dominates the dataset, AGSD achieves a key increase of 0.82% compared to CenterNet [5].

Furthermore, we visualize the experimental results of two sample division methods: FCOS [6] and AGSD. As shown in Fig. 10, the first and second rows are the results of FCOS's division and AGSD. The green and red boxes are the ground-truth and detection results. As shown in the first column, although the FCOS' division method detects the right hand, the positioning accuracy is far less than that of AGSD. In the 2nd, 3rd, 4th columns, mobile phone overlaps hand seriously, and only AGSD can accurately identify correct detection boxes.

The third row shows prediction results of the detector for object distribution. We see that AGSD can divide the samples in the overlapping area, thereby alleviating the semantic conflict. The central area of the overlapping objects can be effectively distinguished by the detector.

(3) Ablation Study for AW Loss

After concatenating the positive and negative samples to the range of [0,1], we design AW Loss for classification head. AW loss uses asymmetric exponential weights for positive and negative samples, which can make the proportion of positive sample loss and negative sample loss more balanced. In the ablation experiment, we compare the performance of AW Loss and Focal Loss [10]. The experimental results are shown in the Table 6.

Comparing Focal loss [10] and AW Loss, we can see that the AP_{50} of AW Loss is 0.60% higher than that of FL 96.23%. At the same time, AW Loss has also made significant improvements in the detection of large objects, increasing the AP_L of Focal loss [10] from 73.21% to 74.04%. Furthermore, AW Loss and AGSD have non-conflicting improvements to overlapping objects. AW Loss can also significantly improve the performance of AP_{50} and AP_L on the basis of AGSD.

6. Conclusion

We develop an anchor-free method for severely overlapping object detection. Extensive experiments have shown the effectiveness of our method. The proposed AGSD effectively divides positive and negative samples into overlapping areas, and alleviates the semantic conflict. The original discrete positive and negative samples are taken to continuous [0,1] interval. AW Loss improves the classification ability of the detector with a more reasonable classification loss. GLD head adds the ability to learn the complex global location distribution for the detector. We hope that these highlights may be useful for other overlapping object tasks.

CRediT authorship contribution statement

Yao Xue: Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. **Yawei Zhang:** Methodology, Writing – original draft. **Yuxiao Liu:** Writing – review & editing. **Xueming Qian:** Funding acquisition, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yao Xue reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[2] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes challenge 2012 (VOC2012) results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[3] X. Wang, T. Xiao, Y. Jiang, S. Shuai, C. Shen, Repulsion loss: detecting pedestrians in a crowd, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[4] Z. Luo, Z. Fang, S. Zheng, Y. Wang, Y. Fu, NMS-loss: learning with non-maximum suppression for crowded pedestrian detection, 2021, arXiv preprint arXiv:2106.02426.

[5] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Object detection with keypoint triplets, 1, (2) 2019, p. 4, arXiv preprint arXiv:1904.08189.

[6] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9627–9636.

[7] X. Li, S. Lai, X. Qian, DBCFace: towards pure convolutional neural network face detection, IEEE Trans. Circuits Syst. Video Technol. (2021).

[8] T. Ma, W. Tian, P. Kuang, Y. Xie, An anchor-free object detector with novel corner matching method, Knowl.-Based Syst. 224 (2021) 107083, http://dx.doi.org/10.1016/j.knosys.2021.107083.

[9] S. Zhang, C. Chi, Y. Yao, Z. Lei, S.Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[11] C. Feng, Y. Zhong, M.R. Scott, W. Huang, TOOD: task-aligned one-stage object detection, 2021, arXiv preprint arXiv:2108.07755.

[12] C. Wei, K. Sohn, C. Mellina, A. Yuille, F. Yang, Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10857–10866.

[13] Y. Yang, Z. Xu, Rethinking the value of labels for improving class-imbalanced learning, 2020, arXiv preprint arXiv:2006.07529.

[14] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9719–9728.

[15] K. Napierala, J. Stefanowski, Types of minority class examples and their influence on learning classifiers from imbalanced data, J. Intell. Inf. Syst. 46 (3) (2016) 563–597.

[16] H. Zhu, P. Tang, J. Park, S. Park, A. Yuille, Robustness of object recognition under extreme occlusion in humans and computational models, 2019, arXiv preprint arXiv:1905.04598.

[17] R. Jin, G. Lin, C. Wen, Online active proposal set generation for weakly supervised object detection, Knowl.-Based Syst. (2021) 107726, http://dx.doi.org/10.1016/j.knosys.2021.107726.

[18] X. Wei, S. Liu, Y. Xiang, Z. Duan, C. Zhao, Y. Lu, Incremental learning based multi-domain adaptation for object detection, Knowl.-Based Syst. 210 (2020) 106420, http://dx.doi.org/10.1016/j.knosys.2020.106420.

[19] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, F. Herrera, Object Detection Binary Classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance, Knowl.-Based Syst. 194 (2020) 105590, http://dx.doi.org/10.1016/j.knosys.2020.105590.

[20] X. Chu, A. Zheng, X. Zhang, J. Sun, Detection in crowded scenes: one proposal, multiple predictions, 2020, arXiv:2003.09163.

[21] S. Zhang, L. Wen, X. Bian, Z. Lei, S.Z. Li, Occlusion-aware R-CNN: Detecting pedestrians in a crowd, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 637–653.

[22] Y. Tian, P. Luo, X. Wang, X. Tang, Deep learning strong parts for pedestrian detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1904–1912.

[23] S. Gilroy, E. Jones, M. Glavin, Overcoming occlusion in the automotive environment—A review, IEEE Trans. Intell. Transp. Syst. 22 (1) (2019) 23–35.

[24] C. Ning, L. Menglu, Y. Hao, S. Xueping, L. Yunhong, Survey of pedestrian detection with occlusion, Complex Intell. Syst. 7 (1) (2021) 577–587.

[25] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, T. Hassner, Precise detection in densely packed scenes, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 5222–5231, http://dx.doi.org/10.1109/CVPR.2019.00537.

[26] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, C. Ma, C. Xu, Dynamic refinement network for oriented and densely packed object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11207–11216.

[27] S. Kant, Learning Gaussian maps for dense object detection, 2020, arXiv preprint arXiv:2004.11855.

[28] Y. Cai, L. Wen, L. Zhang, D. Du, W. Wang, Rethinking object detection in retail stores, 2020, arXiv preprint arXiv:2003.08230.

[29] Q. Xiong, J. Lin, W. Yue, S. Liu, C. Ding, A deep learning approach to driver distraction detection of using mobile phone, in: 2019 IEEE Vehicle Power and Propulsion Conference, VPPC, 2019.

[30] A. Farhadi, J. Redmon, Yolov3: An incremental improvement, in: Computer Vision and Pattern Recognition, 2018, cite as.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

[32] X. Zhou, J. Zhuo, P. Krahenbuhl, Bottom-up object detection by grouping extreme and center points, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[33] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[34] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: exceeding YOLO series in 2021, 2021, CoRR abs/2107.08430, arXiv:2107.08430.

[35] Y. Qin, J. Wen, H. Zheng, X. Huang, J. Yang, N. Song, Y.-M. Zhu, L. Wu, G.-Z. Yang, Varifocal-net: A chromosome classification approach using deep convolutional networks, IEEE Trans. Med. Imaging 38 (11) (2019) 2569–2581.

[36] C. Liu, Y. Liang, Y. Xue, X. Qian, J. Fu, Food and ingredient joint learning for fine-grained recognition, IEEE Trans. Circuits Syst. Video Technol. (2020).

[37] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, L. Zhang, Dynamic head: unifying object detection heads with attentions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7373–7382.

[38] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[39] X. Li, S. Lai, X. Qian, DBCFace: towards pure convolutional neural network face detection, IEEE Trans. Circuits Syst. Video Technol. (2021) 1, http://dx.doi.org/10.1109/TCSVT.2021.3082635.

[40] X. Wang, L. Zhu, H. Wang, Y. Yang, Interactive prototype learning for egocentric action recognition, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 8148–8157.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[42] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 840–849.

[43] X. Zhang, F. Wan, C. Liu, R. Ji, Q. Ye, Freeanchor: Learning to match anchors for visual object detection, in: Advances in Neural Information Processing Systems, 2019, pp. 147–155.