Zone Evaluation: 揭示目标检测的空间偏差性

郑兆晖, 陈宇铭, 侯淇彬, *IEEE* 会员, 李翔, *IEEE* 会员, 王萍,
和程明明, *IEEE* 高级会员

摘要—目标检测器的一个根本性局限性是，它们会遭受“空间偏差”的影响，尤其是在检测图像边界附近的目标时，性能会显著下降。长期以来，目标检测领域一直缺乏有效的方法来测量和识别空间偏差，对于这种偏差的来源和程度也知之甚少。为了解决这个问题，我们提出了一种新的区域评估协议，从传统评估扩展到更通用的评估方法。该协议通过测量不同区域的检测性能，产生一系列区域精度 (Zone Precisions, ZPs)。我们首次提供了数值结果，结果表明，目标检测器在不同区域的性能表现非常不均衡。更令人惊讶的是，检测器在图像 96% 的边界区域的性能甚至达不到 AP 值（平均精度，通常被认为是整个图像区域的平均检测性能）。为了更好地理解空间偏差，我们进行了一系列启发式实验。我们的研究排除了关于空间偏差的两个直观猜想，即目标尺度和目标的绝对位置几乎不会影响空间偏差。我们发现，关键在于不同区域目标数据模式之间，人类难以察觉的差异，这些差异最终导致了区域之间明显的性能差距。基于这些发现，我们最终讨论了目标检测的未来方向，即空间不均衡问题，旨在追求在整个图像区域内实现均衡的检测能力。通过广泛评估 10 种流行的目标检测器和 5 个检测数据集，我们揭示了目标检测器的空间偏差问题。我们希望这项工作能够引起对检测鲁棒性的关注。源代码、评估协议和教程已在 https://github.com/Zzh-tju/ZoneEval 上公开。

Index Terms—目标检测，区域评估（zone evaluation），空间偏差，空间不均衡问题，空间均衡学习。

1 引言

目标检测在过去二十年中取得了令人瞩目的进展 [11], [62], [74], [75]。虽然目标检测器的优化流程已被充分探索，但它们在局部图像区域内的行为仍然是一个谜。检测器的空间鲁棒性尤其重要 [83]，因为目标可能出现在任何位置，并且所有目标都应该被良好地检测到。这对于安全视觉应用尤为重要，例如，火灾/烟雾检测 [37], [76]、自动驾驶汽车中的防撞 [5], [17]、人群计数和定位 [38], [80], [85], [108]、智能监控系统中的武器检测 [7], [70] 以及商店盗窃检测 [48] 等，在这些应用中，边界区域占据了图像区域的很大一部分。

不幸的是，检测器实际上无法在空间区域内均匀地执行检测，通常在图像边界附近表现出明显的性能下降。这种现象，我们称之为“空间偏差”，是目标检测中一个天然的障碍，但在很长一段时间内都被检测社区所忽视。忽视这个问题可能会导致严重的安全隐患和重大财产损失的风险。例如，火灾探测器可能擅长检测中心区域的火灾，但会失去检测图像边界区域火灾的能力。这样的火灾报警系统是不可靠的，因为镜头中心区域仅占图像区域的一小部分。卷积神经网络 (CNN) 在空间鲁棒性方面的一些最新突破 [4], [12], [45], [103], 正朝着难以捉摸的平移不变性方向发展，其基础是理解小的图像变换（例如，颜色抖动、平移）如何影响分类精度。研究发现，即使对于同一个目标，分类器也会随着其空间位置的变化做出完全不同的预测 [45]。除了图像分类，我们在本文中深入研究了目标检测中的空间偏差，揭示了现代目标检测器的局限性。

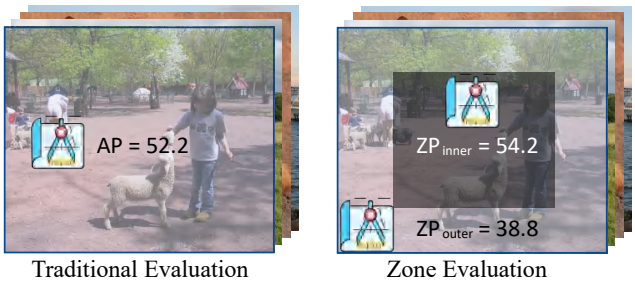


图 1. 传统的评估方法衡量整个图像区域的检测性能，但它忽略了对局部区域的测量，难以反映空间偏差。我们的区域评估 (ZP，区域精度，即区域内约束的平均精度) 弥补了这些问题，表明区域之间存在很大的性能差距。结果由 GFocal [54] 在 VOC 2007 测试集 [25] 上报告。

多年来，一直存在一个开放性问题，即缺乏一种有效的方法来测量和识别空间偏差。传统的评估方法，即 AP 指标，衡量的是整个图像区域的检测性能，这并没有为检测器的空间鲁棒性提供任何指导，人们也很难知

- 郑兆晖, 陈宇铭, 侯淇彬, 李翔, 和程明明均就职于中国天津南开大学计算机科学学院 *VCIP* 研究组 *(通讯作者：侯淇彬)*.
- 王萍就职于中国天津大学数学学院.
- 本研究由国家自然科学基金 *(项目编号：62225604, 62276145, U23B2049)*，中央高校基本科研业务费 *(南开大学, 070-63223049)*，和中国科协青年人才托举工程 *(项目编号：YESS20210377)* 资助。计算资源由南开大学高性能计算中心 *(NKSC)* 提供支持。

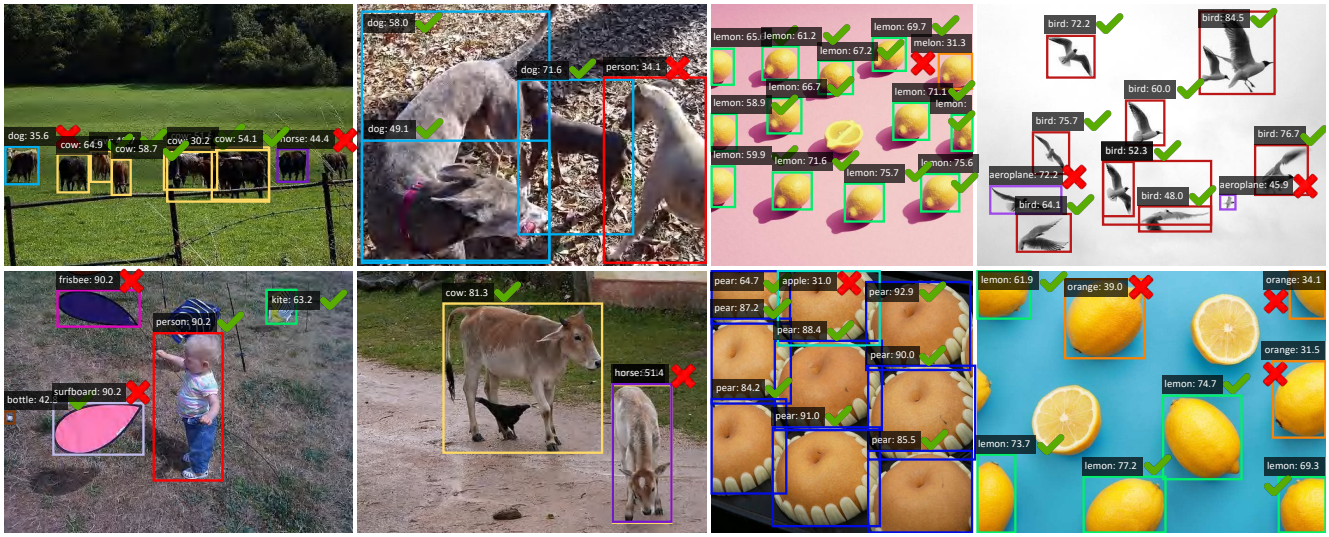


图 2. 检测器在检测边界目标时不太理想。可视化结果由 GFocal [54] 报告。放大以获得更好的视图。

道性能在何处以及下降了多少。因此，评估协议显得尤为重要，因为它可以提供机会更好地理解空间偏差，并为进一步构建方法论提供工具。为此，我们提出了一种系统的方法，称为区域评估，来分析现代目标检测器中是否存在空间偏差，以及如果存在，偏差有多大。

具体而言，我们将传统的全图评估扩展到更通用的评估。我们计算指定区域内的通用平均精度 (AP) [59]，从而得出区域精度 (Zone Precision, ZP)。在评估期间，仅考虑中心位于该区域内的框。借助这些辅助指标，我们首次提供了数值结果，明确揭示了目标检测器实际上在不同区域之间存在相当大的空间偏差。如 图 1 所示，内部区域和外部区域之间的 ZP 差距为 15.4。表面上看，我们可以从这种性能差距中推断出，检测能力似乎与目标的绝对位置高度相关。然而，当我们移动图像中的目标时，这种看似合理的推测在实践中存在许多根本上的不一致之处。我们没有意识到对边界区域性能下降的任何令人满意的解释（见 图 2）。因此，我们希望阐明目标检测器中空间偏差的存在和主要来源。最后，我们提出了未来目标检测研究的一个重点：走向空间均衡。

这项工作的贡献主要包括对空间偏差的三个探索性实验、一个潜在的研究方向以及对现代目标检测器的全面评估。

目标尺度是否在中心区域性能中起关键作用？ (节 4.1) 我们的答案是否定的。虽然大型目标相对频繁地出现在中心区域，但我们观察到，当我们很大程度上消除目标尺度的影响时，不同区域之间的区域性能仍然非常不均匀。空间偏差和目标尺度之间仍然难以建立必然的联系。

检测器是否根据目标的绝对空间位置产生中心区域性能？ (节 4.2) 我们的答案是否定的。检测性能几乎与目标的空间位置无关。我们观察到，当目标移动到中心区域时，这在统计学上不会导致检测质量的提高。

什么才是真正决定目标检测器空间偏差的因素？ (节 4.3) 我们的分析揭示了强有力的证据，表明空间偏差与区域之间目标数据模式的差异高度相关。具体而言，中心目标和边界目标之间在目标数据模式上存在差异。因此，如果一个目标是从中心区域风格的数据分布中采样，而不是从边界区域风格的数据分布中采样，则可以更好地检测到该目标。

一个新的未来方向：空间不均衡问题。 (节 5) 这项工作更进一步，提出了一个迫切需要解决的实际问题，即空间不均衡问题。在这种问题设置下，目标检测器将空间均衡作为重要目标之一，这对鲁棒检测具有至关重要的意义。面对这一挑战，我们还提供了首次尝试，即空间均衡学习，以实现空间均衡目标检测。(节 5.2)

全面的评估。 (节 6) 我们提供了对几种代表性目标检测器的广泛评估和比较。我们通过实验揭示：1) 空间偏差在各种目标检测器和数据集中非常普遍。2) 检测器的空间均衡性差异很大。特别是，我们将展示稀疏检测器在中心区域表现更好，而单阶段密集检测器在边界区域表现更好。3) 提出的空间均衡学习能够缓解空间不均衡问题。

其余部分安排如下：节 2 简要回顾研究背景。节 3 介绍了区域评估。节 4 研究了空间偏差的主要来源。节 5 介绍了空间不均衡问题，并提出了空间均衡学习来缓解这个问题。节 6 给出了区域评估和空间均衡学习的广泛

评估。

2 背景

2.1 与数据不平衡问题的关系

令 $\{X, G\} = \{x_i, g_i\}_{i=1}^n$ 为样本-标签对的集合，其中每个样本 x_i 都有一组真实标签 g_i。模型训练在 $\{X, G\}$ 的子集上进行，网络在每个训练 epoch 遍历训练集。数据不平衡问题通常与 $\{X, G\}$ 的内在属性有关。在目标检测文献中，主要讨论两种广泛存在的不平衡问题。

第一个是类别不平衡问题。在这种情况下，样本 X 根据类别划分为多个子集 $X_1, X_2, \cdots, X_c$，其中跨 c 个类别的样本数量不平衡，从而产生长尾分布 [55], [72], [87], [107]。类别不平衡问题自然会导致训练期间的采样不平衡，阻碍了尾部类别的分类性能。重采样策略 [42], [67] 和代价敏感学习 [21], [110] 是类别重平衡的主流范例。

第二个是前景-背景采样不平衡。这种不平衡也源于数据本身。大量的锚点平铺在背景区域，这些锚点自然地被采样为负样本，因此主导了梯度流。在这种情况下，X 可以分为 X_{neg} 和 X_{pos}，使得 $X = X_{neg} \bigcup X_{pos}$。负样本 X_{neg} 可以看作是正样本 X_{pos} 的互补集，其真实标签是 “背景”，没有边界框注释。解决这个问题的方法类似，包括重采样，例如，OHEM [79]、Guided Anchoring [86] 和 IoU 平衡采样 [73]，以及代价敏感学习，例如，Focal loss [58]、GHM loss [51] 和 PISA [10]。

相比之下，空间偏差也是目标检测中的一个障碍。鉴于此，我们为目标检测建立了一个新的空间不均衡问题。在这种情况下，样本 X 可以根据空间区域划分为多个子集，就像类别划分一样。一般来说，空间不均衡问题与类别不平衡问题具有相似的特征。不同之处在于，后者在类别之间存在长尾分布，而前者则考虑了目标在空间区域上的不均匀分布。我们将在 节 5.1 中展示，这两个问题在形式上是相互等价的。

2.2 CNN 中的鲁棒性

人们已经广泛讨论了平移不变性并未被深度 CNN 完全保持 [4], [40], [45], [93], [103]，因为它们忽略了经典的采样定理。一个小的图像变换可能导致预测的剧烈变化，从而阻碍分类器的鲁棒性。Zhang R. [103] 分析了最大池化算子的缺陷，并提出注入 anti-aliasing 以提高深度网络的鲁棒性。Lopes 等人 [65] 通过提出 patch Gaussian 增强，实现了更好的鲁棒性-准确率权衡。

在更长的空间范围上的鲁棒性方面，最近的研究 [3], [45] 表明，CNN 可以利用目标的绝对位置作为图像分类的附加信息。Islam 等人 [39] 进一步扩展了 CNN 基于通道维度的顺序编码位置信息。在 [18] 中，提出了一个空间上无偏差的 StyleGAN2 [44]，以解决由于人脸数据集 [43] 中摄影师的偏差而导致的图像边界中扭曲的人脸生成问题。Gergely 等人 [83] 经验性地发现，当移动图像以使目标更靠近图像边界时，分类精度会下降。Islam, M. A. 等人 [40] 揭示了语义分割中存在边界效应，其中车辆分割质量与区域内汽车的密度高度相关。Manfredi 等人 [68] 提出了一种通过将图像平移几个像素来测量目标检测器的平移等变性的 AP 变化贪婪近似方法，但这不可避免地需要数倍的推理时间才能完成评估。在这项工作中，我们首次从局部区域的角度数值量化了目标检测器的泛化能力，这有助于我们更好地理解空间偏差的存在和离散幅度。这为目标检测器的可靠性提供了一种新的分析工具。

2.3 从局部角度评估

从局部角度进行评估已被广泛证明在图像质量评估 (IQA) [99] 中具有优势，因为全局评估与人类视觉系统 (HVS) 不一致 [29], [60], [61], [78], [91], [105]。全局值无法反映空间非平稳模型的能力。早在 1982 年的早期研究 [66] 中就有人提出，如果使用局部测量而不是全局测量，质量度量可能会得到改进。在 IQA 中，通常采用两阶段结构。在第一阶段，局部评估图像质量。局部测量过程通常会生成质量图。为了将此类质量图转换为整体质量评分，在 IQA 的第二阶段应用池化算法。

Wang 等人提出了 Mean-SSIM [91] 来获得图像失真评估的空间平滑测量，其关键是计算每个滑动窗口的局部 SSIM，然后求平均值。3-SSIM [52] 为边缘、纹理和平滑区域分配不同的权重。Larson 等人 [50] 引入了可见性加权局部 MSE 来确定感知失真，其中图像被划分为 16×16 的块，相邻块之间有 75% 的重叠。NIQE [69] 索引引入了 patch 选择，以关注信息丰富的图像 patch。Chen 等人 [14] 提出使用 Landmark Distance (LMD) 来关注测量合成唇部运动的质量。Sun 等人 [82] 提出了加权到球形均匀 PSNR (WS-PSNR)，为不同的像素提供不同的权重。GMSD [94] 利用像素级梯度幅度相似性来捕获图像的局部质量。Fan 等人 [26] 为显著性目标检测提出了 S-measure，该方法首先将图像划分为 4 个正方形网格，并为每个局部 SSIM 分配不同的权重。Bosse 等人 [8] 尝试使用基于 CNN 的方法来学习局部图像质量。

一些方法 [28], [31], [105], [106] 将显著性图纳入 IQA 指标，因为显着的位置可以帮助预测人类观察者感知的图像质量。局部指标有助于描述图像 patch 之间的小细节和结构相似性。

尽管局部评估在计算机视觉应用的许多评估系统中已经流行了几十年，但在目标检测中尚未得到充分研究。大多数 IQA 方法主要用于像素预测任务，例如，图像恢复 [32], [56], [98]、显著性/伪装目标检测 [27], [35], [109] 和图像超分辨率 [22], [47]。它们不能直接应用于实例预测任务，例如，通用目标检测。因此，我们提出区域评估来填补这项研究空白。此外，我们的工作进行了一系列启发式实验，为理解现代目标检测器的空间偏差提供了新的见解。我们还研究了评估目标检测器时区域划分的几种形状（环形、条形、正方形），而之前的 IQA 方法很少关注这一点。

3 Zone Evaluation

在本节中，我们将传统的目标检测评估扩展到更通用的区域评估。给定一个测试图像 I 和一组评估指标 $\mathcal{M}$，经典的评估方法同时计算整个图像中所有检测结果和真实标签的指标。$\mathcal{M}$ 中的元素可以是 COCO 风格的 AP（平均精度）[59]，10 个 IoU 阈值上的 mAP，或小/中/大目标的 AP，这些都在目标检测中被广泛使用。这些传统指标衡量了整个图像区域的检测性能，但没有考虑目标检测器的空间鲁棒性。

区域指标。 令 $S = \{z^1, z^2, \cdots, z^n\}$ 为区域划分，使得 $I = \bigcup_S z^i$ 且 $z^i \cap z^j = \emptyset, \forall z^i, z^j \in S, z^i \neq z^j$。我们通过仅考虑中心位于区域 z^i 内的真实目标和检测结果来衡量特定区域 z^i 的检测性能。然后，对于任意评估指标 $m \in \mathcal{M}$，评估过程与传统方式保持一致，产生 n 个区域指标，每个指标表示为 m^i。我们称 $m^S = \{m^1, m^2, \cdots, m^n\}$ 为区域划分 S 的区域指标序列。

环形区域。 在实践中，中心化的摄影师偏差在视觉数据集中普遍存在 [25], [49], [59], [71], [77], [84]。如果目标是具有全方位综合检测能力的检测器，则评估区域可以设计成一系列环形区域：

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

其中 R_i 表示中心区域，由下式给出：

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

其中 $\text{Rectangle}(p, q)$ 表示左上角坐标为 p，右下角坐标为 q 的矩形区域。W 和 H 表示图像的宽度和高度，$r_i = \frac{i}{2n}, i \in \{0, 1, \cdots, n\}$ 控制矩形的大小。评估区域的图示可以在 图 3 中看到，其中 $n = 5$。我们将区域 $z^{i,j}$ 中的平均精度 (AP) 表示为 $\text{ZP}^{i,j}$。通过这种方式，传统评估是我们区域评估中的一个特例，因为可以很容易地得到 $\text{AP} = \text{ZP}^{0,n}$。

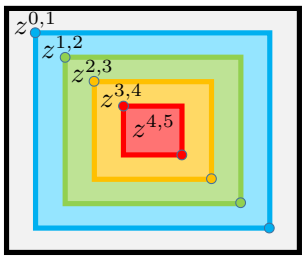
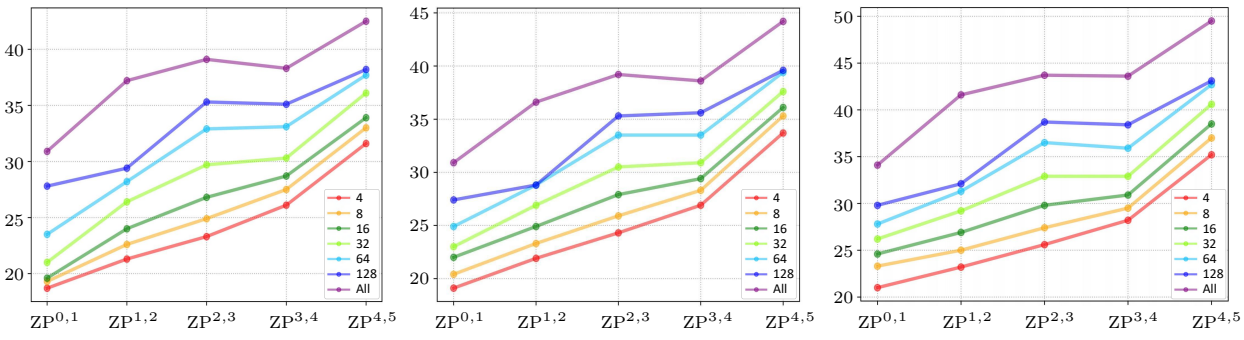


图 3. 当 $n = 5$ 时评估区域的定义。

其他区域划分。 传统评估可以灵活地为不同的应用场景选择不同的 IoU 阈值。对于那些需要精确定位框的应用，可以选择严格的 IoU 阈值，例如，$IoU = 0.75$ (AP_{75})。对于那些对框定位要求较低的应用，AP_{50} 就足够了。例如，旋转目标检测方法通常报告 AP_{50} [95], [96], [97]。类似于 AP 指标，用户可以根据自己的应用灵活地设计各种区域划分。如果用户关心全方位综合检测能力，环形区域划分将是一个不错的选择。如果用户关心某些感兴趣的区域，则可以自定义评估区域。在 节 6 中，我们将展示另外两种特殊区域划分的评估结果。一种是条形区域，即沿 x 轴的 5 个区域和沿 y 轴的 5 个区域（见 节 6.3 “观察 3”），另一种是 11×11 个块的正方形区域（见 节 6.4 “与目标分布的相关性”）。重要的是，由于区域划分保持一致，因此检测器之间的比较仍然是公平的。此属性帮助我们观察不同检测器在感兴趣区域的性能，以便我们可以根据实际应用需求选择检测器。在 节 6 中，我们将展示稀疏检测器在中心区域表现更好，而单阶段密集检测器在边界区域表现更好。

衡量区域指标的离散幅度。 由于检测性能在不同区域之间变化，我们进一步引入一个额外的指标来衡量区域指标之间的离散幅度。给定特定区域划分 S 的所有区域指标 m^S，我们计算区域指标的方差 $\sigma(m^S)$。理想情况下，如果 $\sigma(m^S) = 0$，则目标检测器的泛化能力达到完美的空间均衡 **在当前区域划分下**。在这种情况下，目标可以被很好地检测到，而不会受到其数据模式的影响。还值得一提的是，**空间偏差是目标检测器的一种外部表现**，ZP 方差只能反映给定区域划分的空间均衡性。换句话说，有以下三个概念。

1) 令 S 为区域划分，如果 $\sigma(m^S)$ 足够小，则检测



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

图 4. 具有各种目标尺度范围的平均 ZP。可以看出，对于每个目标尺度范围 r，三种目标检测器的空间偏差都很显著。

器的空间均衡性对于 S 来说是良好的。

2) 令 S_1, S_2 为两个具有 n 个区域的不同区域划分。如果 $\sigma(m^{S_1}) < \sigma(m^{S_2})$，则认为检测器在区域划分 S_1 的方式上更具有空间均衡性。

3) 检测器没有空间偏差，表明 $\forall S$ 为区域划分，$\sigma(m^S)$ 足够小。

4 深入研究空间偏差

在本节中，我们进行探索性实验，以阐明空间偏差的存在及其主要来源。我们采用了三种具有代表性的目标检测器。第一个是流行的单阶段密集目标检测器 GFocal [54]。第二个是经典的多阶段由密集到稀疏的目标检测器 Cascade R-CNN [9]。第三个是稀疏目标检测器 Deformable DETR [111]。

4.1 目标尺度研究

从评估区域的定义（公式 (1)）中，人们可能会问目标尺度是否在中心区域性能中起关键作用。在本实验中，采用环形区域划分，如 图 3 所示。如果目标的中心点坐标位于 $z^{i,j}$ 中，则该目标属于 $z^{i,j}$。为了消除目标尺度的影响，我们将区域评估过程限制在具有相似尺度的目标中。对于每个目标尺度范围 r，我们分别选择所有区域在 $[0, r^2]$、$[r^2, (2r)^2]$、...、$[((kr)^2, \infty]$ 范围内的真实框，其中尺度的最大端点设置为 $kr = 256$，并且 $r \in \{4, 8, 16, 32, 64, 128, \infty\}$。然后，我们计算所有尺度上 ZP 的平均值，如 图 4 所示。我们观察到，无论目标尺度范围选择得多小，空间偏差都非常显著。$\text{ZP}^{4,5}$ 分数仍然是最好的，相比之下，$\text{ZP}^{0,1}$ 分数是最差的。当评估区域更靠近图像边界时，性能下降幅度更大。可以看出，内部区域和外部区域之间的性能差距始终很大，超过 10 个 ZP 差距。这表明中心化的空间偏差可能不是源于目标尺度因素。为简单起见，我们在以下实验中对区域评估采用所有尺度。

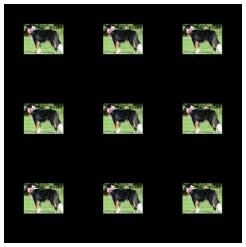
4.2 目标绝对空间位置研究

由于区域性能表现出明显的中心化趋势，因此一个直接的推测是空间偏差与目标的绝对空间位置有关。如果目标被移动到中心区域，则可能会被更好地检测到，反之，如果在边界区域，则会更差。为了分析目标的空间位置是否在其检测质量中起关键作用，我们构建了九宫格的数据集，用于检测在一张 600×600 纯黑色图像上被规则放置的目标。实验基于以下 3 个步骤：(1) 我们首先从 PASCAL VOC 2007 测试集 [25] 中裁剪目标，总共 14,976 个目标。(2) 所有目标都被缩放到固定大小，并以 3×3 网格方式放置。见 图 5(a)。(3) 为了衡量每个网格的检测质量，评估区域被定义为相同的 3×3 区域，表示为 z^{ij}，$i, j \in \{1, 2, 3\}$。从 图 5(b) 可以看出，检测器在中心区域 z^{22} 中表现并非最佳，甚至是最差的。这种现象与 [83] 中的先前观察结果有些不符，后者分析了平移 100 张图像的效果，并得出结论，当目标靠近图像边界时，检测器的性能可能会下降。然而，当我们将样本数量增加到超过 14K 时，我们观察到，将目标移动到中心区域在统计学上不会导致检测质量的提高。因此，检测性能的中心化趋势与目标的绝对位置之间的相关性不太明显。

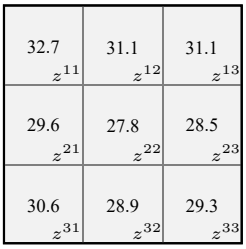
讨论： 图 4 的结论是中心目标比边界目标更容易被检测到，并且这与目标尺度无关，而 图 5 的结论是，目标的绝对位置（通过目标平移）与形成中心化区域性能无关。我们不禁要问：什么才是真正决定目标检测器空间偏差的因素？

4.3 区域之间的目标数据模式

本小节旨在研究目标检测器中心化空间偏差的来源，其背后的灵感是，中心化的空间偏差可能来自于区域之间目标数据模式的差异。如果一个目标是从中心区域风格的数据分布中采样的，则可以更好地检测到该目标，而



(a) 九宫格数据集。



(b) 9 个区域的 ZP。

图 5. (a) 九宫格数据集是通过将测试集的所有目标规则地放置在 600×600 的黑色图像上而构建的。(b) GFocal 的区域评估 (3×3 网格)。

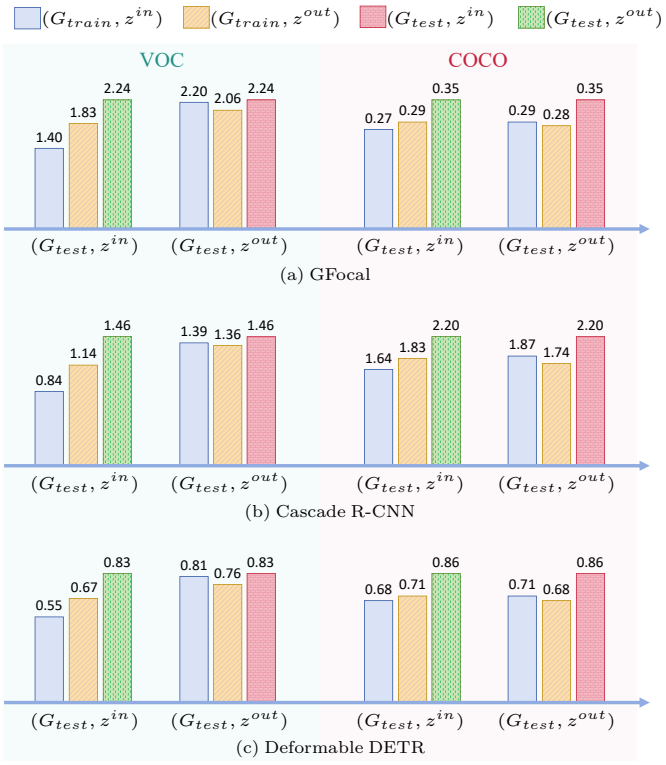


图 6. 特征表示中心之间的平均误差 $\mathcal{E}((G_1, u), (G_2, v))$。例如，蓝色条表示 $\mathcal{E}((G_{test}, z^{in}), (G_{train}, z^{in}))$。从同一区域采样的目标比从不同区域采样的目标具有显著更低的差异。

如果它是从边界区域风格的数据分布中采样的，则检测会更差。本文对目标数据模式的表示非常通用 [6], [16], [102]，并且只需要对检测流程进行少量适应的修改，仅需要 1)输入图像 I,2)所有真实框 G,3)来自预训练目标检测器的特征提取器 $f : I \to \mathbb{R}_{M \times H \times W}$. 在推理过程中，本文使用真实框从 $f(I)$ 中裁剪目标特征，然后沿空间维度对特征值取平均。每个目标都由一个 M 维特征向量 g 表示，该向量编码了高维空间中的目标数据模式。在本实验中，区域数设置为 2。本文将中心区域表示为 z^{in}，它是一个矩形区域，左上角坐标为 $p = (0.25W, 0.25H)$，右下角坐标为 $q = (0.75W, 0.75H)$，其中 W 和 H 是输入图像的宽和高。除此之外的其余部分设置为边界区域，表示为 z^{out}。区域之间目标数据模式的差异表示为：

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

其中 $\bar{g}$ 表示特征表示中心，$u, v \in \{z^{in}, z^{out}\}$ 表示从中心区域或边界区域采样的目标，以及 $G_1, G_2 \in \{G_{train}, G_{test}\}$ 表示从训练集或测试集采样的目标。误差针对每个类别分别计算，然后取平均值。C 是类别的总数。此外本文还引入一个指示函数 $\mathbb{1}_k$，用以消除目标尺度的影响。当目标尺度在范围 $R = \{[((k-1)r)^2, (kr)^2]\} \bigcup \{[((K-1)r)^2, \infty]\}, k \in \{0, 1, \cdots, K-1\}$ 的其中之一时，$\mathbb{1}_k$ 为 1，否则为 0。简而言之，$\mathcal{E}$ 测量四个集合的特征表示中心之间的距离，即来自训练集的中心区域目标、训练集的边界区域目标、测试集的中心区域目标、测试集的边界区域目标。

结果在 图 6 中报告。对于 VOC，训练集是 VOC 2007 trainval，测试集是 VOC 2007 test。对于 COCO，训练集是 COCO train2017，测试集是 COCO val2017。可以看出，从同一区域采样的目标比从不同区域采样的目标具有显著更低的差异。具体而言，测试集的中心目标与训练集的中心目标更相似，而测试集的边界目标与训练集的边界目标更相似。这表明目标数据模式在不同区域之间实际上是不同的，并且网络能够捕捉到这种偏差。如 图 7(a) 所示，区域性能在 VOC 2007 trainval 集上是中心化的，因此它自然地在测试集上继承了相同的趋势。更耐人寻味的是，我们进一步可视化了 图 7(b-f) 中九宫格数据集上的检测性能，其中中心目标和边界目标是分开的。可以看出，无论我们将中心目标放置在哪里，检测器始终可以在检测中心目标方面表现更好。我们注意到，这种现象适用于所有 5 个数据集，包括 PASCAL VOC、MS COCO 和其他 3 个应用数据集（口罩、水果、头盔）。

以上结果证实了我们的直觉，即如果目标是从中心区域风格的数据分布中采样的，则可以更好地检测到该目标，而如果目标是从边界区域风格的数据分布中采样的，则会更差。这表明，当我们人类拍照时，区域之间目标数据模式总是存在差异，尽管这种差异是难以察觉的。当镜头聚焦于目标最有可能出现的感兴趣区域时，不可避免地导致边界区域中目标的采样频率降低，从而导致次优性能。

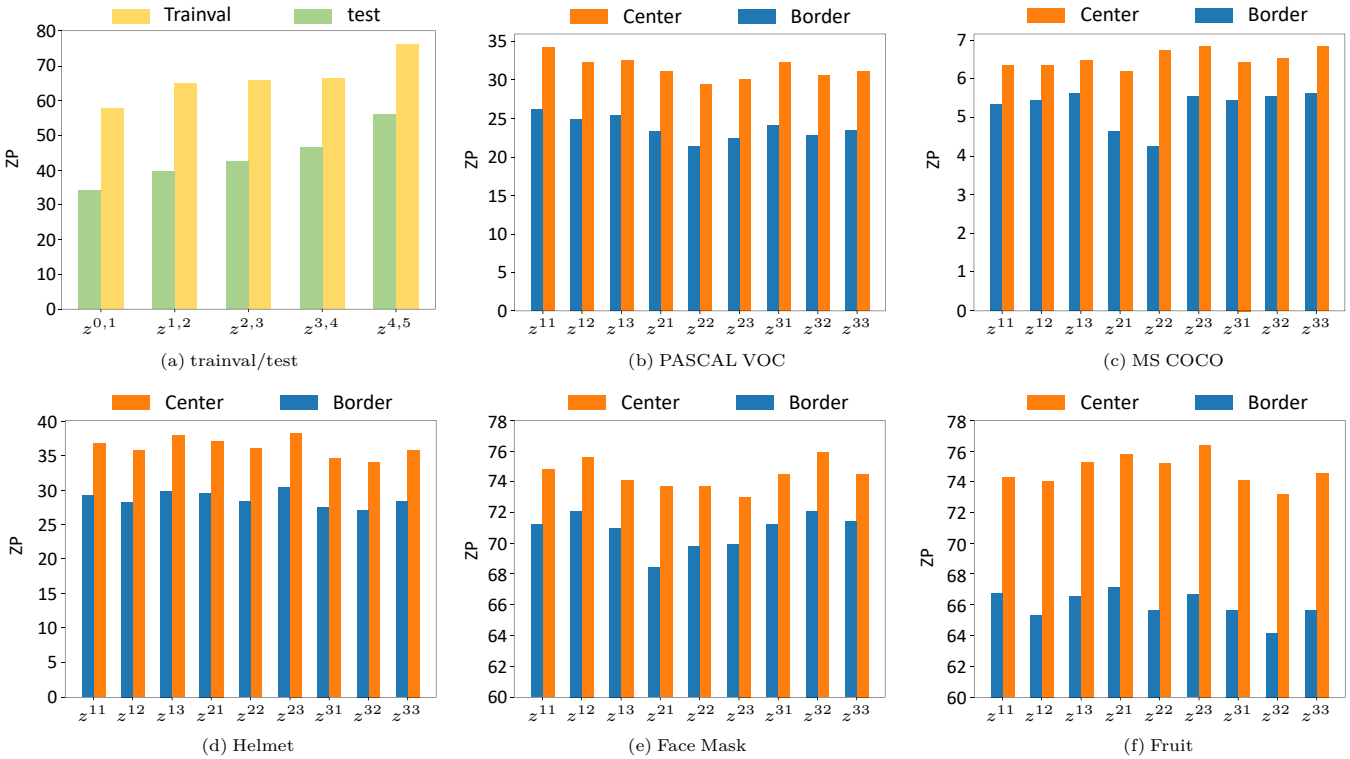


图 7. (a) VOC 2007 trainval 集和 test 集上的 5 区域评估。(b-f) 在九宫格数据集上，分别对中心目标和边界目标进行区域评估。结果表明，无论我们将中心目标放置在哪里，检测器始终可以在检测中心目标方面表现更好。

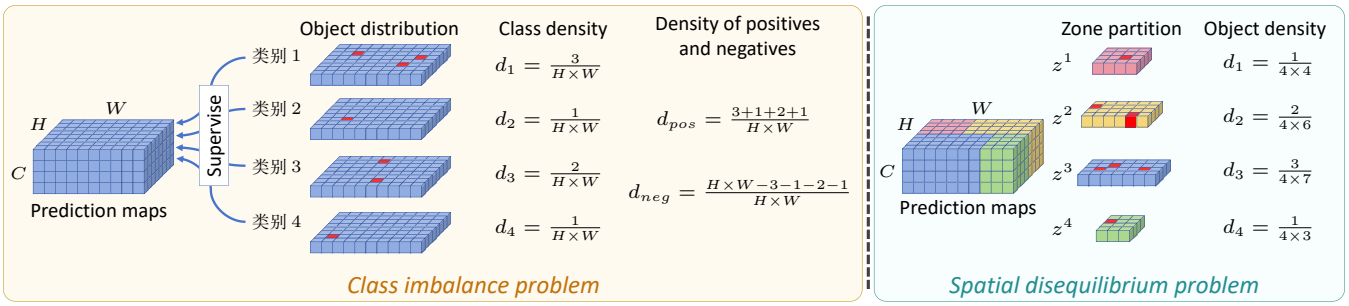


图 8. 类别不平衡问题和空间不均衡问题之间关系的说明。有 7 个 4 个类别的目标，用红色立方体表示。评估区域设置为 4 个区域，颜色分别为粉色、黄色、蓝色和绿色。我们简化了讨论，即每个目标仅包含 1 个正样本，多个正样本的情况类似。在左图中，类别密度是每个类别的目标数量与预测图大小的比率，而在右图中，目标密度是每个区域的区域目标数量与区域大小的比率。空间不均衡问题在形式上等同于类别不平衡问题。

5 空间不均衡问题

至此，我们已经展示了空间偏差的存在和主要来源。边界区域的次优性能阻碍了检测应用的鲁棒性。在本节中，我们为目标检测引入新的空间不均衡问题。

5.1 问题定义

将 S 表示为区域划分，m^S 表示一系列区域指标，$\sigma : \mathbb{R}^n \to \mathbb{R}$ 表示方差计算，空间不均衡问题被定义为最小化区域指标的方差：

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

其中 Θ 是检测器的网络参数集。促进空间均衡的总体目标主要取决于使用哪种区域划分，这取决于应用。因此，对于不同的应用场景，可以自定义区域划分。

讨论： 空间不均衡问题在形式上等同于类别不平衡问题。我们将区域 z^i 中的目标表示为 X^i_{obj}。给定区域 z^i 的目标密度可以表示为 $d_i = |X^i_{obj}|/|z^i|$。直观上，较高的密度表示更多的正样本，从而在区域上产生更大的梯度流。这类似于类别不平衡问题，后者在类别之间具有长尾分布，如 图 8 所示。具体而言，分类分支预测类别分数，这是一个 $C \times H \times W$ 的张量。第 c 个类别的密度

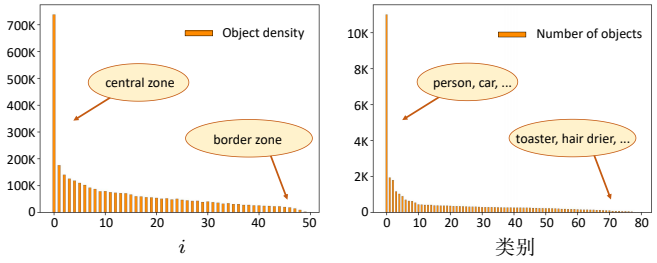


图 9. 左图：COCO val2017 上 50 个区域的目标密度。这些区域被中心定义为 $z^{i,i+1}$，$i = 0, 1, \cdots, 49$。**右图**：COCO val2017 上类别的长尾分布。空间不均衡问题与类别不平衡问题具有相似的特征。

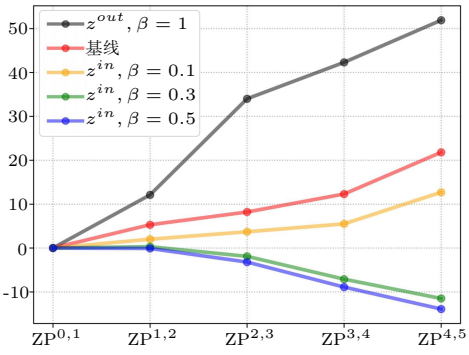


图 10. ZP 相对于 $ZP^{0,1}$。如果边界区域的监督信号强度降低，则区域性能可以进一步极端中心化；反之，如果中心区域的监督信号强度降低，则区域性能可以是反中心化的。

表示为比率 $d_c = |X_c|/(H \times W)$，X_c 是第 c 个类别的目标。由于 $H \times W$ 是一个常数，因此等效于将一个类别的所有样本放置在面积相同的区域中。然后，每个类别都有一个用于模型学习的 $H \times W$ 的单个区域，并且类别之间是不相交的。在 图 9 中可以看出，这两个问题都服从长尾分布。由于这个事实，区域性能也可能与区域中的监督信号强度相关。一种简单的方法是扩大或缩小区域的监督信号强度，以使网络达到新的收敛状态。这里，我们将一个新的参数 β 插入到标签分配算法 [104] 中。如果锚点的 IoU 大于 $\alpha_{pos} + \beta * \mathbb{1}_z$，则该锚点被分配为正样本，其中 α_{pos} 是正 IoU 阈值，$\mathbb{1}_z$ 如果此锚点的中心点位于区域 z 中则为 1，否则为 0。因此，当 β 增加时，正样本的数量 $|X^i_{pos}|$ 会减少。

我们在 图 10 中可视化了相对 ZP，其中所有 ZP 都减去了 $ZP^{0,1}$。可以看出，如果我们通过减少边界区域中正样本的数量来削弱监督信号，则中心化空间偏差会进一步加剧。相反，如果我们在中心区域削弱监督信号，我们甚至可以实现反中心化的空间偏差。这表明监督信号强度确实对区域性能有影响。鉴于以上分析，我们最终讨论了一种可能的解决方案，用于解决环形区域划分下的空间不均衡问题。

5.2 空间均衡学习

大多数现有的目标检测研究都侧重于追求图像级别的更高检测性能，而忽略了区域级别的优化，从而导致检测器出现严重的空间不均衡问题。在本小节中，我们介绍一种可能的解决方案，称为空间均衡学习，作为缓解空间不均衡问题的开始。我们首先介绍空间权重，它通过下式将锚点坐标 (x^a, y^a) 映射到标量 $\alpha(x^a, y^a)$：

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

其中 W 和 H 是图像的宽度和高度。空间权重可以很容易地插入到现有的检测流程中，只需进行少量修改。原理简单且具有多种选择。在这里，我们提供以下两种实现方式。

1) 空间均衡标签分配 (SELA)。 在这种方法中，关键思想是在制定标签分配的判据规则时，将空间权重视为一个额外的约束项。由于大多数标签分配算法都有其自身复杂的实现方式，因此在下文中，我们提供了经典 ATSS [104] 的具体应用描述，仅仅是因为它的简洁性。给定正 IoU 阈值 t，该阈值是通过考虑目标的统计特征来计算的。ATSS 准则遵循与 max-IoU 分配 [58], [74], [75] 相同的规则，即 $\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t$，其中 $\boldsymbol{B}^a$ 和 $\boldsymbol{B}^{gt}$ 分别表示预设的锚框和真实框。SELA 过程表示为：

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

其中 $\gamma \geqslant 0$ 是一个超参数。可以看出，SELA 放宽了图像边界附近目标的正样本选择条件。因此，将选择更多的锚点作为它们的正样本。请注意，上述应用实际上是一种基于频率的方法，就像为长尾类别不平衡问题提出的许多类别重平衡采样策略一样 [42], [67]。

2) 空间均衡损失 (SE 损失)。 在这种方法中，我们采用代价敏感学习方法。我们将空间权重项 $1+\gamma\alpha(x^a, y^a)$ 作为分类和边界框回归损失的附加权重因子。这样，将在边界区域产生更大的梯度流，从而使网络更加关注边界目标。

未来方向： 有更多潜在的、有希望的解决方案可以实现空间均衡，值得未来研究。例如，设计适当的数据增强 [19], [46], [90]，更具体地说，增加数据增强以弥补图像边界附近目标的采样频率不足，可能是一种有希望的解决方案。此外，由于空间不均衡问题在形式上等同于类别不平衡问题，因此一些改进的重平衡方法也可能为空间均衡学习带来收益，例如，类别平衡损失 [21]、迁移学习 [20], [89] 和表示学习 [23], [36], [64] 等。在寻求

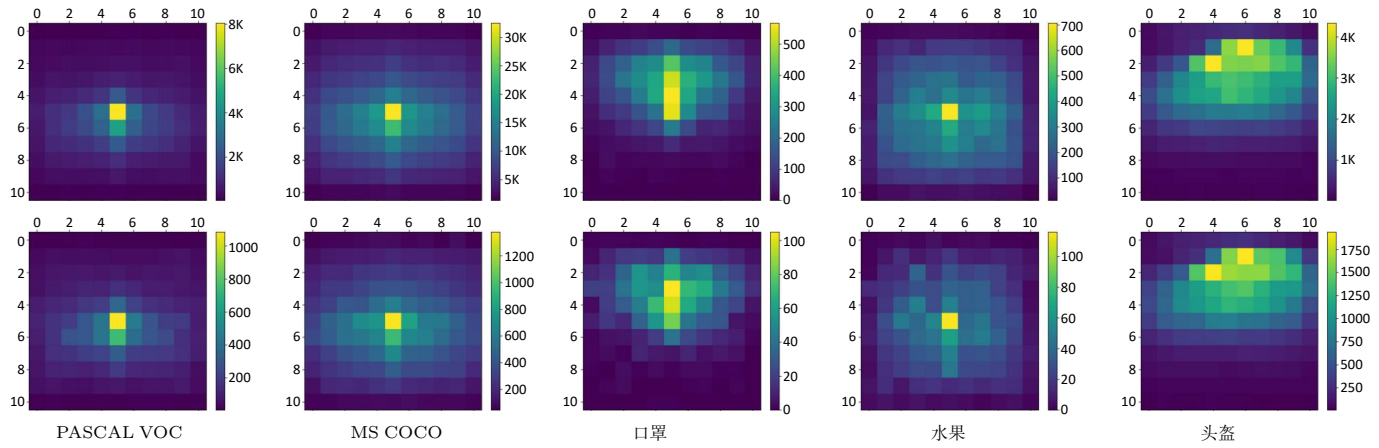


图 11. 5 个目标检测数据集中摄影师的偏差。我们统计所有真实框的中心点。图像被划分为 11 × 11 个区域。第一行：训练集。第二行：测试集。

更多以统一形式解决这两个问题的解决方案方面，也存在一个非常令人兴奋的未来工作领域。此外，我们的方法主要考虑环形区域划分，其旨在平衡中心区域和边界区域之间的检测能力。对于某些更关心某些感兴趣区域的特殊应用，设计理念可以是通用的，并且取决于实际应用。

6 定量评估

在本节中，我们对 10 个流行的目标检测器和 5 个目标检测数据集进行全面的评估。

6.1 实验设置

检测器和指标。 我们评估的所有目标检测器都可以从 MMDetection [13] 或其官方网站下载。我们遵循标准的平均精度评估协议。为了全面评估检测器，报告了各种指标，包括 5 个 ZP、5 个 ZP 的方差以及传统指标 AP。

数据集。 我们使用的所有数据集都是公开可用的，可以从其官方网站或 Kaggle 下载。5 个数据集的目标分布可以在 图 11 中看到。

PASCAL VOC [25] 是最广泛使用的自然场景下目标检测基准之一，包含 20 个类别。我们采用经典的 07+12 训练和测试协议，即，训练集包含 VOC 2007 trainval 和 VOC 2012 trainval 的联合（总共 16551 张图像），测试集包含 VOC 2007 test（4952 张图像）。

MS COCO [59] 是另一个最近流行的基准，规模更大，包含自然场景下的 80 个类别。我们采用 COCO 2017 train（118K 张图像）进行训练，COCO 2017 val（5K 张图像）进行评估。

人脸口罩检测 *[1]*。随着 COVID-19 在世界各地肆虐，人脸口罩检测是一项广泛且必要的视觉应用。该数据集由 5,865 张训练图像和 1,035 张测试图像组成。共有 2 个类别。一个是人脸，另一个是口罩。

水果检测 *[24]* 广泛应用于工业装配线分拣和商品分类。该数据集由 3,836 张训练图像和 639 张测试图像组成。包括 11 种常见水果，例如，苹果、葡萄和柠檬等。

头盔检测 *[2]* 是一种安全视觉应用，常用于建筑工地，以检测工人和访客是否佩戴头盔。它包含 15,887 张训练图像和 6,902 张测试图像。使用头和头盔两个类别。

空间均衡学习的设置。 对于空间均衡学习评估，该实现基于 MMDetection [13] 框架，并且消融研究在 GFocal [54] 上进行，使用 ResNet [34] 主干网络和 FPN [57] 颈部网络。我们对 VOC 07+12 和 3 个应用程序数据集使用 ResNet-18，对 MS COCO 采用 ResNet-50。根据 GPU 的数量，学习率通过线性缩放规则 [30] 进行线性缩放。所有实验的训练 epoch 都设置为 12。我们在 公式 (6) 中设置 $\gamma = 0.2$，并且为了公平比较，所有其他超参数保持不变。

6.2 对各种目标检测器进行区域评估

尽管传统评估为检测器的整体性能提供了良好的指导，但对于检测器的空间偏差以及位置和程度知之甚少。在这里，我们选择了各种目标检测器，它们具有不同的检测流程，但具有相同水平的传统指标。它们是流行的、具有代表性的，并且被认为是现代目标检测的里程碑：单阶段密集检测器 (RetinaNet [58]、GFocal [54]、VFNet [101]、YOLOv5 [41])、多阶段由密集到稀疏的检测器 (R-CNN 系列 [9], [33], [75]) 和稀疏检测器 (DETR 系列 [11], [111] 和 Sparse R-CNN [81])。定量结果在表 1 中报告。

有几个有趣的观察结果：

表 1
对现有流行目标检测器进行区域评估。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。结果在 COCO 2017 val 上报告。"Cas.": Cascade 的缩写. **R**: ResNet [34]. **X**: ResNeXt-32x4d [92]. **PVT-s**: Pyramid vision transformer-small [88]. **CNeXt-T**: ConvNext-T [63].

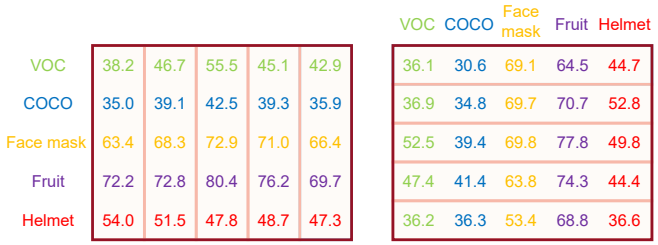
检测器	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
DETR (**R**-50) [11]	40.1	26.9	29.8	36.2	39.8	39.1	45.7
RetinaNet (**PVT-s**) [88]	40.4	19.7	30.8	36.9	39.0	37.4	44.6
Cascade R-CNN (**R**-50) [9]	40.3	18.7	30.9	36.6	39.2	38.6	44.2
GFocal (**R**-50) [54]	40.1	16.9	31.1	37.5	39.4	38.5	43.8
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

表 2
VOC 07+12、COCO 2017 和 3 个应用数据集（即，人脸口罩检测、水果检测和头盔检测）的区域评估。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。结果在 GFocal [54] 上报告。

数据集	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
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
人脸口罩	71.3	13.1	60.4	67.1	69.0	68.8	70.9
水果	76.6	56.2	60.8	69.9	71.2	75.3	83.8
头盔	49.7	3.0	45.9	47.9	50.3	50.6	47.8

1) 空间偏差非常普遍。可以看出，所有检测器都显示出明显的中心化区域性能，即在中心区域 $(z^{3,4}, z^{4,5})$ 表现良好，但在边界区域 $(z^{0,1}, z^{1,2})$ 表现不佳。这证实了空间偏差的存在和普遍性，并且我们首次成功地量化了 图 2 中所示的目标检测器的缺陷。

2) 它们的空间均衡性差异很大。 ZP 方差存在 12.9 到 26.9 的巨大差距。特别是，我们发现稀疏检测器，例如 DETR 系列和 Sparse R-CNN，倾向于产生较大的 ZP 方差，而单阶段密集目标检测器在空间均衡性方面表现更好 (较低的 ZP 方差)。这表明基于 DETR 的检测器在空间均衡性方面与基于 CNN 的检测器不一致。我们推测这可能归因于自注意力机制捕获的全局信息。稀疏检测器首先通过 CNN 提取特征，然后通过一系列注意力模块处理特征。训练更具动态性，我们假设中心目标可能会受到更多关注。显然，可以得出结论，必须有一些因素导致检测器之间的空间均衡性不同，包括但不限于神经网络架构设计、优化和训练策略。然而，目前我们尚不清楚哪些组件或算法设计对空间偏差有影响。我们相信，未来对该主题的进一步研究将很有趣，并且该研究很有可能找到解决空间不均衡问题的关键。



(a) 沿 x 轴的 5 个区域 (b) 沿 y 轴的 5 个区域

图 12. 区域划分的两种设计。报告了 ZP。

3) 传统评估未能捕捉到空间偏差。可以看出，GFocal (R-50) 和 DETR (R-50) 实现了相同的 AP 分数 40.1。然而，传统指标没有提供关于某个区域的检测性能的信息。我们的区域评估表明，GFocal 在边界区域 $z^{0,1}$ 和 $z^{1,2}$ 中表现更好，而 DETR 在区域 $z^{2,3}$、$z^{3,4}$ 和 $z^{4,5}$ 中表现更好。类似地，Deformable DETR (**R**-50) [111] 实现了与 GFocal (**X**-101) 相同的传统 AP。区域评估表明，Deformable DETR 在中心区域 $z^{3,4}$、$z^{4,5}$ 的表现明显优于 GFocal，而在边界区域 $z^{0,1}$、$z^{1,2}$ 中表现较差。这些性能差异被传统评估所掩盖。此外，有趣的是，AP 指标正好介于 $ZP^{3,4}$ 和 $ZP^{4,5}$ 之间，这表明在 96% 的图像区域中的检测性能实际上低于 AP。

启示：以上结果揭示了检测器的性能特征，这有助于我们更好地理解目标检测器的行为，并鼓励我们在部署到应用场景时重新考虑检测器的选择。此外，还值得研究检测流程中的哪些组件导致了这些性能差异，例如，自注意力机制、标签分配等。

6.3 对各种数据集进行区域评估

表 2 报告了 PASCAL VOC 07+12、MS COCO val2017 和 3 个应用数据集的定量检测结果。我们有以下观察结果：

1) 可以看出，检测性能在不同区域之间变化。最靠近图像边界的区域，即 $z^{0,1}$，始终具有最低的检测性能。相比之下，中心区域 $z^{4,5}$ 在几乎所有这些情况下都具有最高的性能。

2) 还有一个代表性的例子，即头盔数据集，其 ZP 方差仅为 3.0。这表明头盔数据集在环形区域划分的情况下实现了最佳的空间均衡性，而其他数据集则存在明显的空间不均衡问题。例如，在 PASCAL VOC 上的 ZP 方差为 53.6，在水果数据集上为 56.2。

3) 如果我们切换到其他区域划分，例如，沿 x 轴的 5 个条形区域和沿 y 轴的 5 个条形区域（见 图 12(a)(b)），它们的空间均衡性会发生变化。在表 3 中，在沿 y 轴的 5 个区域的情况下，人脸口罩和头盔数据集的 ZP 方差

表 3
三种区域划分类型的 ZP 方差。空间偏差是检测器的外部表现，而空间均衡性对应于给定的区域划分。

区域划分	VOC	COCO	人脸口罩	水果	头盔
5 个环形区域	53.6	16.9	13.1	56.2	3.0
沿 x 轴的 5 个区域	32.3	7.2	11.2	13.7	6.4
沿 y 轴的 5 个区域	46.7	14.0	39.6	20.8	30.5

表 4
SELA 中超参数 γ 的评估。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。$\gamma = 0$ 表示基线 GFocal。方差越低，空间均衡性越好。(数据集：VOC 07+12)

γ	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
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

分别增加到 39.6 和 30.5，而在这两种情况下，水果数据集的 ZP 方差都显着降低。

启示：区域评估提供了一个新的视角，揭示了目标检测器的局限性。可以看出，空间偏差也是目标检测器的自然特征，并且对于任意区域划分，它们很难实现完美的空间均衡。以上结果表明，ZP 方差是与区域划分相关的集合函数。环形区域划分主要考虑内部区域和外部区域之间检测能力的平衡，这在实践中是一个不错的选择，因为中心化摄影师偏差在视觉数据集中普遍存在。然而，应该注意的是，区域划分是灵活的，并且能够根据应用场景定制为任何形状。

6.4 空间均衡学习评估

最后，我们提供空间均衡学习的评估。消融研究是使用 GFocal 进行的，并且我们默认采用第一种方法，即空间均衡标签分配 (SELA)。

超参数 γ。 回想一下，SELA 的实现仅涉及 公式 (6) 中的一个超参数 γ。γ 控制空间权重的大小。较大的 γ 会为图像边界附近的目标增加更多的正样本。如表 4 所示，我们观察到我们的 SELA 可以为 γ 的所有选项实现持续的空间均衡改进（较低的方差）。过大的 γ，例如 0.4，将为所有区域增加更多的正样本，导致性能下降。因此，我们将 PASCAL VOC 的 γ 设置为 0.2。可以看出，我们的 SELA 可以显着提高外部区域（例如，$ZP^{0,1}$、$ZP^{1,2}$、$ZP^{2,3}$ 和 $ZP^{3,4}$）的检测性能。如 图 14(a) 所示，虽然中心区域 $z^{4,5}$ 的性能略有下降，但 ZP 的改善在边界区域非常显着，边界区域占据了总图像区域的

表 5
空间权重的分析。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。$\gamma = 0.2$。方差越低，空间均衡性越好。(数据集：VOC 07+12)

权重	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
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

表 6
PASCAL VOC 07+12、MS COCO 2017 和 3 个应用数据集（包括人脸口罩检测、水果检测和头盔检测）的区域评估。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。检测器为 GFocal [54]。

数据集	SELA	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
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
人脸口罩		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
水果		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
头盔		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

96 这对于监控系统和自动驾驶汽车中的安全应用尤为重要，因为目标可能出现在任何地方。边界区域的性能在鲁棒性检测中起着重要作用。在实践中，我们将所有其他数据集的 γ 设置为 0.1，但应注意，对于不同的应用场景，可能存在更好的 γ。

空间权重。 人们可能想知道，如果我们直接放宽正样本的选择条件而不考虑其空间位置，性能会如何变化。在这里，我们进行实验以研究空间权重的影响。定量结果在表 5 中报告。如果空间权重设置为常数 1，则意味着我们直接将正 IoU 阈值 t 降低为 $\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t - \gamma$，并且将选择更多正样本而没有空间区分。可以看出，尽管性能有所提高，但 5 个 ZP 的方差很大。这表明从正 IoU 阈值中减去一个常数不能显着改变采样频率，因为在中心区域会生成更多正样本。相比之下，我们的 SELA 可以显着降低方差，并实现更好的空间均衡性。

各种数据集上的 SELA。 表 6 向我们展示了有希望的结果，我们的 SELA 可以为目标检测实现更好的空间均衡性。特别是，我们在 ZP 方面大幅降低了方差。例如，在 PASCAL VOC、MS COCO 和人脸口罩/水果检测方面，我们成功地将 ZP 的方差降低了 -15.9、-2.5、-1.0 和 -22.6。这表明我们的 SELA 可以提高多种应用场景的空间均衡性，而不会牺牲 AP。

空间均衡学习的通用性。 我们进一步提供了更多实

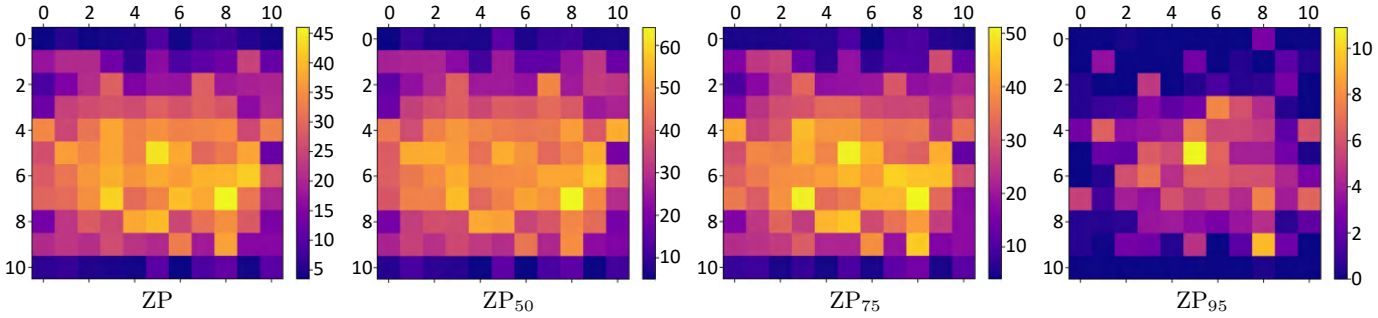
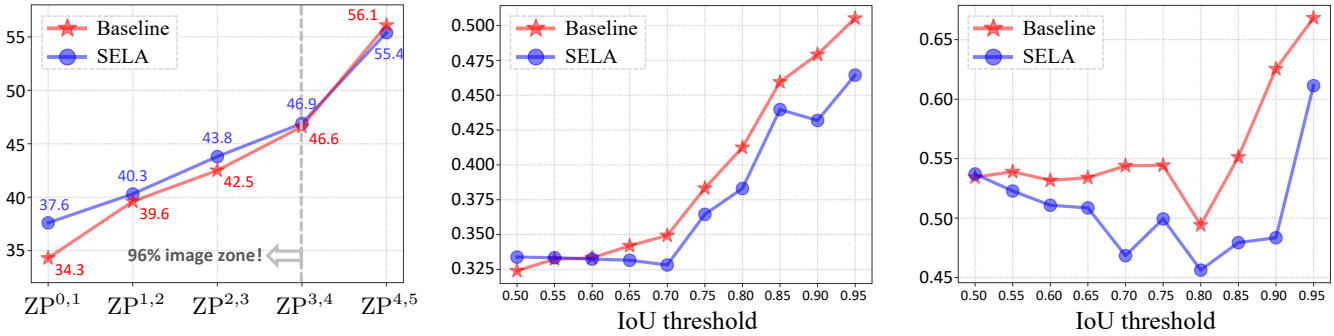


图 13. 在 11×11 个正方形区域上进行区域评估。模型是 GFocal。结果在 VOC 07+12 上报告。



(a) SELA 的效果 (b) PCC (c) SCC

图 14. (a) ZP 与区域的关系。(b) mZP 与目标分布（中心计数）之间针对 IoU 阈值的 Pearson 相关系数 (PCC)。(c) mZP 与目标分布之间针对 IoU 阈值的 Spearman 相关系数 (SCC)。我们的 SELA 可以在大多数 IoU 阈值下大幅降低这些相关性，表明空间均衡性更好。基线模型为 GFocal。结果在 VOC 07+12 上报告。

表 7
具有各种主干网络的 SELA 评估。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。**X**: ResNeXt [92]。(数据集：VOC 07+12)

模型	SELA	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
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

表 8
空间均衡学习的通用性。采用基于代价敏感学习的方法，SE 损失。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。(数据集：VOC 07+12)

检测器	SE 损失	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
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

验，以验证空间均衡学习在各种主干网络上的有效性。表 7 表明，我们的 SELA 可以显着提高所有 3 个主干网络（即，较低方差）的空间均衡性。我们还进行了实验，通过将空间均衡学习结合到 3 个更多的检测器 DW [53]、DDOD [15] 和类 DETR 检测器 DINO [100] 中，来检验空间均衡学习的通用性。在这里，我们采用空间均衡损失（SE 损失），并且我们扩大了图像边界附近目标的训练损失。表 8 报告了这 4 个目标检测器的 SE 损失的定量结果。如图所示，我们的方法可以显着降低 4 个检测器的 ZP 方差，表明实现了更好的空间均衡性。这表明我们的方法在无需任何花里胡哨的情况下即可提高检测器空间鲁棒性的通用能力。

我们还注意到，与基于 CNN 的目标检测器相比，我们的方法在 DINO 上产生了轻微的空间均衡性改进。这可能归因于类 DETR 检测器和其他检测器之间不同的优化过程。类 DETR 检测器中的正样本数量非常有限，因为它们使用一对一的匈牙利匹配，而密集目标检测器中则丰富得多，因为它们采用一对多的分配。因此，我们的 SE 损失更有助于缓解密集目标检测器上监督信号强度不平衡的问题。这意味着改善类 DETR 检测器的空间均衡性可能更具挑战性。我们希望这项工作可以启发更多解决方案，以解决未来类 DETR 检测器的不均衡问题。

与目标分布的相关性。 我们进一步提供了区域指标与目标分布之间的相关性。我们定义了更精细的区域划

分，这与用于计数目标中心的区域划分相同，即 11×11 个正方形区域（见 图 11）。然后，我们逐个评估 121 个区域的检测性能。我们在图. 13 中绘制了 121 个区域的 ZP。可以看出，ZP 分布与目标分布（图. 11）相似，即相同的中心化趋势。为了研究区域指标与目标分布之间的相关性，我们进一步计算了 mZP 与测试集目标分布之间的 Pearson 相关系数 (PCC) 和 Spearman 相关系数 (SCC)。如图. 14(b) 和图. 14(c) 所示，我们对空间偏差有了以下深刻的反思。我们首先注意到，图 14(b) 中的所有 PCC > 0.3，这表明检测性能与目标分布呈中等线性相关。提醒一下，PCC 仅反映两个给定向量的线性相关性，而当它们呈曲线相关时，PCC 可能会失效。在 图 14(c) 中，Spearman 相关性反映了 mZP 和目标分布之间更高的排序相关性，所有 SCC > 0.45。这说明检测性能与目标分布具有中等到高度的相关性。总的来说，我们的 SELA 大大降低了这些相关性，表明与目标分布的相关性较低，空间均衡性更好。

检测可视化。 我们在图. 15 中可视化了 SELA 的检测结果。我们的方法可以提高边界区域的检测性能。我们认为，进一步探索空间均衡性对于鲁棒的检测应用显然是值得且重要的。

6.5 其他实现空间均衡的尝试

正如我们在 节 4 中讨论的那样，我们的结果表明，目标尺度和目标的绝对位置几乎不影响空间偏差。区域之间目标数据模式的差异在空间偏差中起着重要作用。在本小节中，我们研究更多可能的因素，看看空间均衡性如何变化。第一个是填充操作。我们遵循 [45] 的工作，将填充方式设置为 full-conv，正如 [45] 所证明的那样，它是平移不变的。我们用 full-conv 替换了头部网络的所有卷积核。第二个是图像边界附近超大尺寸锚框的影响。基线模型的默认设置保留了所有锚框。我们移除了所有边缘超出有效图像范围的锚框。第三个是图像分辨率。基线模型的默认分辨率为 1333×800，我们训练了一个分辨率较小的模型，例如 640×640。结果在表 9 中报告。可以看出，这三个修改都导致了 AP 的显着下降。full-conv 填充可以降低 ZP 方差，但对检测精度没有帮助。此外，通过移除越界锚框或设置不同的图像分辨率无法获得更好的空间均衡性，因为区域之间的监督信号仍然不平衡。找到一种既能缓解空间不均衡问题又不会导致性能下降的解决方案具有挑战性。

表 9
空间偏差的 3 个潜在因素的评估。(1) 我们在头部网络中使用 full-conv [45] 填充；(2) 我们移除了超出有效图像边界的超大锚框；(3) 我们将图像分辨率设置为 640×640。报告了 5 个区域精度 (ZP)、ZP 的方差和传统指标 AP。结果在 COCO val2017 上的 GFocal [54] 上报告。

修改	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
基线	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 结论、挑战与展望

在本文中，我们提出了区域评估，以揭示现代目标检测器中空间偏差的存在和离散幅度。我们发现，空间偏差与目标尺度和目标的绝对位置的相关性较小，而与区域之间目标数据模式的差距密切相关。基于对空间偏差起源的深入研究，我们最终提出了空间不均衡问题，旨在实现跨区域的鲁棒检测。作为缓解此问题的开始，我们还展示了一条通往空间均衡目标检测的路径。广泛的实验证明了空间偏差的存在和主要来源，这在各种现代检测器和数据集中普遍存在。

意义。空间偏差是目标检测中的一个天然障碍，检测器通常在边界区域表现出性能下降，而边界区域占据了图像区域的很大一部分。虽然经典的 AP 指标仍然被认为是主要的衡量标准，但它很难揭示空间偏差，并且难以全面反映目标检测器的真实性能。最大化 AP 指标并不能完全表明鲁棒检测，并且在所有区域都表现良好。区域评估补充了一系列区域指标，弥补了传统评估的缺点，并捕获了更多关于检测性能的信息。我们希望这项工作能够启发社区重新思考目标检测器的评估，并激发对空间偏差以及空间不均衡问题解决方案的进一步探索。

这项工作留下了一些挑战：

*各种目标检测器中空间偏差的可解释性。*本文主要揭示了目标检测器中空间偏差的存在和离散幅度，而不同检测器表现差异很大的具体原因仍然扑朔迷离。神经网络架构设计、预训练数据、优化、训练策略，甚至超参数都可能在空间偏差中发挥作用。进一步探索以回答上述问题至关重要。

*其他潜在因素对空间偏差的影响。*目前，我们指出了不平衡的目标分布与区域性能之间存在明显的关联。还有一些复杂而隐含的因素，例如图像模糊、目标遮挡、边界效应、噪声等，也可能导致空间偏差。然而，当前检测数据集几乎缺乏对上述因素的注释，这使得难以建立定量分析。

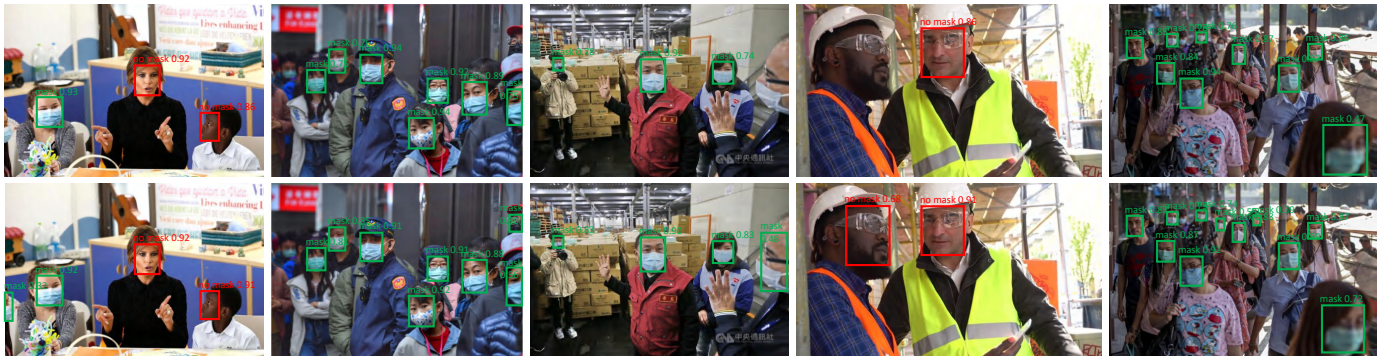


图 15. GFocal（第一行）和 GFocal + SELA（第二行）的检测结果图示。我们的方法提升了边界区域的检测性能。放大以获得更好的视图。

其他视觉任务的区域评估。研究人员发现了一些线索，表明图像生成器可能会在图像边界附近生成失真内容 [18]。因此，空间偏差也可能存在于许多视觉任务中。我们的区域评估可能具有巨大的潜力来揭示空间偏差，无论是对于高级还是低级视觉任务。

参考文献

[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.

郑兆晖 于 2021 年在天津大学获得计算数学硕士学位。他目前是南开大学媒体计算实验室的博士研究生，导师是程明明教授。他的研究兴趣包括目标检测、实例分割和知识蒸馏。他于 2023 年获得 CCF-CV Academic Emerging Scholar Award 。

陈宇铭 于 2022 年在兰州大学获得计算机科学学士学位。他目前是南开大学媒体计算实验室的硕士研究生，导师是程明明教授和侯淇彬教授。他的研究兴趣包括目标检测和知识蒸馏。

侯淇彬 于南开大学计算机学院获得博士学位。之后，他在新加坡国立大学担任研究员。现在，他是南开大学计算机学院的副教授。他在顶级会议/期刊上发表了 30 多篇论文，包括 T-PAMI、CVPR、ICCV、NeurIPS 等。他的研究兴趣包括深度学习和计算机视觉。

李翔 是南开大学计算机学院副教授。于 2020 年在南京理工大学获得博士学位。他的研究兴趣包括 CNN/Transformer 主干网络、目标检测、知识蒸馏和自监督学习。他在 TPAMI、CVPR、NeurIPS 等顶级期刊和会议上发表了 30 多篇论文。

王萍 分别于 1988 年、1991 年和 1998 年在天津大学获得计算机科学学士、硕士和博士学位。她目前是天津大学数学学院教授。她的研究兴趣包括图像处理和机器学习。

程明明 于 2012 年在清华大学获得博士学位。然后在牛津大学师从 Philip Torr 教授进行了 2 年的博士后研究。他现在是南开大学的教授，领导媒体计算实验室。他的研究兴趣包括计算机图形学、计算机视觉和图像处理。他获得的科研奖项包括国家杰出青年科学基金和 ACM 中国新星奖。他是 IEEE TPAMI 和 IEEE TIP 的编委会成员。