Zone Evaluation: Revealing Spatial Bias in Object Detection

Zhaohui Zheng, Yuming Chen, Qibin Hou, *Member, IEEE*, Xiang Li, *Member, IEEE*, Ping Wang, and Ming-Ming Cheng, *Senior Member, IEEE*

Abstract—A fundamental limitation of object detectors is that they suffer from "spatial bias", and in particular perform less satisfactorily when detecting objects near image borders. For a long time, there has been a lack of effective ways to measure and identify spatial bias, and little is known about where it comes from and what degree it is. To this end, we present a new zone evaluation protocol, extending from the traditional evaluation to a more generalized one, which measures the detection performance over zones, yielding a series of Zone Precisions (ZPs). For the first time, we provide numerical results, showing that the object detectors perform quite unevenly across the zones. Surprisingly, the detector's performance in the 96% border zone of the image does not reach the AP value (Average Precision, commonly regarded as the average detection performance in the entire image zone). To better understand spatial bias, a series of heuristic experiments are conducted. Our investigation excludes two intuitive conjectures about spatial bias that the object scale and the absolute positions of objects barely influence the spatial bias. We find that the key lies in the human-imperceptible divergence in data patterns between objects in different zones, thus eventually forming a visible performance gap between the zones. With these findings, we finally discuss a future direction for object detection, namely, spatial disequilibrium problem, aiming at pursuing a balanced detection ability over the entire image zone. By broadly evaluating 10 popular object detectors and 5 detection datasets, we shed light on the spatial bias of object detectors. We hope this work could raise a focus on detection robustness. The source codes, evaluation protocols, and tutorials are publicly available at https://github.com/Zzh-tju/ZoneEval.

Index Terms—Object detection, zone evaluation, spatial bias, spatial disequilibrium problem, spatial equilibrium learning.

1 Introduction

Object detection has shown impressive progress over the past two decades [11], [62], [74], [75]. While the optimization pipelines of object detectors are well-explored, their behaviors in a local image zone remain a mystery. The spatial robustness of a detector is especially fundamental [83] as objects may appear anywhere and all of them should be detected well. This is particularly important for safety vision applications, e.g., fire/smoke detection [37], [76], collision prevention in self-driving cars [5], [17], crowd counting and localization [38], [80], [85], [108], weapon detection in smart surveillance system [7], [70], and shoplifting detection [48], etc., where the border zone occupies a large proportion of the image area.

Unfortunately, the detectors are in fact unable to perform uniformly across the spatial zones, commonly showing an evident performance drop near the image border. This phenomenon, which we refer to as "spatial bias", is a natural obstacle in object detection, yet somehow being ignored for a long time by the detection community. Setting aside the issue may result in serious safety hazards and the risk of substantial property damage. For example, a fire detector may be good at detecting fire in the central zone while losing its ability to detect fires in the border areas of the image. Such a fire alarm system is unreliable since the central zone of the lens only occupies a small proportion of the image area. Some of the recent breakthroughs in spatial robustness of Convolution Neural Networks (CNNs) [4], [12], [45], [103], edging towards that ever-elusive translation invariance, have their basis for understanding how a small image transformation (e.g., color jitter, translation) does impact classification accuracy. It has been found that even for the same object, the classifier can make completely different predictions as its spatial position changes [45]. Beyond image classification, we in this work, delve into spatial bias in object detection, shedding light on the limitations of modern

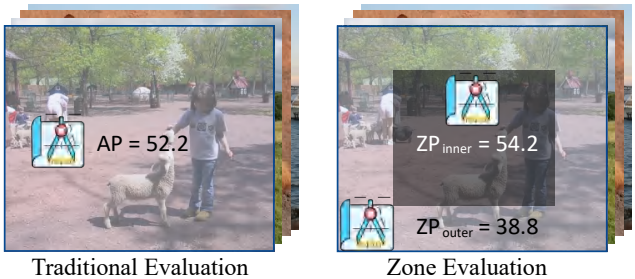


Fig. 1. The traditional evaluation measures the detection performance over the entire image zone, but it neglects the measurement of local zone and can hardly reflect the spatial bias. Our zone evaluation (ZP, the average precision constrained in the zone) compensates for these issues, indicating a large performance gap between zones. The results are reported by GFocal [54] on the VOC 2007 test set [25].

- *Z. Zheng, Y. Chen, Q. Hou, X. Li, and M.M. Cheng are with VCIP, School of Computer Science, Nankai University, Tianjin, China (Corresponding author: Qibin Hou).*
- *P. Wang is with the School of Mathematics, Tianjin University, Tianjin, China.*
- *This research was supported by NSFC (NO. 62225604, No. 62276145, No. U23B2049), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63223049), and CAST through Young Elite Scientist Sponsorship Program (No. YESS20210377). Computations were supported by the Supercomputing Center of Nankai University (NKSC).*

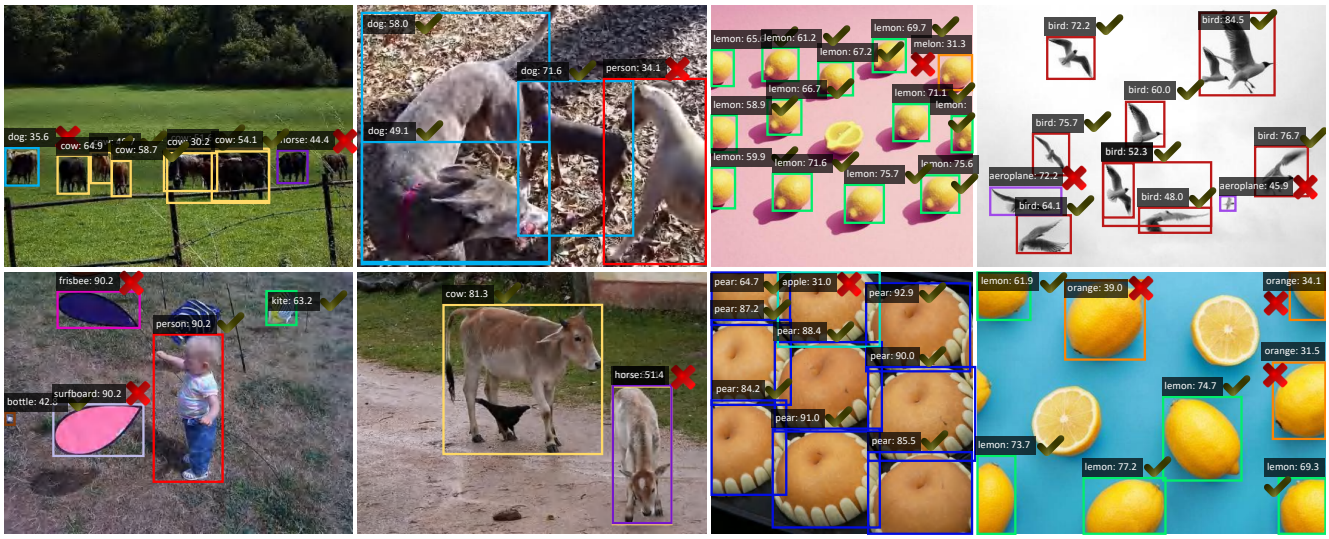


Fig. 2. The detector is less sanguine in detecting border objects. The visualizations are reported by GFocal [54]. Zoom in for a better view.

object detectors.

It has been an open issue for years that there is a lack of an effective way for measuring and identifying spatial bias. The traditional evaluation, i.e., AP metric, measures the detection performance over the entire image zone, which provides nothing guided for the spatial robustness of detectors, and one can hardly know where and how much the performance drops. Therefore, an evaluation protocol is particularly essential as it can provide the opportunity to better understand the spatial bias and provide tools to further build on methodology. To this end, we present a systematic way, termed zone evaluation, to analyze whether there is spatial bias in modern object detectors, and if so, how much.

Specifically, we extend the traditional full-map evaluation to a more generalized one. We calculate the common Average Precision (AP) [59] within a designated zone, yielding Zone Precision (ZP). During evaluation, only the boxes whose centers lie in the zone are considered. With these auxiliary metrics, we for the first time provide the numerical results that explicitly reveal the object detectors are in fact quite spatially biased across the zones. As shown in Fig. 1, the ZP gap is 15.4 between the inner zone and the outer one. Ostensibly, we can infer from this performance gap that the detection ability seems to be highly related to the absolute position of the object. However, this plausible conjecture has fundamentally many inconsistencies in practice when we shift the object in an image. We are unaware of any satisfactory explanation for the performance drop in the border zone (see Fig. 2). Thus, we would like to shed light on the existence and the major source of spatial bias in object detectors. Finally, we suggest a focus in future object detection research: toward spatial equilibrium.

The contributions of this work mainly involve three exploratory experiments on spatial bias, one potential research direction, and a comprehensive evaluation of modern object detectors.

Does object scale play a key role in centralized zone performance? (Sec. 4.1) Our answer is no. While large objects appear relatively frequently in the central zone, we observe that the zone performance remains quite uneven across different zones when we largely eliminate the effect of object scale. It is still difficult to forge a necessary connection between spatial bias and object scale.

Does the detector produce the centralized zone performance according to the absolute spatial position of the object? (Sec. 4.2) Our answer is no. The detection performance is barely relevant to the spatial position of objects. We observe that it can not statistically lead to an increase in detection quality when the objects move to the central zone.

What will actually determine the spatial bias in object detectors? (Sec. 4.3) Our analysis reveals strong evidence that spatial bias is highly related to the discrepancy of object data patterns between the zones. Specifically, there is a distinction in object data patterns between the central objects and the border objects. As a consequence, an object could be better detected if it is sampled from a central-zone-style data distribution rather than a border-zone-style one.

A new future direction: spatial disequilibrium problem. (Sec. 5) The work takes a step further and presents a practical issue that urgently needs to be addressed, namely, spatial disequilibrium problem. In such a problem setting, the object detector incorporates spatial equilibrium as one of the important goals, which has vital significance to robust detection. Facing this challenge, we also provide our first attempt, spatial equilibrium learning, toward spatial equilibrium object detection. (Sec. 5.2)

A comprehensive evaluation. (Sec. 6) We provide extensive evaluation and comparison of several representative object detectors. We reveal through experiments that 1) Spatial bias is quite common in various object detectors and datasets. 2) The spatial equilibria of detectors vary significantly. In particular, we will show that the sparse detectors perform better in the central zone, while the one-stage dense detectors perform better in the border zone. 3) The proposed spatial equilibrium learning is able to alleviate the spatial disequilibrium problem.

The remainders are organized as follows: Sec. 2 briefly reviews the study background. Sec. 3 presents zone evalu-

ation. Sec. 4 studies the main origin of spatial bias. Sec. 5 introduces the spatial disequilibrium problem and proposes spatial equilibrium learning to alleviate this issue. Sec. 6 gives extensive evaluations of zone evaluation and spatial equilibrium learning.

2 Background

2.1 Relationship with data imbalance problem

Let $\{X, G\} = \{x_i, g_i\}_{i=1}^n$ be a collection of sample-label pairs, where each sample x_i has a set of ground-truth labels g_i. The model training is conducted on the subset of $\{X, G\}$, and the network glances through the training set at each training epoch. The data imbalance problems are usually related to the inherent properties of $\{X, G\}$. In the literature of object detection, there are mainly two widely discussed imbalance problems.

The first one is class imbalance problem. In this case, the sample X is composed of multiple subsets $X_1, X_2, \cdots, X_c$ according to the class division, where the number of samples is imbalanced across c classes, thereby yielding a long-tail distribution [55], [72], [87], [107]. The class imbalance problem naturally causes imbalanced sampling during training, hindering the classification performance for those tail classes. Re-sampling strategies [42], [67] and cost-sensitive learning [21], [110] are the mainstream paradigms for class re-balancing.

The second is foreground-background sampling imbalance. This imbalance is also derived from the data itself. A large number of anchor points are tiled on the background area, which are naturally sampled to be the negatives and hence dominant the gradient flows. In this case, X can be split into X_{neg} and X_{pos}, s.t., $X = X_{neg} \bigcup X_{pos}$. The negative samples X_{neg} can be seen as the complementary set of the positives X_{pos}, whose ground-truth labels are "background" without bounding box annotations. The solutions to this problem are similar, including re-sampling, e.g., OHEM [79], Guided Anchoring [86] and IoU-balanced sampling [73], as well as cost-sensitive learning, e.g., Focal loss [58], GHM loss [51], and PISA [10].

In comparison, the spatial bias is also an obstacle in object detection. Given this, we establish a novel spatial disequilibrium problem for object detection. In this case, the sample X can be divided into multiple subsets according to the spatial zones, just like the class division. Generally speaking, spatial disequilibrium problem shares similar characteristics to class imbalance problem. The difference is that the latter has a long-tail distribution across classes, while the former considers the uneven distribution of objects over spatial zones. And we will show in Sec. 5.1, that the two problems are formally equivalent to each other.

2.2 Robustness in CNNs

It has been widely discussed that translation invariance is not fully held by deep CNNs [4], [40], [45], [93], [103] since they neglect the classical sampling theorem. A small image transformation could cause dramatic changes in predictions, thereby hindering the robustness of the classifier. Zhang R. [103] analyzed the flaws of the max-pooling operator and proposed to inject anti-aliasing for improving the robustness of deep networks. Lopes et al. [65] achieved a better robustness-accuracy trade-off by presenting a patch Gaussian augmentation.

Speaking of robustness over longer spatial ranges, recent studies [3], [45] reveal that CNNs can exploit the absolute positions of objects as additional information for image classification. Islam et al. [39] further extended that CNNs encode the position information based on the ordering of the channel dimensions. In [18], a spatially unbiased StyleGAN2 [44] is proposed to tackle the distorted face generation problem in the image border due to the photographer's bias in face dataset [43]. Gergely et al. [83] empirically identified there is a drop in classification accuracy when shifting the image to make the objects closer to the image border. Islam, M. A. et al. [40] revealed that there is a boundary effect on semantic segmentation, where the vehicle segmentation quality shows a high correlation to the density of cars in a region. Manfredi et al. [68] proposed a greedy approximation of AP variations by shifting the image by a few pixels to measure the translation equivariance of object detectors, but it inevitably requires several times the inference time to complete the evaluation. In this work, we numerically quantify the generalization ability of object detectors from the perspective of the local zone for the first time, which helps us better comprehend the existence and the discrete amplitude of spatial bias. This offers a novel analysis tool for the reliability of object detectors.

2.3 Evaluation from a local perspective

Evaluation from a local perspective has been widely demonstrated to have benefits in image quality assessment (IQA) [99] since the global evaluation is not aligned with the human visual system (HVS) [29], [60], [61], [78], [91], [105]. A global value cannot reflect the spatially non-stationary model capability. It can be dated to early research [66] in 1982 that the authors put forward a possible improvement in quality measure if local rather than global measurements were used. In IQA, a two-stage structure is commonly adopted. In the first stage, image quality is evaluated locally. The local measurement process typically produces a quality map. To convert such quality maps into an overall quality score, a pooling algorithm is applied in the second stage of the IQA.

Wang et al. proposed Mean-SSIM [91] to obtain a spatial-smooth measurement for image distortion assessment, the key of which is to compute the local SSIM for every sliding window, and then average. 3-SSIM [52] assigns different weights for the edges, textures, and smooth regions. Larson et al. [50] introduced a visibility-weighted local MSE to determine perceived distortion, where the image is divided into 16×16 blocks with 75% overlap between neighboring blocks. NIQE [69] index introduces patch selection to focus on the informative image patches. Chen et al. [14] proposed to use Landmark Distance (LMD) to focus on measuring the quality of the synthesized lip movement. Sun et al. [82] proposed weighted-to-spherically-uniform PSNR (WS-PSNR) to deliver different weights for different pixels. GMSD [94] exploits pixel-wise gradient magnitude similarity to capture image local quality. Fan et al. [26] proposed an S-measure for salient object detection, which first divides the image into 4 square grids and assigns different weights

to each local SSIM. Bosse et al. [8] attempted to use a CNN-based approach for learning the local image quality. Some methods [28], [31], [105], [106] incorporate saliency maps into IQA metrics since the conspicuous location can help predict the image quality perceived by human observers. The local metrics are helpful for describing small details and structural similarity between image patches.

Although the local evaluation has been popular for decades in many assessment systems of computer vision applications, it has not been fully investigated in object detection. Most IQA methods are mainly for pixel prediction tasks, e.g., image restoration [32], [56], [98], salient/camouflaged object detection [27], [35], [109], and image super-resolution [22], [47]. They cannot be directly applied to instance prediction tasks, e.g., generic object detection. Therefore, we propose zone evaluation to fill this research blank. In addition, our work conducts a series of heuristic experiments, which provide new insights for understanding the spatial bias of modern object detectors. We also study several shapes of zone partitions in evaluating object detectors (annular, strip, square), while previous IQA methods pay little attention to this.

3 Zone Evaluation

In this section, we extend the traditional object detection evaluation to a more generalized zone evaluation. Given a test image I and a set of evaluation metrics $\mathcal{M}$, the classic evaluation methods simultaneously calculate the metrics for all the detections and the ground-truths over the whole image. The elements in $\mathcal{M}$ can be the COCO-style AP (Average Precision) [59], mAP across 10 IoU thresholds, or AP for the small/medium/large objects, which have been widely used in object detection. These traditional metrics measure the detection performance over the entire image zone but consider nothing about the spatial robustness of object detectors.

Zone metric. Let $S = \{z^1, z^2, \cdots, z^n\}$ be a zone partition s.t. $I = \bigcup_S z^i$ and $z^i \cap z^j = \emptyset$, $\forall z^i, z^j \in S$, $z^i \neq z^j$. We measure the detection performance for a specific zone z^i by only considering the ground-truth objects and the detections whose centers lie in the zone z^i. Then, for an arbitrary evaluation metric $m \in \mathcal{M}$, the evaluation process stays the same as the conventional ways, yielding n zone metrics, each of which is denoted by m^i. We call $m^S = \{m^1, m^2, \cdots, m^n\}$ the series of zone metrics for the zone partition S.

Annular zones. In practice, the centralized photographer's bias is ubiquitous in visual datasets [25], [49], [59], [71], [77], [84]. If one aims at a detector with comprehensive detection ability for all directions, the evaluation zones can be designed into a series of annular zones:

$$z^{i,j} = R_i \setminus R_j, \quad i < j, \tag{1}$$

where R_i denotes a centralized region, which is given by:

$$R_i = \text{Rectangle}((r_i W, r_i H), ((1 - r_i)W, (1 - r_i)H)), \tag{2}$$

where $\text{Rectangle}(p, q)$ represents the rectangle region with the top-left coordinate p and the bottom-right coordinate q. W and H denote the width and the height of the image and $r_i = \frac{i}{2n}, i \in \{0, 1, \cdots, n\}$ controls the sizes of rectangles.

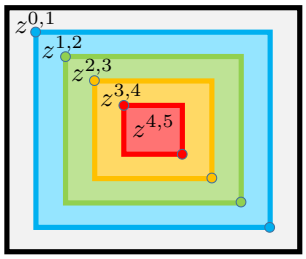


Fig. 3. Definition of evaluation zone when $n = 5$.

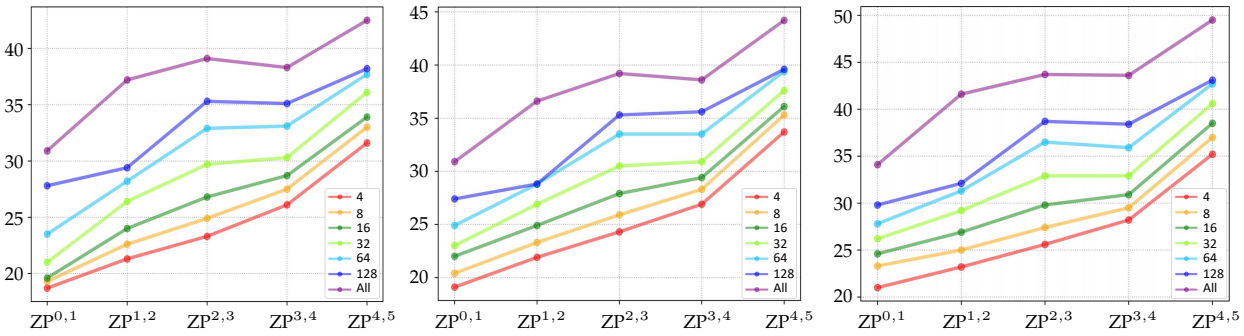
An illustration of the evaluation zones can be seen in Fig. 3, where $n = 5$. We denote the average precision (AP) in the zone $z^{i,j}$ as $\text{ZP}^{i,j}$. In this way, the traditional evaluation is a special case in our zone evaluation as one can easily get $\text{AP} = \text{ZP}^{0,n}$.

Other zone partition. The traditional evaluation is flexible to select different IoU thresholds for different application scenarios. For those requiring precise box localization, a rigorous IoU threshold can be selected, e.g., $IoU = 0.75$ (AP_{75}). For those less requiring box localization, AP_{50} is good enough. For example, rotated object detection methods usually report AP_{50} [95], [96], [97]. Analogous to the AP metric, users can flexibly design various zone partitions based on their own applications. If the user cares about the comprehensive detection ability for all directions, the annular zone partition would be a good choice. If the user cares about some regions of interest, the evaluated zones can be customized. In Sec. 6, we will show the evaluation results on two more special zone partitions. One is the strip zones, i.e., 5 zones along x-axis and 5 zones along y-axis (See Sec. 6.3 "Observations 3"), and the other is square zones of 11×11 blocks (See Sec. 6.4 "Correlation with object distribution". Importantly, as the zone partition holds consistently, the comparison among the detectors stays fair. This property helps us observe the performance of different detectors in the regions of interest so that we can choose detectors based on practical application needs. In Sec. 6, we will show that the sparse detectors perform better in the central zone, while the one-stage dense detectors perform better in the border zone.

Measuring the discrete amplitude for zone metrics. As the detection performance varies across the zones, we further introduce an additional metric to gauge the discrete amplitude among the zone metrics. Given all the zone metrics m^S for a specific zone partition S, we calculate the variance $\sigma(m^S)$ of the zone metrics. Ideally, if $\sigma(m^S) = 0$, the generalization ability of the object detector reaches perfect spatial equilibrium **under the current zone partition**. In this situation, an object can be well detected without being influenced by its data pattern. It is also worth mentioning that **spatial bias is an external manifestation of object detectors**, and the ZP variance can only reflect the spatial equilibrium for the given zone partition. In other words, there are three concepts as follows.

1) Let S be a zone partition, the spatial equilibrium of the detector is good for S if $\sigma(m^S)$ is sufficiently small.

2) Let S_1, S_2 be two different zone partitions with n zones. If $\sigma(m^{S_1}) < \sigma(m^{S_2})$, then the detector is considered to be more spatially equilibrated in the way of zone partition



(a) GFocal (b) Cascade R-CNN (c) Deformable DETR

Fig. 4. Mean ZP with various object scale ranges. One can see that for each range of the object scale r, the spatial bias is significant on the three object detectors.

S_1.

3) The detector has no spatial bias, indicating $\forall S$ be a zone partition, $\sigma(m^S)$ is sufficiently small.

4 Delving Deep into Spatial Bias

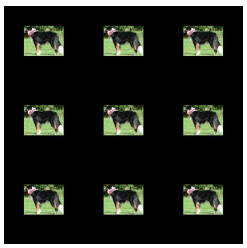
In this section, we conduct exploratory experiments to clarify the spatial bias about its existence and the major source. We adopt three representative object detectors. The first one is the popular one-stage dense object detector GFocal [54]. The second is the classic multi-stage dense-to-sparse object detector Cascade R-CNN [9]. The third is the sparse object detector Deformable DETR [111].

4.1 Study on object scale

From the definition of evaluation zones (Eq. (1)), one may ask whether the object scale plays a key role in centralized zone performance. In this experiment, the annular zone partition is adopted as shown in Fig. 3. An object belongs to $z^{i,j}$ if its central point coordinates are in $z^{i,j}$. To eliminate the effect of object scale, we restrict the zone evaluation process in the objects with a similar scale. For each range of object scale r, we select all the ground-truth boxes whose areas are in the range of $[0, r^2]$, $[r^2, (2r)^2]$, ..., $[(kr)^2, \infty]$, respectively, where the maximum endpoint of the scale is set to $kr = 256$, and $r \in \{4, 8, 16, 32, 64, 128, \infty\}$. Then, we calculate the mean value of ZP over all the scales, which is shown in Fig. 4. We observe that the spatial bias is significant no matter how small the range of the object scale is chosen. The $\text{ZP}^{4,5}$ score continues to be the best and in contrast, the $\text{ZP}^{0,1}$ score is the worst. It shows a steeper drop-off in performance as the evaluation zones get closer to the image border. One can see that the performance gap between the inner zone and the outer zone is consistently large, with more than 10 ZP gap. This indicates that the centralized spatial bias is probably not derived from the object scale factor. For simplicity, we adopt all scales for zone evaluation in the following experiments.

4.2 Study on the absolute spatial position of objects

As the zone performance exhibits a clear centralization trend, one straightforward conjecture is that the spatial bias is related to the absolute spatial positions of objects. The object may be better detected if it is shifted to the central zone, and reversely, be worse if it is in the border zone. To analyze whether the spatial positions of objects play a key role in their detection quality, we construct the Sudoku-style dataset for object detection of the regularly placed objects on a black 600×600 image. The experiment is based on the following 3 steps: (1) We first crop the objects from PASCAL VOC 2007 test set [25], 14,976 objects in total. (2) All the objects are scaled to a fixed size and placed in a 3×3 grid manner. See Fig. 5(a). (3) To measure the detection quality of each grid, the evaluation zones are defined as the same 3×3 zones, denoted by z^{ij}, $i, j \in 1, 2, 3$. It can be seen from Fig. 5(b) that the detector does not perform the best and even the worst in the central zone z^{22}. This phenomenon is somewhat incompatible with previous observation in [83], which analyzes the effect of shifting 100 images and concludes that the detector may have a performance drop when the objects close to the image border. However, when we increase the number of samples up to more than 14K, we observe that shifting objects to the central zone cannot statistically eventuate an increase in detection quality. Therefore, it is less discernible for the relevance between the centralization tendency of detection performance and the absolute position of objects.



(a) Sudoku-style dataset.

32.7 z^{11}	31.1 z^{12}	31.1 z^{13}
29.6 z^{21}	27.8 z^{22}	28.5 z^{23}
30.6 z^{31}	28.9 z^{32}	29.3 z^{33}

(b) ZPs of 9 zones.

Fig. 5. (a) The Sudoku-style dataset is constructed by regularly placing all the objects of the test set on a 600×600 black image. (b) Zone evaluation on GFocal (3×3 grids).

Discussion: The conclusion of Fig. 4 is that the central objects can be better detected than the border objects and it is irrelevant to object scale, while the conclusion of Fig. 5 is that the absolute position of the object (by object shifting) is irrelevant to forming the centralized zone performance. We ask: What will actually determine the spatial bias in object detectors?

4.3 Object data patterns between the zones

This subsection aims to investigate the origin of the centralized spatial bias in object detectors. Our inspiration behind it is that the centralized spatial bias may come from the differences of the object data patterns between the zones. An object can be better detected if sampled from a central-zone-style data distribution, and worse if it is sampled from a border-zone-style one. The representation of the object data patterns is very general [6], [16], [102], and only a few adaptive modifications need to be made to the detection pipeline, requiring only 1) the input image I, 2) all the ground-truth boxes G, 3) a feature extractor $f : I \rightarrow \mathbb{R}_{M\times H\times W}$ from a pre-trained object detector. During inference, we crop the object features from $f(I)$ by using the ground-truth boxes, and then average the feature values along the spatial dimensions. Each object is represented by a M-dimensional feature vector g, which encodes the object data patterns in high-dimensional space. The number of zones is set to 2 in this experiment. We denote the central zone as z^{in}, which is a rectangle region with top-left coordinate $p = (0.25W, 0.25H)$ and bottom-right coordinate $q = (0.75W, 0.75H)$, where W and H are the width and the height of the input image. The rest part of the region is set as the border zone, denoted by z^{out}. The differences of the object data patterns between the zones are represented as:

$$\mathcal{E}((G_1,u),(G_2,v)) = \frac{1}{KCM}\sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{m=1}^{M}\mathbb{1}_k||\bar{g}_{m,c}^{u,G_1} - \bar{g}_{m,c}^{v,G_2}||, \tag{3}$$

where $\bar{g}$ denotes the feature representation center, $u, v \in \{z^{in}, z^{out}\}$ represent the objects sampled from the central zone or the border zone, as well as $G_1, G_2 \in \{G_{train}, G_{test}\}$ represent the objects sampled from the train set or the test set. The errors are calculated for each class and then averaged. C is the total number of classes. An indicator function $\mathbb{1}_k$ is also introduced to eliminate the effect of object scale, which is 1 when the object scale is in one of the range $R = \{[((k-1)r)^2, (kr)^2]\}\bigcup\{[((K-1)r)^2, \infty]\}$, $k \in \{0, 1, \cdots, K-1\}$, and 0 otherwise. In a nutshell, $\mathcal{E}$ measures the distance between the feature representation centers of the four sets, i.e., the central/border objects from the train/test set.

The results are reported in Fig. 6. For VOC, the train set is VOC 2007 trainval, and the test set is VOC 2007 test. For COCO, the train set is COCO train2017, and the test set is COCO val2017. One can see that the objects sampled from the same zone have significantly lower differences than those sampled from different zones. Specifically, the central objects of the test set are more similar to the central objects of the train set, and the border objects of the test set are more similar to the border objects of the train set. This indicates that the object data patterns are actually distinct across the zones, and the network is able to capture such deviation. As illustrated in Fig. 7(a), the zone performance is centralized on VOC 2007 trainval set, thus it naturally inherits the same trend on the test set. More intriguingly, we further visualize the detection performance on the Sudoku-style dataset in Fig. 7(b-f), where the central objects and the border objects are separated. It can be seen that the

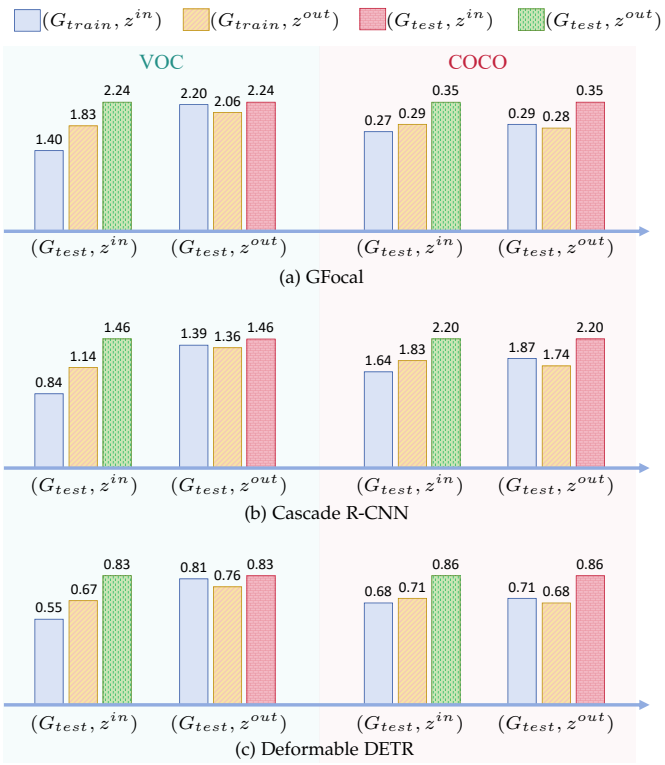


Fig. 6. Average error $\mathcal{E}((G_1,u),(G_2,v))$ between the feature representation centers. For instance, the blue bar denotes $\mathcal{E}((G_{test}, z^{in}), (G_{train}, z^{in}))$. The objects sampled from the same zone have significantly lower differences than those sampled from different zones.

detector can consistently perform better in detecting the central objects no matter where they are. We notice that this phenomenon holds for all the 5 datasets, including PASCAL VOC, MS COCO, and 3 other application datasets (face mask, fruit, helmet).

The above results confirm the intuition that the objects would be better detected if they are sampled from the central-zone-style data distribution, and be worse if they are sampled from the border-zone-style one. This shows that when we humans take photos, there is always a divergence in the object data patterns between zones, though it is imperceptible. When the lens are focused on the regions of interest where objects are most likely to appear, it inevitably leads to a decrease in the sampling frequency of objects in the border zone, resulting in sub-optimal performance.

5 Spatial Disequilibrium Problem

Thus far, we have shown the existence and the major source of spatial bias. The sub-optimal performance in the border zone impedes the robustness of detection applications. In this section, we introduce the new spatial disequilibrium problem for object detection.

5.1 Problem definition

Denoting S be a zone partition, m^S be a series of zone metrics, and $\sigma : \mathbb{R}^n \rightarrow \mathbb{R}$ be a variance calculation, the spatial disequilibrium problem is defined to minimize the variance of the zone metrics:

$$\min_{\Theta} \sigma(m^S|\Theta), \tag{4}$$

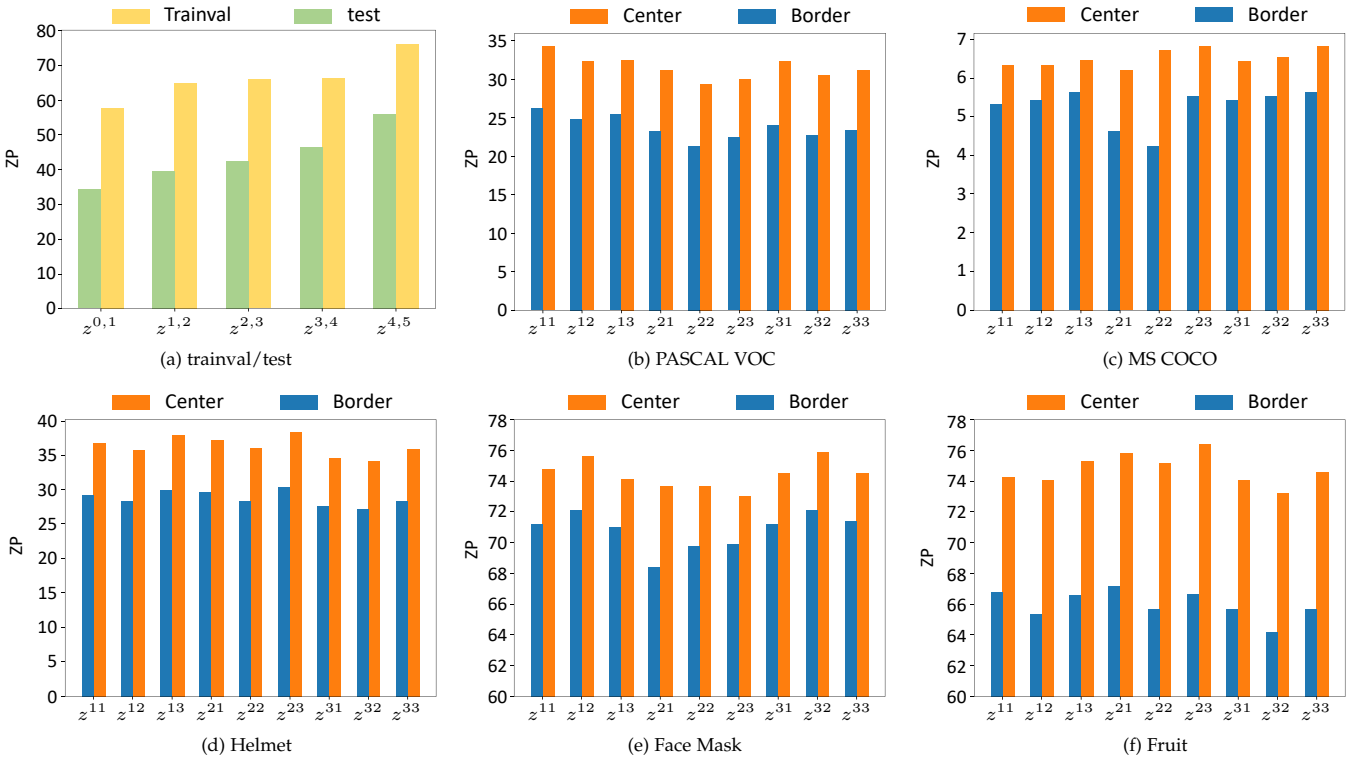


Fig. 7. (a) The 5-zone evaluation on VOC 2007 trainval set and test set. (b-f) The zone evaluation on Sudoku-style dataset with the center objects and the border objects separately. It shows that the central objects can always be better detected no matter where we place them.

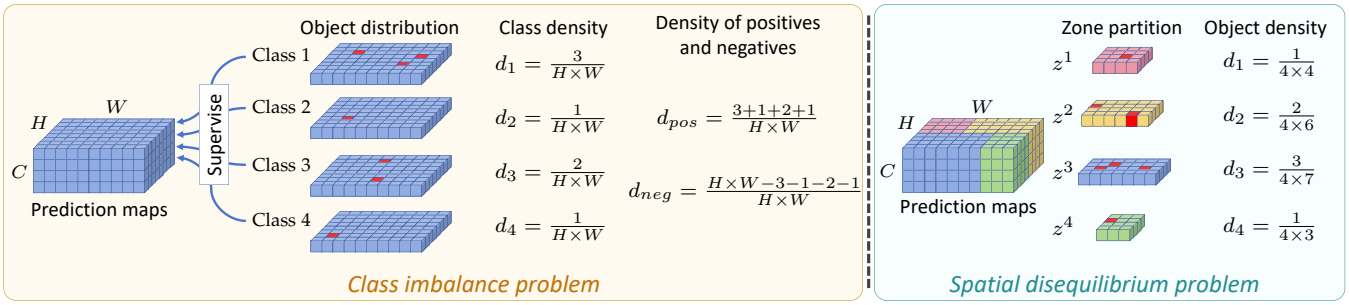


Fig. 8. Illustration of the relationship between class imbalance problem and spatial disequilibrium problem. There are 7 objects of 4 classes, denoted by red cubes. The evaluated zone is set to 4 zones, colored pink, yellow, blue, and green. We simplify the discussion that each object contains only 1 positive sample, and similarly for the case of multiple positives. In the left diagram, the class density is the ratio of the number of objects and the size of prediction maps for each class, whereas, in the right diagram, the object density is the ratio of the number of objects and the size of zones for each zone. The spatial disequilibrium problem is formally equivalent to the class imbalance problem.

where Θ is the set of network parameters of the detector. The general goal of facilitating spatial equilibrium is primarily decided by which zone partition to use, which is application-oriented. Thus, for different application scenarios, the zone partition can be customized.

Disucssion: *The spatial disequilibrium problem is formally equivalent to the class imbalance problem.* We denote the objects in the zone z^i as X^i_{obj}. The density of objects for a given zone z^i can be represented as $d_i = |X^i_{obj}|/|z^i|$. Intuitively, higher density indicates more positive samples, thereby generating a larger gradient flow on the zone. It is analogous to the class imbalance problem which has a long-tail distribution among the classes, as shown in Fig. 8. Specifically, the classification branch predicts the class scores which is a $C \times H \times W$ tensor. The density of the c-th class is represented as the ratio $d_c = |X_c|/(H \times W)$, X_c is the object of the c-th class. Since $H \times W$ is a constant, it is equivalent that all the samples of a class are placed in a zone with an identical area. Then, each class has a single $H \times W$ zone for model learning and is disjoint among classes. In Fig. 9, it can be seen that both problems are subject to a long-tail distribution. Due to that fact, the zone performance may also be correlated to the supervision signal strength in the zone. A simple approach is to enlarge or reduce the supervision signal strength for the zones so that the network reaches a new convergence state. Here we plug a new parameter β into label assignment algorithm [104]. An anchor is assigned to be a positive sample if its IoU is larger than $\alpha_{pos} + \beta * \mathbb{1}_z$, where α_{pos} is the positive IoU threshold, $\mathbb{1}_z$ is 1 if the central point of this

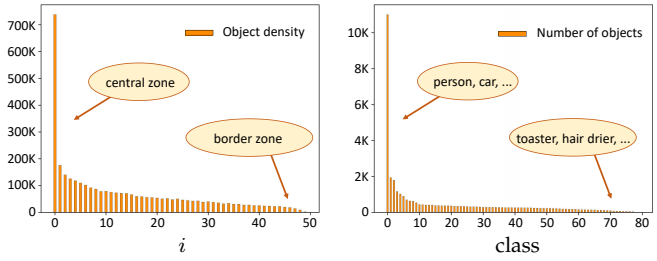


Fig. 9. **Left**: The object density against 50 zones on COCO val2017. The zones are centrally defined as $z^{i,i+1}, i = 0, 1, \cdots, 49$. **Right**: The long-tail distribution of classes on COCO val2017. The spatial disequilibrium problem shares a similar characteristic to the class imbalance problem.

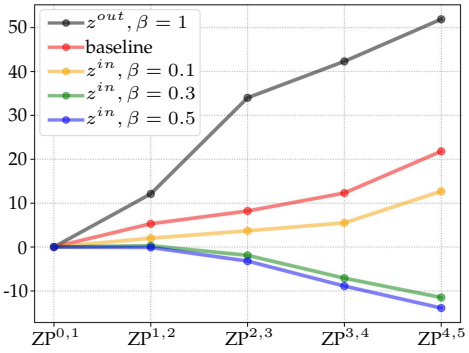


Fig. 10. ZP relative to $ZP^{0,1}$. The zone performance can be further extremely centralized if the supervision signal strength is reduced in the border zone, and reversely, be anti-centralized if the supervision signal strength is reduced in the central zone.

anchor lies in the zone z and 0 otherwise. Thus, the number of the positive samples $|X^i_{pos}|$ decreases when β increases.

We visualize the relative ZP in Fig. 10, where all the ZPs are subtracted by $ZP^{0,1}$. One can see that the centralized spatial bias can be further aggravated if we weaken the supervision signals by reducing the number of positive samples in the border zone. Reversely, we can even achieve an anti-centralized spatial bias if we weaken the supervision signals in the central zone. This indicates that the supervision signal strength does have an impact on the zone performance. Given the above analysis, we finally discuss one possible solution for addressing the spatial disequilibrium problem under the annular zone partition.

5.2 Spatial equilibrium learning

Most existing research on object detection focuses on pursuing higher detection performance at the image level while neglecting optimization at the zone level, resulting in serious spatial disequilibrium issues for detectors. In this subsection, we introduce a possible solution, termed Spatial Equilibrium Learning as the beginning of relieving the spatial disequilibrium problem. We first introduce the spatial weight, which maps the anchor point coordinates (x^a, y^a) to a scalar $\alpha(x^a, y^a)$ by

$$\alpha(x, y) = 2 \max \left\{ ||x - \frac{W}{2}||_1 \frac{1}{W}, ||y - \frac{H}{2}||_1 \frac{1}{H} \right\} \in [0, 1], \tag{5}$$

where W and H are the width and the height of the image. The spatial weight can be easily plugged into the existing detection pipelines with few modifications. The principle is simple and multi-optional. Here, we offer two implementations as follows.

1) Spatial equilibrium label assignment (SELA). In this approach, the key idea is to consider the spatial weight as an additional constraint term when making the criterion rule for label assignment. Since most of the label assignment algorithms have their own sophisticated implementations, in the following, we provide a specific application description of the classic ATSS [104], simply because of its brevity. Given the positive IoU threshold t, which is calculated by considering the statistical characteristics of the objects. The ATSS criterion follows the same rule as the max-IoU assignment [58], [74], [75], i.e., $\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t$, where $\boldsymbol{B}^a$ and $\boldsymbol{B}^{gt}$ denote the preset anchor boxes and the ground-truth boxes. The SELA process is represented as:

$$\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t - \gamma\alpha(x^a, y^a), \tag{6}$$

where $\gamma \geqslant 0$ is a hyperparameter. It can be seen that SELA relaxes the positive sample selection conditions for objects near the image borders. Therefore, more anchor points will be selected as positive samples for them. Notice that the above application is actually a frequency-based approach, just like many of the class rebalance sampling strategies proposed for the long-tail class imbalance problem [42], [67].

2) Spatial equilibrium loss (SE loss). In this approach, we adopt the cost-sensitive learning method. We take the spatial weight term $1 + \gamma\alpha(x^a, y^a)$ as an additional weight factor for the classification and the bounding box regression losses. As such, a larger gradient flow will be generated in the border zone so that the network pays more attention to the border objects.

Future directions: There are more potentially promising solutions toward spatial equilibrium that warrant future study. For example, designing an appropriate data augmentation [19], [46], [90], more specifically, increasing data augmentation to make up for the insufficient sampling frequency for the objects near the image borders, might be a promising solution. Besides, since the spatial disequilibrium problem is formally equivalent to the class imbalance problem, some improved re-balancing methods may also bring gains to spatial equilibrium learning, e.g., class-balanced loss [21], transfer learning [20], [89], and representation learning [23], [36], [64], etc. There is also a very exciting area for future work in pursuing more solutions to address the two problems in a unified form. Furthermore, our approaches mainly consider the annular zone partition, which is designed to balance the detection ability between the central zone and the border zone. For some special applications that care more about some regions of interest, the design ethos can be versatile and contingent on the practical application.

6 Quantitative Evaluation

In this section, we conduct comprehensive evaluations on 10 popular object detectors and 5 object detection datasets.

6.1 Experimental setups

Detectors and metrics. All the object detectors we evaluate can be downloaded from MMDetection [13] or their

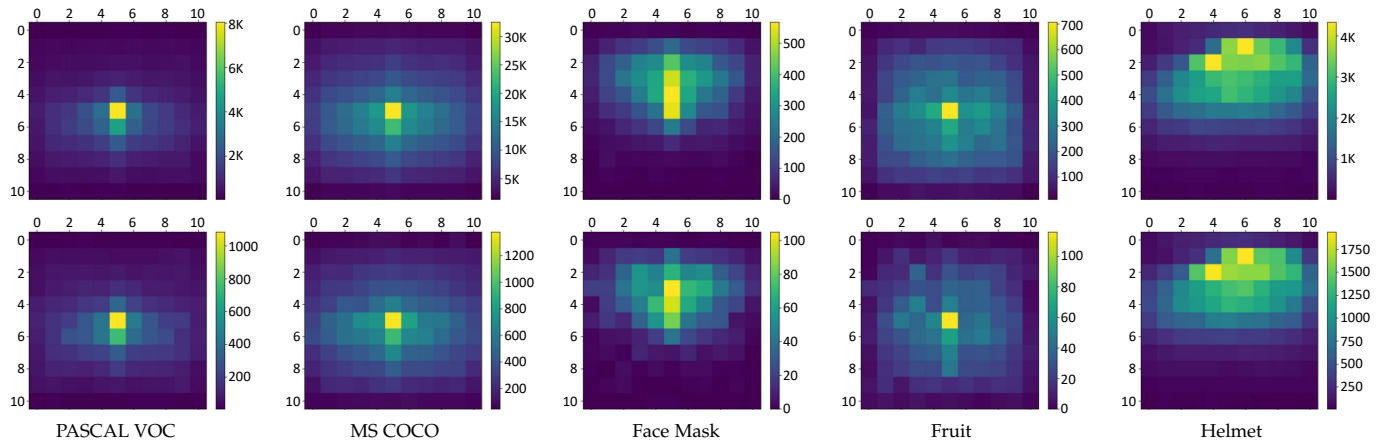


Fig. 11. The photographer's bias in the 5 object detection datasets. We count the center points for all the ground-truth boxes. The images are divided into 11×11 zones. First row: Train set. Second row: Test set.

official websites. We follow the standard Average Precision evaluation protocol. To comprehensively evaluate the detectors, various metrics are reported, including 5 ZPs, the variance of 5 ZPs, and the traditional metric AP.

Datasets. All the datasets we used are publicly available and can be downloaded from their official websites or Kaggle. The object distributions of the 5 datasets can be seen in Fig. 11.

PASCAL VOC [25] is one of the most widely used object detection benchmark under natural scenes, which contains 20 classes. We adopt the classical 07+12 training and testing protocol, i.e., the train set contains the union of VOC 2007 trainval and VOC 2012 trainval (totally 16551 images) and the test set contains VOC 2007 test (4952 images).

MS COCO [59] is another recently popular benchmark with much larger scale, containing 80 classes under natural scenes. We adopt COCO 2017 train (118K images) for training and COCO 2017 val (5K images) for evaluation.

Face mask detection [1]. With COVID-19 raging around the world, face mask detection is a widespread and necessary visual application. The dataset consists of 5,865 images for training and 1,035 images for testing. There are 2 classes. One is face and the other is mask.

Fruit detection [24] is widely used in industrial assembly line sorting and commodity classification. The dataset consists of 3,836 train images and 639 test images. 11 common fruits are included, e.g., apple, grape, and lemon, etc.

Helmet detection [2] is a safety vision application that is often used on construction sites to detect whether workers and visitors are wearing helmets. It contains 15,887 images for training and 6,902 images for testing. Two classes, head and helmet, are used.

Setups of spatial equilibrium learning. For spatial equilibrium learning evaluation, the implementation is based on the MMDetection [13] framework, and the ablation study is conducted on GFocal [54] with ResNet [34] backbone and FPN [57] neck. We use ResNet-18 for VOC 07+12 and 3 application datasets, and adopt ResNet-50 for MS COCO. The learning rate is linearly scaled by the linear scaling rule [30] according to the number of GPUs. The training epochs are set to 12 for all the experiments. We set $\gamma = 0.2$ in Eq. (6), and all the other hyper-parameters remain unchanged for a fair comparison.

TABLE 1
Zone evaluation on the existing popular object detectors. 5 zone precisions (ZP), the variance of ZP, and the traditional metric AP are reported. The results are reported on COCO 2017 val. "Cas.": Cascade. **R**: ResNet [34]. **X**: ResNeXt-32x4d [92]. **PVT-s**: Pyramid vision transformer-small [88]. **CNeXt-T**: ConvNext-T [63].

Detector	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
DETR (**R**-50) [11]	40.1	26.9	29.8	36.2	39.8	39.1	45.7
RetinaNet (**PVT-s**) [88]	40.4	19.7	30.8	36.9	39.0	37.4	44.6
Cascade R-CNN (**R**-50) [9]	40.3	18.7	30.9	36.6	39.2	38.6	44.2
GFocal (**R**-50) [54]	40.1	16.9	31.1	37.5	39.4	38.5	43.8
Cas. Mask R-CNN (**R**-101) [9]	45.4	22.4	34.7	41.6	44.3	44.4	49.1
Sparse R-CNN (**R**-50) [81]	45.0	21.6	35.8	41.9	43.4	44.0	50.3
YOLOv5-m [41]	45.2	12.9	36.0	42.3	44.5	43.2	46.7
Deform. DETR (**R**-50) [111]	46.1	23.2	36.3	42.6	45.6	45.1	51.2
Sparse R-CNN (**R**-101) [81]	46.2	21.2	36.9	42.9	44.9	44.7	51.3
Cas. Mask R-CNN (**X**-101) [9]	46.1	21.1	36.1	42.0	44.8	45.9	49.9
Mask R-CNN (**CNeXt-T**) [63]	46.2	17.6	36.7	41.9	44.5	43.6	49.7
GFocal (**X**-101) [54]	46.1	15.7	37.0	43.5	45.0	44.4	49.3
VFNet (**R**-101) [101]	46.2	15.6	36.7	43.0	45.0	44.5	48.8

6.2 Zone evaluation on various object detectors

Although traditional evaluation provides good guidance for the overall performance of detectors, little is known about the spatial bias of detectors, as well as the location and degree. Here, we select various object detectors that have different detection pipelines but with the same level of traditional metrics. They are popular, representative, and considered to be the milestones in modern object detection: one-stage dense detectors (RetinaNet [58], GFocal [54], VFNet [101], YOLOv5 [41]), multi-stage dense-to-sparse detectors (R-CNN series [9], [33], [75]), and sparse detectors (DETR series [11], [111] and Sparse R-CNN [81]). The quantitative results are reported in Table 1.

There are several interesting observations:

1) *Spatial bias is quite common*. One can see that all the detectors show a clear centralized zone performance, i.e., performing well in the central zones ($z^{3,4}, z^{4,5}$) but unsatisfactorily in the border zones ($z^{0,1}, z^{1,2}$). This confirms the existence and universality of spatial bias, and we successfully quantify the defects of object detectors shown in Fig. 2 for the first time.

2) *Their spatial equilibria vary significantly.* There is a large gap of 12.9 to 26.9 in ZP variance. Particularly, we find that

TABLE 2
Zone evaluation on VOC 07+12, COCO 2017, and 3 application datasets, i.e., face mask detection, fruit detection, and helmet detection. 5 zone precisions (ZP), the variance of ZP, and the traditional metric AP are reported. The results are reported on GFocal [54].

Dataset	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
VOC 07+12	52.2	53.6	34.3	39.6	42.5	46.6	56.1
COCO 2017	40.1	16.9	31.1	37.5	39.4	38.5	43.8
Face Mask	71.3	13.1	60.4	67.1	69.0	68.8	70.9
Fruit	76.6	56.2	60.8	69.9	71.2	75.3	83.8
Helmet	49.7	3.0	45.9	47.9	50.3	50.6	47.8

	VOC	COCO	Face mask	Fruit	Helmet
VOC	38.2	46.7	55.5	45.1	42.9
COCO	35.0	39.1	42.5	39.3	35.9
Face mask	63.4	68.3	72.9	71.0	66.4
Fruit	72.2	72.8	80.4	76.2	69.7
Helmet	54.0	51.5	47.8	48.7	47.3

(a) 5 zones along x-axis

	VOC	COCO	Face mask	Fruit	Helmet
VOC	36.1	30.6	69.1	64.5	44.7
COCO	36.9	34.8	69.7	70.7	52.8
Face mask	52.5	39.4	69.8	77.8	49.8
Fruit	47.4	41.4	63.8	74.3	44.4
Helmet	36.2	36.3	53.4	68.8	36.6

(b) 5 zones along y-axis

Fig. 12. Two designs of zone partition. ZP is reported.

TABLE 3
ZP variance on three types of zone partitions. Spatial bias is an external manifestation of detectors, whereas spatial equilibrium corresponds to a given zone partition.

Zone partition	VOC	COCO	Face Mask	Fruit	Helmet
5 Annular zones	53.6	16.9	13.1	56.2	3.0
5 zones along x-axis	32.3	7.2	11.2	13.7	6.4
5 zones along y-axis	46.7	14.0	39.6	20.8	30.5

TABLE 4
Evaluation for hyper-parameter γ in SELA. 5 zone precisions (ZP), the variance of ZPs, and the traditional metric AP are reported. $\gamma = 0$ denotes the baseline GFocal. Lower variance, better spatial equilibrium. (Dataset: VOC 07+12)

γ	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
0	52.2	53.6	34.3	39.6	42.5	46.6	**56.1**
0.1	52.5	44.5	35.9	40.6	42.1	46.6	55.6
0.2	**52.8**	**37.7**	**37.6**	40.3	**43.8**	**46.9**	55.4
0.3	**52.8**	37.3	37.4	**41.5**	43.6	**46.9**	55.6
0.4	52.0	46.3	35.0	38.9	42.6	46.6	54.8

the sparse detectors, e.g., DETR series and Sparse R-CNN, tend to produce a large ZP variance, while the one-stage dense object detectors perform better in spatial equilibrium (lower ZP variance). It shows that the DETR-based detectors do not align with CNN-based detectors in terms of spatial equilibrium. We conjecture that it may be attributed to the global information captured by the self-attention mechanism. The sparse detectors first extract features by CNNs and then process the features by a series of attention modules. The training is more dynamic, and we hypothesize that the central objects may receive more attention. Clearly, one can conclude that there must be some factors that lead to different spatial equilibria among detectors, including but not limited to the neural network architecture designs, optimization, and training strategies. Yet currently, we are unaware of which components or algorithm designs have an impact on spatial bias. We believe that further research on this topic in the future will be interesting, and the study is of great possibility to find the key to addressing the spatial disequilibrium problem.

3) *Traditional evaluation fails to capture spatial bias.* It can be seen that GFocal (R-50) and DETR (R-50) achieve the same AP score of 40.1. However, the traditional metrics provide nothing about how much the detection performance is in a zone. Our zone evaluation shows that GFocal performs better in the border zones $z^{0,1}$ and $z^{1,2}$, while DETR performs better in the zones $z^{2,3}$, $z^{3,4}$, and $z^{4,5}$. Similarly, Deformable DETR (**R**-50) [111] achieves the same traditional AP as GFocal (**X**-101). The zone evaluation shows that Deformable DETR performs significantly better than GFocal in the central zones $z^{3,4}$, $z^{4,5}$, whereas worse in the border zones $z^{0,1}$, $z^{1,2}$. And these performance discrepancies are concealed by the traditional evaluation. In addition, it is interesting to see that the AP metrics are exactly between $ZP^{3,4}$ and $ZP^{4,5}$, which indicates that the detection performance in 96% of the image area is actually lower than AP.

Implications: The above results reveal the performance characteristics of the detectors, which help us better understand the behavior of the object detectors, and encourage us to rethink the choice of detectors when deploying to the application scenarios. Furthermore, it is also worth studying which components in the detection pipelines lead to such performance discrepancies, e.g., self-attention mechanism, label assignment, etc.

6.3 Zone evaluation on various datasets

Table 2 reports the quantitative detection results on PASCAL VOC 07+12, MS COCO val2017, and 3 application datasets. We have the following observations:

1) One can see that the detection performance varies across the zones. The zone closest to the image border, i.e., $z^{0,1}$, has consistently the lowest detection performance. In contrast, the central zone $z^{4,5}$ has the highest performance in almost all of these cases.

2) There is also a representative case, i.e., the helmet dataset, having only 3.0 ZP variance. This indicates that the helmet dataset achieves the best spatial equilibrium under the scenario of annular zone partition, whereas the other datasets have a clear spatial disequilibrium problem. For example, the variance of ZP is 53.6 on PASCAL VOC and 56.2 on the fruit dataset.

3) If we switch to other zone partitions, e.g., 5 strip zones along x-axis and 5 strip zones along y-axis (see Fig. 12(a)(b)), their spatial equilibria change. In Table 3, the face mask and helmet datasets increase ZP variance to 39.6 and 30.5 in the scenario of 5 zones along y-axis, respectively, while the fruit dataset decreases the ZP variance significantly in both cases.

Implications: The zone evaluation provides a fresh perspective that sheds light on the limitations of object detectors. It can be seen that spatial bias is also a natural characteristic of object detectors and they are difficult to perform perfect spatial equilibrium for an arbitrary zone partition. The above results indicate that the ZP variance is a set function related to zone partition. The annular zone partition mainly considers the balance of detection ability between the inner zone and the outer one, which is a good choice in practice since the centralized photographer's bias is ubiquitous in visual datasets. Nevertheless, it should be

TABLE 5
Analysis of spatial weight. 5 zone precisions (ZP), the variance of ZPs, and the traditional metric AP are reported. $\gamma = 0.2$. Lower variance, better spatial equilibrium. (Dataset: VOC 07+12)

weight	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
0	52.2	53.6	34.3	39.6	42.5	46.6	56.1
1	**52.8**	48.3	35.7	40.2	43.3	**47.1**	**56.2**
$\alpha(x^a, y^a)$	**52.8**	**37.7**	**37.6**	**40.3**	**43.8**	46.9	55.4

TABLE 6
Zone evaluation on PASCAL VOC 07+12, MS COCO 2017, and 3 application datasets, including face mask detection, fruit detection, and helmet detection. 5 zone precisions (ZP), the variance of ZPs, and the traditional metric AP are reported. The detector is GFocal [54].

Dataset	SELA	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
VOC 07+12		52.2	53.6	34.3	39.6	42.5	46.6	56.1
	✓	52.8	37.7	37.6	40.3	43.8	46.9	55.4
COCO 2017		40.1	16.9	31.1	37.5	39.4	38.5	43.8
	✓	40.3	14.4	31.2	37.7	39.5	38.3	42.9
Face Mask		71.3	13.1	60.4	67.1	69.0	68.8	70.9
	✓	71.6	12.1	60.6	68.0	69.5	69.3	69.8
Fruit		76.6	56.2	60.8	69.9	71.2	75.3	83.8
	✓	77.0	33.6	65.7	69.8	72.0	76.2	82.7
Helmet		49.7	3.0	45.9	47.9	50.3	50.6	47.8
	✓	49.9	3.1	45.9	48.5	50.5	50.6	47.9

noted that the zone partition is flexible and is able to be customized to any shape based on the application scenarios.

6.4 Evaluation on spatial equilibrium learning

Finally, we provide the evaluation of spatial equilibrium learning. The ablation studies are conducted by using GFocal and we adopt the first approach, i.e., spatial equilibrium label assignment (SELA), by default.

Hyperparameter γ. Recall that the implementation of SELA only involves one hyper-parameter γ in Eq. (6). γ controls the magnitude of the spatial weight. A larger γ increases more positive samples for objects near the image borders. As shown in Table 4, we observe that our SELA can achieve a consistent spatial equilibrium improvement (lower variance) for all the options of γ. Too large γ, e.g. 0.4, will increase much more positive samples for all the zones, leading to a performance drop. Thus, we set γ to 0.2 for PASCAL VOC. One can see that our SELA can significantly improve the detection performance for the outer zones, e.g., $ZP^{0,1}$, $ZP^{1,2}$, $ZP^{2,3}$ and $ZP^{3,4}$. As shown in Fig. 14(a), while the central zone $z^{4,5}$ has a slight performance drop, the ZP improvement is remarkable in the border zone, which occupies 96% of the total image area. This is particularly important for the safety applications in surveillance systems and self-driving cars, as objects may appear anywhere. The performance of the border zone plays a significant role in robustness detection. In practice, we set $\gamma = 0.1$ for all the other datasets, but it should be noted that there might be a better γ for different application scenarios.

Spatial weight. One may wonder how the performance would go if we directly loose the selection condition for the positive samples without considering their spatial positions. Here, we conduct an experiment to investigate the effect of the spatial weight. The quantitative results are reported in Table 5. If the spatial weight is set to a constant of 1, it means that we directly lower the positive IoU threshold t by $\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t - \gamma$, and more positive samples will be selected without spatial discrimination. One can see that although the performance is boosted, the variance of the 5 ZPs is large. This indicates that subtracting a constant from the positive IoU threshold cannot change the sampling frequency much, because more positive samples are generated in the central zone. In contrast, our SELA can significantly reduce the variance, and achieve a much better spatial equilibrium.

TABLE 7
Evaluation of SELA with various backbone networks. 5 zone precisions (ZP), the variance of ZPs, and the traditional metric AP are reported. **X**: ResNeXt [92]. (Dataset: VOC 07+12)

Model	SELA	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
ResNet-18		52.2	53.6	34.3	39.6	42.5	46.6	56.1
	✓	52.8	37.7	37.6	40.3	43.8	46.9	55.4
ResNet-50		56.1	41.5	40.9	44.6	46.7	51.0	59.7
	✓	56.2	32.2	43.3	44.6	47.3	50.4	59.2
X-101-32x4d-DCN		64.0	37.1	48.7	53.1	55.0	58.0	66.9
	✓	64.3	31.0	50.2	54.1	55.9	57.7	66.9

TABLE 8
Generality of spatial equilibrium learning. The cost-sensitive learning based method, SE loss, is adopted. 5 zone precisions (ZP), the variance of ZPs, and the traditional metric AP are reported. (Dataset: VOC 07+12)

Detector	SE loss	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
GFocal [54]		52.2	53.6	34.3	39.6	42.5	46.6	56.1
	✓	52.5	41.6	37.1	40.6	42.9	46.5	56.0
DW [53]		51.8	32.6	38.4	39.9	43.3	45.7	54.6
	✓	52.7	25.9	39.8	41.2	44.4	46.8	54.2
DDOD [15]		51.1	22.6	38.4	40.0	42.2	45.2	51.9
	✓	51.5	20.8	40.9	40.1	42.6	45.8	52.7
DINO [100]		61.5	47.6	47.1	48.4	53.0	57.1	66.2
	✓	61.7	46.7	47.4	48.5	53.4	57.1	66.3

SELA on various datasets. Table 6 shows us promising results that our SELA can achieve a better spatial equilibrium for object detection. In particular, we reduce the variance by a large margin in terms of ZP. For example, we successfully lower the variance of ZP by -15.9, -2.5, -1.0, and -22.6 on PASCAL VOC, MS COCO, and face mask/fruit detection. It demonstrates that our SELA can improve the spatial equilibrium for multiple application scenarios without the sacrifice of AP.

Generality of spatial equilibrium learning. We further provide more experiments to verify the effectiveness of spatial equilibrium learning on various backbone networks. Table 7 exhibits that our SELA can notably improve the spatial equilibrium for all the 3 backbone networks, i.e., lower variance. We also conduct experiments to check out the generality of spatial equilibrium learning by incorporating it into 3 more detectors, DW [53], DDOD [15], and DETR-like detector DINO [100]. Here, we adopt spatial equilibrium loss (SE loss) and we enlarge the training losses for the objects near the image border. Table 8 reports the quantitative results of SE loss for these 4 object detectors. As shown, our method can substantially reduce the ZP variance for the 4 detectors, indicating that a better spatial equilibrium is achieved. This shows the generalized ability of our method to improve the spatial robustness of detectors without bells and whistles.

We also note that compared to the CNN-based object detectors, our method produces a slight improvement of

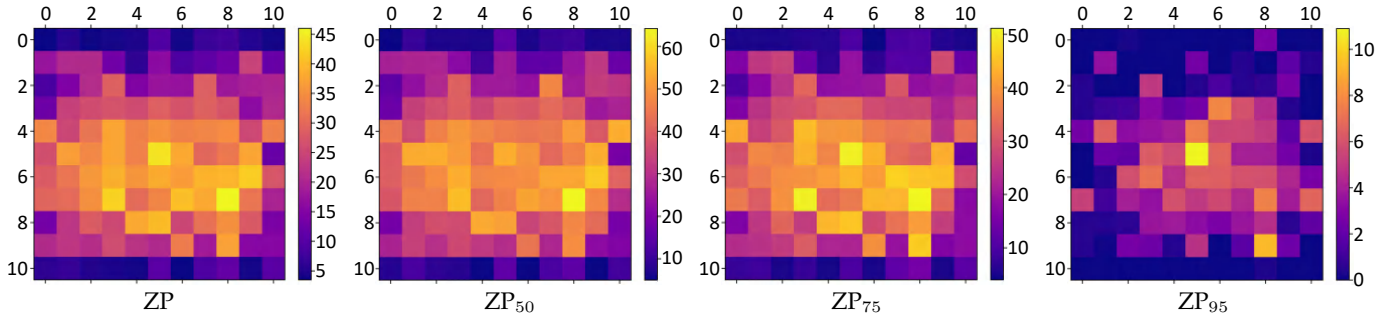
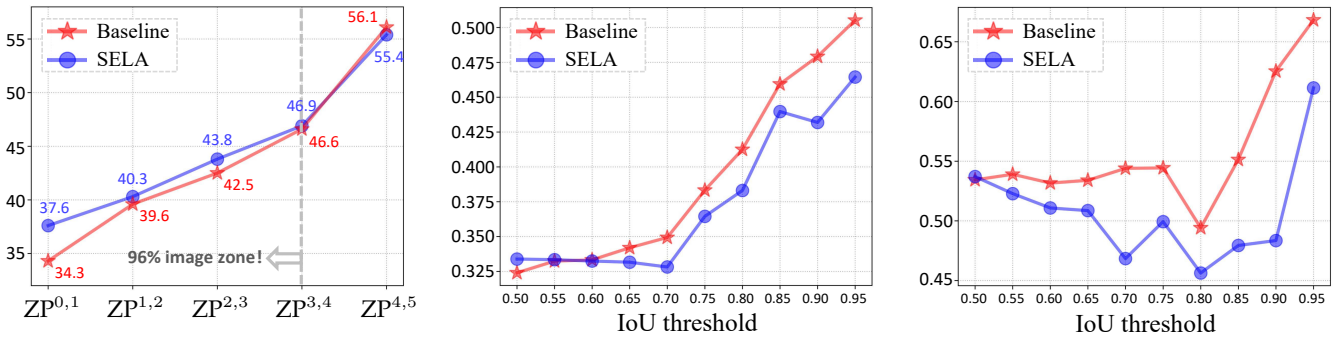


Fig. 13. Zone evaluation on 11×11 square zones. The model is GFocal. The results are reported on VOC 07+12.



(a) Effect of SELA (b) PCC (c) SCC

Fig. 14. (a) ZP against the zone. (b) Pearson Correlation Coefficient (PCC) between the mZP and the object distribution (center counts) against the IoU threshold. (c) Spearman Correlation Coefficient (SCC) between the mZP and the object distribution against the IoU threshold. Our SELA can substantially reduce these correlations under most of IoU thresholds, indicating a better spatial equilibrium. The baseline model is GFocal. The results are reported on VOC 07+12.

spatial equilibrium on DINO. This may be attributed to the different optimization processes between DETR-like detectors and others. The number of positive samples is quite limited in DETR-like detectors since they use one-to-one Hungarian matching, while it is much more abundant in dense object detectors as they adopt one-to-many assignments. Thus, our SE loss is more helpful in alleviating the imbalance of supervision signal strength on dense object detectors. This implies that improving spatial equilibrium for DETR-like detectors could be more challenging. We hope this work could inspire more solutions to address the disequilibrium problem of DETR-like detectors in the future.

Correlation with object distribution. We further provide the correlation between the zone metrics and the object distribution. We define a finer zone partition, which is the same as the zone partition for counting the centers of objects, i.e., 11×11 square zones (see Fig. 11). Then we evaluate the detection performance in the 121 zones one by one. We plot the ZP of the 121 zones in Fig. 13. It can be seen that the ZP distribution is similar to the object distribution (Fig. 11), i.e., the same centralized trend. To investigate the correlation between the zone metrics and the object distribution, we further calculate the Pearson Correlation Coefficient (PCC) and the Spearman Correlation Coefficient (SCC) between the mZPs and the object distribution of the test set. As shown in Fig. 14(b) and Fig. 14(c), we get the following deep reflections on the spatial bias. We first note that all the PCCs > 0.3 in Fig. 14(b), which indicates that the detection performance is moderately linear correlated with the object distribution. As a reminder, the PCC only reflects the linear correlation of two given vectors, while it may fail when they are curvilinearly correlated. In Fig. 14(c), the Spearman correlation reflects a higher ranking correlation between the mZPs and the object distribution with all the SCCs > 0.45. This illustrates that the detection performance has a moderate-to-high correlation with the object distribution. In general, our SELA substantially reduces these correlations, indicating a lower correlation with object distribution and better spatial equilibrium.

Visualization of Detection. We visualize the detection results of SELA in Fig. 15. Our method can improve the detection performance of the border zone. We believe that further exploration of spatial equilibrium is clearly worthy and important for robust detection applications.

6.5 Other attempts toward spatial equilibrium

As we discussed in Sec. 4, our results show that the object scale and the absolute positions of objects barely influence the spatial bias. The discrepancy of object data patterns between the zones plays a significant role in spatial bias. In this subsection, we investigate more possible factors and see how the spatial equilibrium changes. The first one is the padding operation. We follow the work [45] to set the padding way as full-conv, which is translation-invariant as [45] demonstrated. We replace all the convolution kernels of the head networks with full-conv. The second is the effect of the oversized anchor near the image border. The default setting of the baseline model keeps all the anchors. We remove all the anchors whose edges are outside the valid image. The third is the image resolution. The default resolution of the baseline model is 1333×800, and we train a model with a small one, e.g., 640×640. The results are reported

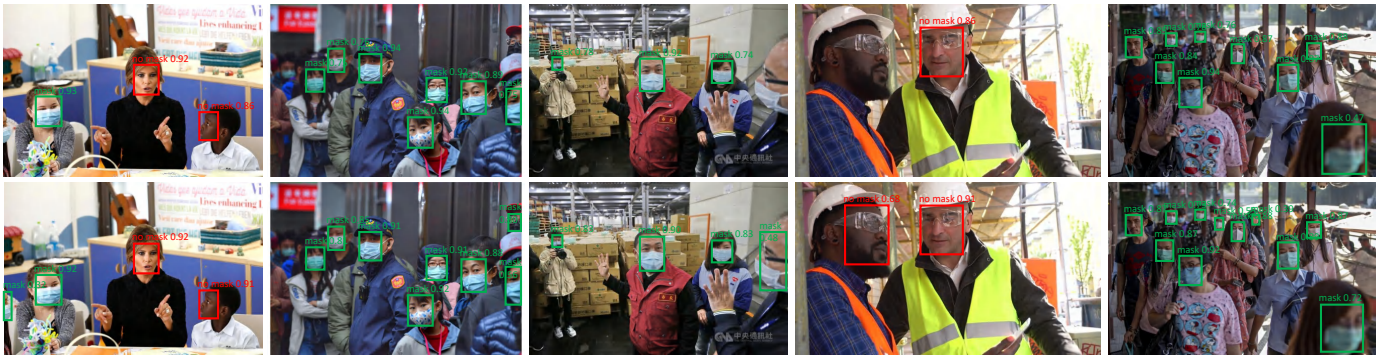


Fig. 15. Illustration of detection results for GFocal (first row) and GFocal + SELA (second row). Our method boosts the detection performance for the border zone. Zoom in for a better view.

in Table 9. One can see that the three modifications result in a significant AP drop. The full-conv padding can lower the ZP variance, but it is not helpful for detection accuracy. In addition, it cannot obtain a better spatial equilibrium by removing the out-of-bounds anchors or setting different image resolutions, because the supervision signal remains imbalanced between the zones. It is challenging to find a solution that can alleviate spatial disequilibrium problem without causing performance degradation.

TABLE 9
Evaluation of 3 potential factors for spatial bias. (1) We use full-conv [45] padding in the head network; (2) We remove the oversized anchor boxes that exceed the valid image border; (3) We set the image resolution as 640×640. 5 zone precisions (ZP), the variance of ZP, and the traditional metric AP are reported. The results are reported on GFocal [54] on COCO val2017.

Modification	AP	Var.	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
Baseline	40.1	16.9	31.1	37.5	39.4	38.5	43.8
(1)	38.6	12.4	30.4	36.0	37.5	37.8	41.2
(2)	38.5	16.7	28.8	35.8	37.8	37.2	41.3
(3)	36.9	18.8	26.7	33.6	36.2	34.9	39.9

7 Conclusions, Challenges and Outlook

In this paper, we present zone evaluation to reveal the existence and the discrete amplitude of spatial bias in modern object detectors. We find that the spatial bias is less relevant to the object scale and the absolute positions of objects, but closely related to the gap of object data patterns between the zones. Based on the thorough study of the origin of spatial bias, we finally introduce the spatial disequilibrium problem, aiming at a robust detection across the zones. We also show a path to spatial equilibrium object detection, as a start to alleviate this problem. Extensive experiments manifest the existence and the major source of spatial bias, which is prevalent in various modern detectors and datasets.

Significance. Spatial bias is a natural obstacle in object detection that the detectors usually show a performance drop in the border zone, which occupies a large proportion of the image area. While the classic AP metric is still considered to be the primary measurement, it is difficult to reveal spatial bias and is challenging to comprehensively reflect the real performance of object detectors. Maximizing the AP metric does not fully indicate a robust detection and performs well in all zones. Zone evaluation supplements a series of zone metrics, compensates for the drawbacks of traditional evaluation, and captures more information about detection performance. We hope this work could inspire the community to rethink the evaluation of object detectors and stimulate further explorations on spatial bias, and the solutions to the spatial disequilibrium problem.

There are several challenges left behind in this work:

Interpretability of spatial bias in various object detectors. This paper mainly reveals the existence and the discrete amplitude of spatial bias in object detectors, whereas the specific reason why different detectors perform quite differently is still frozen in the ice. The neural network architecture designs, pre-training data, optimization, training strategies, and even hyper-parameters may play a role in the spatial bias. Further exploration to answer the above question is of paramount importance.

The effect of other potential factors on the spatial bias. Currently, we pinpointed an evident correlation between imbalanced object distribution and zone performance. There are some complicated yet implicit factors such as image blur, object occlusion, border effect, noise, etc., that may also contribute to spatial bias. However, current detection datasets almost lack such annotations for the above factors, making it difficult to establish a quantitative analysis.

Zone evaluation for other vision tasks. Researchers have found some clues that the image generator may generate distorted content near the image border [18]. Hence, the spatial bias may also exist in many vision tasks. Our zone evaluation may have great potential to reveal spatial bias, whether for high-level or low-level vision tasks.

References

[1] Ahmad Abdulkader. https://www.kaggle.com/datasets/parot99/face-mask-detection-yolo-darknet-format.

[2] Alexander. https://www.kaggle.com/datasets/vodan37/yolo-helmethead/metadata.

[3] Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. Mind the pad–cnns can develop blind spots. In *Int. Conf. Learn. Represent.*, 2021.

[4] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

[5] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M Paixao, Filipe Mutz, et al. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021.

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6541–6549, 2017.

[7] Muhammad Tahir Bhatti, Muhammad Gufran Khan, Masood Aslam, and Muhammad Junaid Fiaz. Weapon detection in real-time cctv videos using deep learning. *IEEE Access*, 9:34366–34382, 2021.

[8] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

[9] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6154–6162, 2018.

[10] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin. Prime sample attention in object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11583–11591, 2020.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229, 2020.

[12] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3773–3783, 2021.

[13] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[14] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Eur. Conf. Comput. Vis.*, pages 520–535, 2018.

[15] Zehui Chen, Chenhongyi Yang, Qiaofei Li, Feng Zhao, Zheng-Jun Zha, and Feng Wu. Disentangle your dense object detector. In *ACM Int. Conf. Multimedia*, pages 4939–4948, 2021.

[16] Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12925–12935, 2020.

[17] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Int. Conf. Comput. Vis.*, pages 502–511, 2019.

[18] Jooyoung Choi, Jungbeom Lee, Yonghyun Jeong, and Sungroh Yoon. Toward spatially unbiased generative models. In *Int. Conf. Comput. Vis.*, pages 14233–14242, 2021.

[19] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *Eur. Conf. Comput. Vis. 2020 Workshops*, pages 95–110, 2020.

[20] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *Eur. Conf. Comput. Vis.*, pages 694–710, 2020.

[21] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9268–9277, 2019.

[22] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.

[23] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *Int. Conf. Comput. Vis.*, pages 1851–1860, 2017.

[24] Eunpyohong. https://www.kaggle.com/datasets/eunpyohong/fruit-object-detection.

[25] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88:303–338, 2010.

[26] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Int. Conf. Comput. Vis.*, pages 4548–4557, 2017.

[27] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022.

[28] Xin Feng, Tao Liu, Dan Yang, and Yao Wang. Saliency based objective quality assessment of decoded video affected by packet losses. In *2008 15th IEEE International Conference on Image Processing*, pages 2560–2563. IEEE, 2008.

[29] Rony Ferzli and Lina J Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE transactions on image processing*, 18(4):717–728, 2009.

[30] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[31] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia*, 18(6):1098–1110, 2016.

[32] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Eur. Conf. Comput. Vis.*, pages 126–143, 2022.

[33] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[35] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):815–828, 2019.

[36] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5375–5384, 2016.

[37] Lida Huang, Gang Liu, Yan Wang, Hongyong Yuan, and Tao Chen. Fire detection in video surveillances using convolutional neural networks and wavelet transform. *Engineering Applications of Artificial Intelligence*, 110:104737, 2022.

[38] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Eur. Conf. Comput. Vis.*, pages 532–546, 2018.

[39] Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G Derpanis, and Neil DB Bruce. Global pooling, more than meets the eye: Position information is encoded channel-wise in cnns. In *Int. Conf. Comput. Vis.*, pages 793–801, 2021.

[40] Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G Derpanis, and Neil DB Bruce. Position, padding and predictions: A deeper look at position information in cnns. *arXiv preprint arXiv:2101.12322*, 2021.

[41] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, Zeng Yifu, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Max Strobel, Mrinal Jain, Lorenzo Mammana, and xylieong. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations, August 2022.

[42] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Int. Conf. Learn. Represent.*, 2020.

[43] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019.

[44] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020.

[45] Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14274–14285, 2020.

[46] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13896–13905, 2020.

[47] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1646–1654, 2016.

[48] Lyudmyla Kirichenko, Tamara Radivilova, Bohdan Sydorenko, and Sergiy Yakovlev. Detection of shoplifting on video using a hybrid network. *Computation*, 10(11):199, 2022.

[49] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[50] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010.

[51] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *AAAI Conf. on Artif. Intel.*, pages 8577–8584, 2019.

[52] Chaofeng Li and Alan C Bovik. Three-component weighted structural similarity index. In *Image quality and system performance VI*, volume 7242, pages 252–260. SPIE, 2009.

[53] Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. A dual weighting label assignment scheme for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9387–9396, 2022.

[54] 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. In *Adv. Neural Inform. Process. Syst.*, pages 21002–21012, 2020.

[55] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10991–11000, 2020.

[56] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Int. Conf. Comput. Vis. Workshops*, pages 1833–1844, 2021.

[57] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2117–2125, 2017.

[58] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, pages 2980–2988, 2017.

[59] 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 *Eur. Conf. Comput. Vis.*, pages 740–755, 2014.

[60] Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011.

[61] Tsung-Jung Liu and Kuan-Hsien Liu. No-reference image quality assessment by wide-perceptual-domain scorer ensemble method. *IEEE Transactions on Image Processing*, 27(3):1138–1151, 2017.

[62] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Eur. Conf. Comput. Vis.*, pages 21–37, 2016.

[63] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11976–11986, 2022.

[64] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2537–2546, 2019.

[65] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. In *Int. Conf. on Mach. Learn. Worksh.*, 2019.

[66] F Lukas and Z Budrikis. Picture quality prediction based on a visual model. *IEEE Transactions on Communications*, 30(7):1679–1692, 1982.

[67] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Eur. Conf. Comput. Vis.*, pages 181–196, 2018.

[68] Marco Manfredi and Yu Wang. Shift equivariance in object detection. In *Eur. Conf. Comput. Vis. Workshops*, pages 32–45, 2020.

[69] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[70] Sanam Narejo, Bishwajeet Pandey, Doris Esenarro Vargas, Ciro Rodriguez, and M Rizwan Anjum. Weapon detection using yolo v3 for smart surveillance system. *Mathematical Problems in Engineering*, 2021:1–9, 2021.

[71] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3388–3415, 2020.

[72] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 864–873, 2016.

[73] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: Towards balanced learning for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 821–830, 2019.

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

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, pages 91–99, 2015.

[76] Sergio Saponara, Abdussalam Elhanashi, and Alessio Gagliardi. Real-time video fire/smoke detection based on cnn in antifire surveillance systems. *Journal of Real-Time Image Processing*, 18:889–900, 2021.

[77] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Int. Conf. Comput. Vis.*, pages 8430–8439, 2019.

[78] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[79] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 761–769, 2016.

[80] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Int. Conf. Comput. Vis.*, pages 3365–3374, 2021.

[81] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14454–14463, 2021.

[82] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE signal processing letters*, 24(9):1408–1412, 2017.

[83] Gergely Szabó and András Horváth. Mitigating the bias of centered objects in common datasets. In *ICPR*, pages 4786–4792, 2022.

[84] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1521–1528, 2011.

[85] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1974–1983, 2021.

[86] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2965–2974, 2019.

[87] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, and Ming Tang. Adaptive class suppression loss for long-tail object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3103–3112, 2021.

[88] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In *Int. Conf. Comput. Vis.*, pages 568–578, 2021.

[89] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, pages 7032–7042, 2017.

[90] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep

networks. In *Adv. Neural Inform. Process. Syst.*, pages 12635–12644, 2019.

[91] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[92] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1492–1500, 2017.

[93] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13569–13578, 2021.

[94] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013.

[95] Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15819–15829, 2021.

[96] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *Int. Conf. on Mach. Learn.*, pages 11830–11841, 2021.

[97] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. In *Adv. Neural Inform. Process. Syst.*, pages 18381–18394, 2021.

[98] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5728–5739, 2022.

[99] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020.

[100] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: Detr with improved denoising anchor boxes for end-to-end object detection. In *Int. Conf. Learn. Represent.*, 2023.

[101] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: An iou-aware dense object detector. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8514–8523, 2021.

[102] Quanshi Zhang, Wenguan Wang, and Song-Chun Zhu. Examining cnn representations with respect to dataset bias. In *AAAI Conf. on Artif. Intel.*, pages 4464–4473, 2018.

[103] Richard Zhang. Making convolutional networks shift-invariant again. In *Int. Conf. on Mach. Learn.*, pages 7324–7334, 2019.

[104] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9759–9768, 2020.

[105] Wei Zhang, Ali Borji, Zhou Wang, Patrick Le Callet, and Hantao Liu. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1266–1278, 2015.

[106] Wei Zhang and Hantao Liu. Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications. *IEEE Transactions on Image Processing*, 26(5):2424–2437, 2017.

[107] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

[108] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 589–597, 2016.

[109] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Int. Conf. Comput. Vis.*, pages 8779–8788, 2019.

[110] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2005.

[111] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *Int. Conf. Learn. Represent.*, 2021.

Zhaohui Zheng received the M.S. degree in computational mathematics from Tianjin University in 2021. He is currently a Ph.D. candidate at Media Computing Lab, Nankai University, supervised by Prof. Ming-Ming Cheng. His research interests include object detection, instance segmentation and knowledge distillation. He received the CCF-CV Academic Emerging Scholar Award in 2023.

Yuming Chen received his B.E. degree in computer science from Lanzhou University in 2022. He is currently a master student at the Media Computing Lab, Nankai University, under the supervision of Prof. Ming-Ming Cheng and Prof. Qi-Bin Hou. His research interests include object detection and knowledge distillation.

Qibin Hou received his Ph.D. degree from the School of Computer Science, Nankai University. Then, he worked at the National University of Singapore as a research fellow. Now, he is an associate professor at School of Computer Science, Nankai University. He has published more than 30 papers on top conferences/journals, including T-PAMI, CVPR, ICCV, NeurIPS, etc. His research interests include deep learning and computer vision.

Xiang Li is an Associate Professor in College of Computer Science, Nankai University. He obtained the Ph.D. degree from Nanjing University of Science and Technology, Jiangsu, China, in 2020. His research interests include CNN/Transformer backbone, object detection, knowledge distillation and self-supervised learning. He has published 30+ papers in top journals and conferences such as TPAMI, CVPR, NeurIPS, etc.

Ping Wang received the B.S., M.S., and Ph.D. degrees in computer science from Tianjin University, Tianjin, China, in 1988, 1991, and 1998, respectively. She is currently a Professor with the School of Mathematics, Tianjin University. Her research interests include image processing and machine learning.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests include computer graphics, computer vision, and image processing. He received research awards including National Science Fund for Distinguished Young Scholars and ACM China Rising Star Award. He is on the editorial boards of IEEE TPAMI and IEEE TIP.