DEEP OBJECT DETECTION WITH EXAMPLE ATTRIBUTE BASED PREDICTION MODULATION

Zhihao Wu[1,2] *Chengliang Liu*[1,2] *Chao Huang*[1,2,3] *Jie Wen*[1,2] *Yong Xu*[⋆1,2,3]

[1] Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen
[2] Shenzhen Key Laboratory of Visual Object Detection and Recognition [3] Peng Cheng Laboratory

ABSTRACT

Deep object detectors suffer from the gradient contribution imbalance during training. In this paper, we point out that such imbalance can be ascribed to the imbalance in example attributes, e.g., difficulty and shape variation degree. We further propose example attribute based prediction modulation (EAPM) to address it. In EAPM, first, the attribute of an example is defined by the prediction and the corresponding ground truth. Then, a modulating factor w.r.t the example attribute is introduced to modulate the prediction error. Finally, the new prediction and the ground-truth are input into the loss function. Essentially, we adjust the gradients of examples with specific attributes to reweight their contribution on the global gradients. We apply EAPM with focal loss and balanced L1 loss to simultaneously solve the imbalance in classification and localization. The experimental results on MS COCO demonstrate that EAPM can bring substantial improvement for deep object detectors.

Index Terms— Deep object detection, gradient contribution imbalance, example attribute, classification, localization

1. INTRODUCTION

Object detection is a fundamental topic in the computer vision field. Recently, deep object detectors, including two-stage [1, 2, 3, 4, 5] and one-stage [6, 7, 8, 9], have achieved great success.

Both two families of the detectors usually have a common training process, i.e., joint classification and box regression under the guidance of a multi-task loss function. However, they face the gradient contribution imbalance during this typical training process. Some existing researches [10, 6, 11] ascribe such imbalance to the imbalance in example difficulty. For example, one-stage detectors suffer from the extreme imbalance problem caused by background examples during the classification training [6, 12, 13]. Due to the dense sampling of object candidates, a huge number of negative examples in the background dominate the contribution on the global gradients. Therefore, the negatives tend to be easily classified. However, the easy negatives make nearly no contribution to the precision because of useless learning signal. Another example of the imbalance is caused by outliers. During the classification training of one-stage detectors, some very hard examples still exist even when the model converges, so they are regarded as outliers [11]. The gradients of the outliers are larger than those of the other examples (inliers) when cross entropy loss is used. However, the outliers will degenerate the model if the parameters are tuned to fit them well. Previous studies [11, 2] also point out that outliers contribute much in the box regression training. To handle the issues, some previous works propose hard example mining [14, 15, 10] or novel loss functions [6, 16, 11, 2, 17, 18, 19, 20, 21] to change the number and the gradients of examples respectively.

We observe that the imbalance problem also appears in the box regression training of anchor-based detectors, which is caused by severe shape variations of examples. Anchor-based detectors are difficult to handle candidates with large shape variations owing to fixed scales and aspect ratios of anchor boxes [9]. Furthermore, the loss functions [1, 2] cannot distinguish the gradients of examples with different shape variations. In other words, these loss functions ignore the contribution compensation of examples with large shape variations as

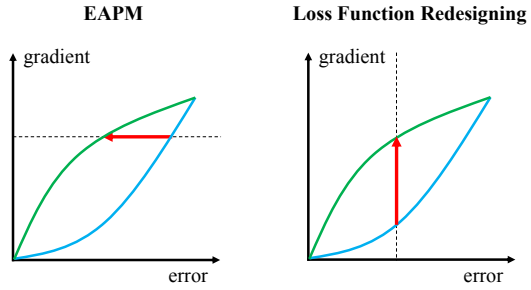

Fig. 1: The difference between loss function redesigning and EAPM. The blue curve represents the original gradient, the green curve represents the new one, and the red arrow represents the idea of modulation.

⋆ Corresponding Author. This work is partially supported by National Key Research and Development Program of China under Project Number 2018AAA0100100, and Establishment of Key Laboratory of Shenzhen Science and Technology Innovation Committee under Grant ZDSYS20190902093015527.

a hedging for the detector shortcoming. In this paper, we ascribe this imbalance and the imbalance in example difficulty to the imbalance in example attributes. To mitigate such imbalance, we propose example attribute based prediction modulation (EAPM). In EAPM, first, the attribute is defined by the prediction and the corresponding ground-truth. Next, a novel modulating factor is introduced to adjust the prediction error according to the attribute. Finally, the new prediction and the corresponding ground-truth are taken as inputs of loss functions. Similar to loss function redesigning, the nature behind our strategy is to reformulate the gradients of examples according to their attributes, thereby adjusting their contribution on the global gradients. However, we do not need to design new gradients but only use the original ones by changing errors (see Fig. 1). We apply EAPM with focal loss and balanced L1 loss to simultaneously solve the example attribute imbalance in classification and localization. The experimental results on MS COCO demonstrate that EAPM can bring substantial improvement for deep object detection.

2. METHOD

The training pipeline of deep object detectors with EAPM is shown in Fig. 2. Our goal is to simultaneously address the example attribute imbalance in classification and regression by EAPM with loss functions.

2.1. EAPM

For the training of deep object detectors, a multi-task loss function L is designed as follows for classification and box regression, where the loss is computed for each example: $L_{p,p^*,t,t^*} = L_{cls}\left(p,p^*\right) + \lambda p^* L_{reg}\left(t,t^*\right)$. L_{cls} and L_{reg} are loss functions for classification and box regression respectively. $p \in [0,1]$ and $p^* \in \{0,1\}$ denote the predicted probability of an example being from a certain class and the corresponding label. t is the regression result, and t^* is its target. The term $p^* L_{reg}$ means that the regression loss is activated only for positives ($p^* = 1$). λ is used for balancing the loss weight in the multi-task learning.

We ascribe the gradient contribution imbalance to the imbalance in example attribute. We further propose EAPM to modify the gradients by example attributes associated prediction modulation. EAPM is detailed as follows.

First, two descriptors, EA_{cls} and EA_{reg}, are introduced to denote specific attributes of examples in classification and regression respectively. These two descriptors are defined as: $EA_{cls} = f_1\left(p,p^*\right), EA_{reg} = f_2\left(t,t^*\right)$, where f_1 and f_2 are functions designed according to specific example attributes. Actually, they are general extensions to example difficulty in the previous works [6, 2]. When $f_1 = |p - p^*|$ and $f_2 = |t - t^*|$, f_1 and f_2 represent the difficulty.

Next, two corresponding modulating factors, MF_{cls} and MF_{reg}, are designed to adjust the prediction error.

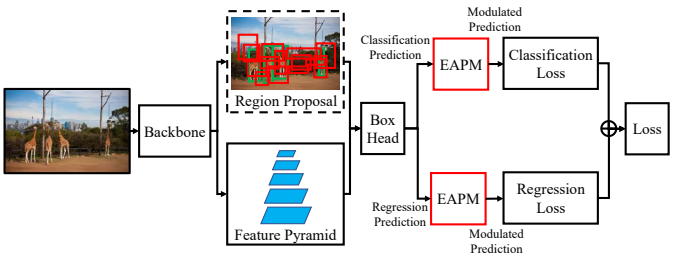

Fig. 2: The training pipeline of deep object detectors with EAPM. The dotted bordered rectangle denotes that one-stage detectors do not have the region proposal stage, but two-stage detectors have.

The two modulating factors are designed as: $MF_{cls} = g_1\left(EA_{cls}\right), MF_{reg} = g_2\left(EA_{reg}\right)$, where g_1 and g_2 are functions that determine modulation degrees of examples with different attribute values. Note that they can be altered flexibly for various deep detectors and datasets. Then, the predictions are updated as: $p' = p^* + MF_{cls} \times \left(p - p^*\right), t' = t^* + MF_{reg} \times \left(t - t^*\right)$, where $p - p^*$ and $t - t^*$ denote the prediction errors of classification and regression respectively. p' and t' are the corresponding updated predictions.

Finally, the loss functions, L_{cls} and L_{reg}, are recalculated by inputting the new predictions and the targets. This step can be summarized as: $L_{p',p^*,t',t^*} = L_{cls}\left(p',p^*\right) + \lambda p^* L_{reg}\left(t',t^*\right)$. So the gradients are reformulated as well.

We can see how EAPM with loss functions reformulates the gradients of examples according to their attribute values, thereby changing the contribution on the overall gradients. In particular, let x be the direct network output for classification and $p = \text{sigmoid}\left(x\right)$. $\frac{\partial L_{cls}}{\partial x}\Big|_{x=x_0}$ is adjusted to $\frac{\partial L_{cls}}{\partial x}\Big|_{x=x'_0}$, where $x_0 = -\log\left(1/p_0 - 1\right)$ and $x'_0 = -\log\left(1/p'_0 - 1\right) = -\log\left(1/(p^*_0 + MF_{cls} \times (p_0 - p^*_0)) - 1\right)$. Also, $\frac{\partial L_{reg}}{\partial t}\Big|_{t=t_0}$ is tuned to $\frac{\partial L_{reg}}{\partial t}\Big|_{t=t'_0}$, where $t'_0 = t^*_0 + MF_{reg} \times (t_0 - t^*_0)$. In this way, we can easily use the original gradients instead of designing new ones. It is worth noting that it requires the loss functions to assign different gradients to samples with different errors.

2.2. EAPM with Focal Loss

We apply EAPM with focal loss in the classification training of one-stage detectors. Focal loss is defined as:

$$L_{FL}\left(p,p^*\right) = \begin{cases} -(1-p)^{\gamma}\log\left(p\right) & if\, p^* = 1 \\ -p^{\gamma}\log\left(1-p\right) & if\, p^* = 0. \end{cases} \quad (1)$$

To apply EAPM, we first introduce Dif as EA_{cls} to define example difficulty: $Dif = |p - p^*|$. Dif equals to the absolute value of probability prediction error. Thus a larger Dif denotes that the example is harder.

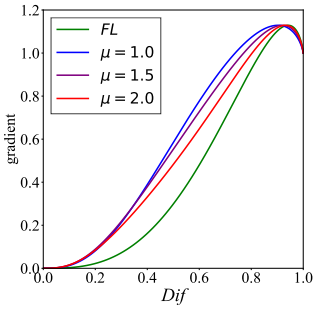

Fig. 3: The gradient of EAPM with focal loss.

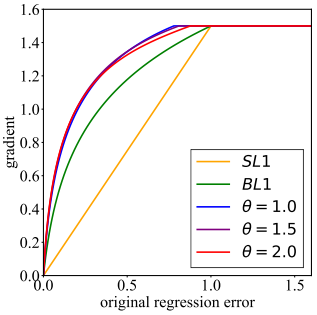

Fig. 4: The gradient of EAPM with balanced L1 Loss.

Then, we define MF_{cls} as: $MF_{cls} = 1 + \tanh(\mu) - \tanh(\mu Dif)$, where $\tanh$ denotes the hyperbolic tangent function and μ is a positive tunable parameter. The new prediction is calculated by $p' = p^* + MF_{cls} \times (p - p^*)$.

Finally, the new prediction and its label are input into Eq. (1) to reformulate the loss and the corresponding gradient. Fig. 3 shows the gradients of focal loss ($\gamma = 2$) and EAPM with focal loss. For convenience, we adopt their absolute values. We can see two properties of the combination compared with focal loss. (a) When examples are easy or very hard, the gradients are nearly unaffected while the ones of other examples increase. (b) The tunable parameter μ smoothly adjusts the rate at which examples are up-weighted.

Intuitively, EAPM with focal loss increases the gradients of all examples except for easy and very hard ones. This in turn reduces the importance of background examples and outliers. In this way, it avoids these two kinds of examples overwhelming the classification training of one-stage detectors.

2.3. EAPM with Balanced L1 Loss

For anchor-based detector regression, let $t = (t_x, t_y, t_w, t_h)$ be a vector representing the four parameters of a predicted bounding box, in which x, y, w and h denote the horizontal and vertical coordinates of the box's center, its width and height respectively. $t^* = (t_x^*, t_y^*, t_w^*, t_h^*)$ is the corresponding ground-truth, and $d = t - t^*$ represents the error. Balanced L1 loss is defined as:

$$L_b(d) = \begin{cases} \frac{\alpha}{b}(b|d|+1)\log(b|d|+1) - \alpha|d| & if\ |d| < 1 \\ \delta|d| + \frac{\delta}{b} - \alpha & otherwise, \end{cases} \quad (2)$$

where examples with $|d| < 1$ are seen as inliers, α controls their increase of gradients, δ represents the upper bound of regression gradient, and $b = e^{\delta/\alpha} - 1$ is used to ensure that $L_b(d = 1)$ has the equal value for both formulae in Eq. (2).

To perform EAPM, we first introduce Gap-over-Sum (GoS) as EA_{reg} to define the shape variation degree. For box regression, anchor-based detectors usually adopt the parameterizations of the four coordinates following [1]: $t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a, t_w = \log(w/w_a), t_h = \log(h/h_a), t_x^* = (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a, t_w^* = \log(w^*/w_a), t_h^* = \log(h^*/h_a)$, where variables x, x_a and x^* are for the prediction, the anchor and the ground-truth respectively (likewise for y, w and h). We can see that a large absolute value of t_w, t_w^*, t_h or t_h^* denotes a large shape variation relative to w_a or h_a. Based on this observation, we define GoS as: $GoS = \frac{|t_j - t_j^*|}{|t_j| + |t_j^*|}$, where $j \in \{w, h\}$. t_j represents the height or the width of the prediction, and t_j^* is the corresponding ground-truth. We can see that GoS ranges from 0 to 1, and its value decreases as the shape variation degree increases under the same error.

Next, we define MF_{reg} as: $MF_{reg} = 1 + \tanh(\theta) - \tanh(\theta GoS)$, where $\tanh$ denotes the hyperbolic tangent function and θ is a positive tunable parameter. We recalculate the regression prediction as: $t' = t^* + MF_{reg} \times (t - t^*)$.

Finally, the recalculated prediction and its ground-truth are input into Eq. (2) to recalculate the loss and the gradient. The gradients of smooth L1 loss, balanced L1 loss ($\alpha = 0.5$ and $\gamma = 1.5$) and EAPM with balanced L1 loss for examples with large shape variations ($|t_j| + |t_j^*| = 1.5$) are visualized in Fig. 4. Compared with smooth L1 loss, balanced L1 loss increases the gradients of inliers to increase their importance. With EAPM, the gradients of inliers with large shape variations are up-weighted. And the tunable parameter θ determines the increasing rate of their gradients.

We argue that in box regression of anchor-based detectors, examples with large shape variations should be put more focus on. It is difficult for the detectors to deal with these examples, so better predictions of them will more obviously improve the quality of final candidates.

Table 1: Comparison of EAPM with focal loss and the corresponding baselines on MS COCO *val-2017*.

Settings	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Libra RetinaNet	37.6	56.9	40.1	21.7	41.5	49.6
$\mu = 1.0$	37.9	**58.1**	40.2	**22.4**	**41.9**	49.0
$\mu = 1.5$	**38.2**	58.0	**40.7**	**22.4**	41.5	**50.3**
$\mu = 2.0$	38.0	58.0	40.5	22.3	41.6	49.7
RetinaNet	35.6	55.5	38.3	20.0	39.6	46.8
$\mu = 1.0$	37.0	**56.6**	39.5	20.4	40.4	**49.1**
$\mu = 1.5$	**37.1**	**56.6**	**39.6**	20.5	**40.6**	**49.1**
$\mu = 2.0$	36.7	55.9	39.3	**20.7**	40.0	48.2
FCOS	35.7	55.8	38.4	20.2	39.2	47.0
$\mu = 1.0$	36.9	56.7	39.2	21.3	40.7	**48.2**
$\mu = 1.5$	**37.1**	**56.8**	**39.4**	**22.3**	**40.8**	47.8
$\mu = 2.0$	36.8	56.5	39.0	21.0	40.5	47.6

3. EXPERIMENTS

3.1. Dataset and Evaluation Metrics

We evaluate EAPM on MS COCO [22]. We train all models on *train-2017*, report the results of EAPM with focal loss and EAPM with balanced L1 loss on *val-2017*, and report main

results on *test-dev*. All results follow the standard MS COCO metrics which include AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L.

3.2. Implementation Details

We implement all experiments on mmdetection [23]. We train the models on 4 GPUs with 2 images per GPU. All models are trained for 12 epochs with an initial learning rate of 0.005 for Libra RetinaNet [2], RetinaNet [6] and FCOS [9], and 0.01 for Libra R-CNN [2]. The learning rates are decreased by 0.1 at the 9th epoch and again at the 12th epoch. All other hyper-parameters follow the settings in mmdetection.

3.3. Experimental Results

3.3.1. EAPM with Focal Loss

To evaluate the effectiveness of EAPM for one-stage detector classification, we apply EAPM with focal loss on several classic one-stage detectors, including Libra RetinaNet, RetinaNet and FCOS. All baselines use ResNet-50-FPN as the backbone and focal loss for classification. Table 1 shows the results of EAPM with focal loss and the corresponding baselines.

We can see that all baselines have better performance by using EAPM with focal loss. Specifically, EAPM ($\mu = 1.5$) brings gains of 0.6 AP, 1.5 AP and 1.4 AP to Libra RetinaNet, RetinaNet and FCOS, respectively.

3.3.2. EAPM with Balanced L1 Loss

To evaluate the effectiveness of EAPM for anchor-based detector regression, the experiments here first adopt the best configuration of EAPM for one-stage anchor-based detector classification. So the first baseline is the Libra RetinaNet model using EAPM ($\mu = 1.5$) with focal loss. Furthermore, EAPM for regression is not limited to one-stage detectors. Therefore, we also verify the effect on the two-stage anchor-based detector. The second baseline is the Libra R-CNN model using ResNet-50-FPN as the backbone. These two baselines use balanced L1 loss for regression. Table 2 shows the results of EAPM with balanced L1 loss and the baselines.

Table 2: Comparison of EAPM with balanced L1 loss and the corresponding baselines on MS COCO *val-2017*.

Settings	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Libra RetinaNet ($\mu = 1.5$)	38.2	58.0	40.7	22.4	41.5	**50.3**
$\theta = 1.0$	**38.4**	**58.2**	40.8	**22.5**	**41.9**	50.0
$\theta = 1.5$	38.2	57.9	40.6	22.3	41.8	50.0
$\theta = 2.0$	38.3	58.0	**40.9**	22.2	41.8	49.8
Libra R-CNN	38.3	**59.5**	41.9	22.1	42.0	48.5
$\theta = 1.0$	38.5	59.1	41.8	21.8	42.1	49.2
$\theta = 1.5$	**38.6**	59.2	41.9	22.4	**42.3**	49.1
$\theta = 2.0$	**38.6**	59.2	**42.1**	**22.5**	42.0	**49.6**

For the first baseline, we can see a gain of 0.2 AP by applying EPAM with balanced L1 loss. And it can also improve the performance of Libra R-CNN by 0.3 AP. It verifies our proposition that examples with large variations are more important for accurate localization.

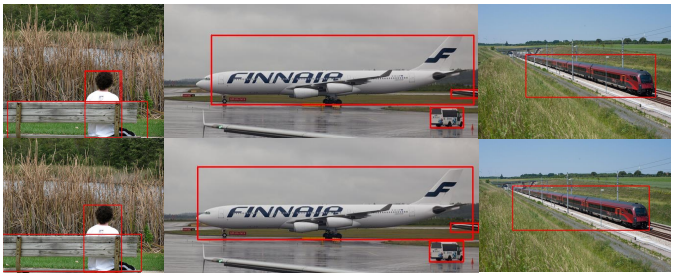

Fig. 5: Some qualitative results on *test-dev*. For the same image, the lower one corresponds to EAPM with balanced L1 loss, and the upper one corresponds to balanced L1 loss.

Furthermore, some qualitative results are shown in Fig. 5. We can see that EAPM with balanced L1 loss can better handle objects with large shape variations.

3.3.3. Main Results

We apply both EAPM with focal loss and EAPM with balanced L1 loss on Libra RetinaNet. The experiments are performed on MS COCO *test-dev*. Table 3 shows our main results compared with the state-of-the-art detectors. EAPM achieves 38.5 AP and 58.4 AP_{50} with ResNet-50-FPN, which surpasses Libra RetinaNet by 0.7 and 1.5 respectively. With ResNet-101-FPN, it achieves 40.4 AP and 60.5 AP_{50}.

Table 3: Comparison with the state-of-the-art detectors on MS COCO *test-dev*.

Backbone	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50-FPN	Faster R-CNN	37.7	58.7	40.8	21.7	40.6	46.7
	Mask R-CNN	38.3	**59.3**	**41.6**	**22.2**	41.2	47.5
	RetinaNet	36.9	56.2	39.3	20.5	39.9	46.3
	GHM	37.5	56.4	39.9	21.1	39.8	47.0
	Libra RetinaNet	37.8	56.9	40.5	21.2	40.9	47.7
	EAPM (ours)	**38.5**	58.4	41.1	21.7	**41.4**	**48.1**
ResNet-101-FPN	Faster R-CNN	39.7	60.7	43.2	22.5	42.9	49.9
	Mask R-CNN	**40.4**	**61.2**	**44.1**	23.1	43.5	50.8
	RetinaNet	39.0	58.6	41.7	21.9	42.2	49.3
	GHM	39.5	58.6	42.0	22.2	42.4	50.1
	Libra RetinaNet	40.0	59.7	42.8	22.8	43.1	50.7
	EAPM (ours)	**40.4**	60.5	43.2	**23.2**	**43.6**	**51.3**

4. CONCLUSION

In this paper, we focus on the gradient contribution imbalance in the training of deep object detectors and ascribe this imbalance to the example attribute imbalance. Furthermore, we propose EAPM, a strategy in cooperation with loss functions to simultaneously solve such imbalance in classification and box regression. Extensive experiments show that EAPM is very applicable for both one-stage detectors and two-stage detectors using various backbones. The detectors using EAPM can easily achieve the state-of-the-art performance.

5. REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[2] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin, "Libra r-cnn: Towards balanced learning for object detection," in *CVPR*, 2019, pp. 821–830.

[3] Zengsheng Kuang, Xian Fang, Ruixun Zhang, Xiuli Shao, and Hongpeng Wang, "A plug and play fast intersection over union loss for boundary box regression," in *ICASSP*. IEEE, 2021, pp. 1705–1709.

[4] Jiangmiao Pang, Kai Chen, Qi Li, Zhihai Xu, Huajun Feng, Jianping Shi, Wanli Ouyang, and Dahua Lin, "Towards balanced learning for instance recognition," *IJCV*, vol. 129, no. 5, pp. 1376–1393, 2021.

[5] Zhihao Wu, Jie Wen, Yong Xu, Jian Yang, and David Zhang, "Multiple instance detection networks with adaptive instance refinement," *TMM*, 2021.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[7] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[8] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," in *ECCV*, 2018, pp. 734–750.

[9] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in *ICCV*, 2019, pp. 9627–9636.

[10] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016, pp. 761–769.

[11] Buyu Li, Yu Liu, and Xiaogang Wang, "Gradient harmonized single-stage detector," in *AAAI*, 2019, vol. 33, pp. 8577–8584.

[12] Joya Chen, Qi Wu, Dong Liu, and Tong Xu, "Foreground-background imbalance problem in deep object detectors: A review," in *MIPR*. IEEE, 2020, pp. 285–290.

[13] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas, "Imbalance problems in object detection: A review," *TPAMI*, 2020.

[14] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001, vol. 1, pp. 511–518.

[15] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester, "Cascade object detection with deformable part models," in *CVPR*, 2010, pp. 2241–2248.

[16] Jie Wen, Yong Xu, Zuoyong Li, Zhongli Ma, and Yuanrong Xu, "Inter-class sparsity based discriminative least square regression," *NN*, vol. 102, pp. 36–47, 2018.

[17] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin, "Prime sample attention in object detection," in *CVPR*, 2020, pp. 11583–11591.

[18] Kean Chen, Weiyao Lin, John See, Ji Wang, Junni Zou, et al., "Ap-loss for accurate one-stage object detection," *TPAMI*, 2020.

[19] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *arXiv preprint arXiv:2006.04388*, 2020.

[20] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang, "Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection," in *CVPR*, 2021, pp. 11632–11641.

[21] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan, "Rank & sort loss for object detection and instance segmentation," in *ICCV*, 2021.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[23] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al., "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.