在对象检测中使用少量样本复制粘贴添加新类别

Boyang Deng, Meiyan Lin, Shoulun Long

摘要—开发高效的数据实例检测模型，能够处理罕见的对象类别，仍然是计算机视觉中的一个重要挑战。然而，现有的研究往往忽视了针对涉及神经网络的真实世界场景而设计的数据收集策略和评估指标。在这项研究中，我们系统地调查了专注于对象遮挡的数据收集和增强技术，旨在模拟实际应用中观察到的遮挡关系。令人惊讶的是，我们发现即使是一个简单的遮挡机制，在引入新的对象类别时也足以达到强大的性能。值得注意的是，通过向包含超过五十万张图像并涵盖数百个类别的大型训练数据集中添加仅 15 张新类别的图像，模型在未见过的新类别实例的测试集上实现了 95%的准确率。

***Index Terms*—深度学习；数据增强；少量样本检测**

I. 介绍

目标检测是计算机视觉中的一个基本任务，具有众多实际应用。然而，基于卷积神经网络的最先进的目标检测模型通常需要大量数据 [1]。为对象检测标注大规模数据集既昂贵又耗时。例如，在我们的智能货架数据集中，四个工人花费大约一个小时才能标注出 3,000 个对象边界框。这突显了开发能够提高现代目标检测模型数据效率的方法的紧迫性。

许多研究试图通过架构创新来提高检测性能 [1]–[3]，这类方法通常会引入权衡，包括增加推理时间或增加模型复杂性。相比之下，我们的重点是开发通用策略，通过数据增强技术提升模型性能，如在 [4] 中所示。我们探索了如何高效地将新类别添加到现有数据集中，同时尽可能减少图像收集和标注的工作量。一个有前景的方法是基于真实图像的少样本学习，在这种方法中，使用有限数量的真实图像来代表新类别，同时仍然力求保持高检测精度。

我们提出，有效管理数据收集和增强是提高目标检测模型数据效率的直接且有效的方法。在多样化图像分布上训练检测网络已显示出显著的好处 [5]，并且通过引入物体遮挡可以进一步丰富训练数据中的挑战性场景 [6]。在数据收集阶段，使用真实物体捕获自然遮挡，在增强阶段，则通过将提取的边界框叠加到目标物体上合成生成遮挡。

尽管边界框注释比分割掩码更容易获得，但它们可能包含部分背景，导致粘贴到新图像时出现不一致。这使得基于遮挡的边界框增强方法不如基于分割的方法优化。然而，我们的实验表明，通过精心设计，使用边界框的遮挡增强仍能显著提高检测精度。

受最近的数据增强技术 [7], [8] 启发，我们提出了一种新的基于复制粘贴的方法，仅使用边界框注释来训练目标检测网络。我们的方法与在 [9] 中引入的增量学习概念一致，旨在高效地纳入新类别。然而，与仅使用合成数据的方法 [10] 不同，我们发现完全依赖于合成数据会在真实世界场景中产生次优结果。

II. 方法

本研究的核心思想是在构建训练数据集时模拟物体在现实世界场景中发生的遮挡。该方法能够创建多样性和组合性的遮挡关系，包括各种可能性：

1）选择相互部分遮挡的多个对象；
2）定义这些物体之间的遮挡关系；
3）确定物体位置和相机视角以捕捉预期场景。

我们的基于复制-粘贴的数据生成方法引入了不同程度的遮挡，以模拟现实的对象交互。我们假设对象之间的遮挡关系是神经网络学习的关键因素，特别是在使用新类别中的少量标注示例进行训练时。通过让模型接触目标对象的部分视图，该方法鼓励在遮挡情况下的鲁棒特征学习。

实验结果表明，遮挡的结构，包括其严重程度、视角和可见区域，比遮挡物体的具体类别更具影响力。这表明复制现实中的遮挡模式可以使即使在标注数据有限的情况下也能实现有效学习。我们还发现，仅标注物体的可见部分而忽略遮挡区域会导致更快的收敛和提高检测精度。

Fig. 1. 饮料仅安排示例。

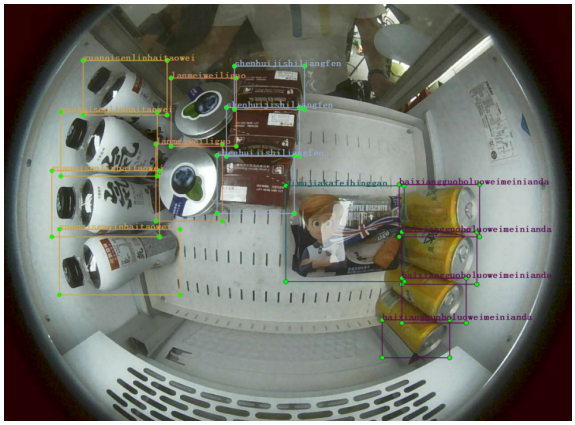

Fig. 2. 饮料和小吃的排列示例。

A. 遮挡物

在现实世界中的物体遮挡关系中，小物体通常只会遮挡大物体的部分区域，而大物体则可以遮挡住小物体的较大区域。这些遮挡关系在自然环境中遵循相对固定的分布，这为我们通过瞄准重要样本点来在合成数据集中复制这种分布提供了机会。

例如，考虑来自类别 X 的目标对象 A。在一个给定的遮挡分布中，如果对象 A 的底部 50%被来自类别 Y 的对象 B 遮挡，并且没有来自类别 Y 的对象可用，我们可以使用来自类别 Z 的对象 C 来遮挡 A 的相同部分。这会产生类似的遮挡效果，表明遮挡物的具体类别不如复制遮挡关系重要。

我们通过货架上商品的典型摆放方式来说明这一概念，如图 1和 2所示。由于我们使用了带有超广角镜头的鱼眼摄像头，该镜头会引入强烈的视觉失真以创建半球形图像，因此我们必须仔细安排物品的位置以确保所有商品都在摄像头的视野范围内。高大的物品，如饮料，应放在靠近货架墙壁的位置，而较短的物品则应放置在中央。随着物品尺寸的增加，它们应该被放置得更靠外侧。这确保了所有的物体都能被鱼眼摄像头看到，并且可以由神经网络模型检测到。

准备各种尺寸的物体对于模拟广泛的遮挡关系至关重要。例如，在真实世界场景中，目标物体可能被较小的物体部分遮挡，而较大的物体可以遮挡目标物体更显著的部分。

遮挡可以在数据采集或数据增强阶段引入。在数据采集过程中，每个对象类别的大小对于生成真实的遮挡关系起着关键作用，因为不同大小的对象自然会产生不同类型的比例。相比之下，在数据增强阶段，对象的大小不太重要，因为我们可以用任何类别来遮挡目标对象。可以使用复制-粘贴、剪切-粘贴、图像缩放和图像平移等技术有效地模拟这些遮挡。

B. 遮挡关系

准确识别真实世界中的物体遮挡分布对于有效的模仿至关重要。在特定场景中，物体的遮挡依赖于多种因素，包括相机视角、物体大小和物体位置。物体之间的关系必须合理；例如，在室内环境中，根据视角的不同，桌子上的杯子可能会部分或完全被纸夹遮挡，而电视遥控器更可能放在杯子旁边而不是上面。因此，我们优先收集常见的遮挡关系，确保每种类型的遮挡由一个或两个案例代表，这足以在真实世界测试案例中实现高精度。

为了模拟这种遮挡分布，我们应用蒙特卡罗方法来采样数据点。首先，我们在实际场景中识别新类别的遮挡分布，并单独处理每个类别。然后，我们通过应用复制粘贴技术根据这种分布遮挡住新类别的对象来生成合成图像。这些合成图像与少量真实图像配对以训练检测网络。

在数据收集阶段，我们缺乏新类别特定的遮挡分布信息，但我们可能能够访问来自以前数据的类似大小的类别。在生成遮挡时，新类别的表面材料或纹理并不重要。然而，新类别的放置取决于其尺寸和内在特征。例如，在 FVSS 数据集中，较小的物品如包装零食或罐装饮料被放置在货架层的中心，而较大的物品，如零食袋，则被放在外围。

为了最大化货架上的空间利用，货物被排列以确保它们都能从顶部中央的鱼眼摄像头处可见，并且每个项目的可视区域足够明显以便人类识别。每个项目的顶部或侧顶部部分必须可见，避免任何堆叠货物的情况。在

一个完全装满的一层中，物体的下部通常会被相邻物品遮挡。较小的物品放置在中心位置，而较大的物品则被放置在该层边缘或靠近货架墙壁的位置。这种排列方式最小化了同类或不同类物品之间的遮挡。

例如，当添加一个新的大件物品，如水瓶时，通常会将其放置在边缘或靠近货架壁的位置以减少被附近物体遮挡。遮挡比例取决于相邻的物体：如果一个水瓶被另一个相同的瓶子遮挡，可能只有瓶盖可见，而一个小牛奶盒可能会遮挡瓶子的三分之二。一个平放的零食袋可能只会遮挡大约三分之一。此外，应避免将较大的物品放置在鱼眼摄像头中心附近，以防完全遮挡较小的物品。

在数据遮挡阶段，我们通过使用已标注的图像生成新的遮挡图像，遵循为新目标类别识别出的遮挡分布。第一步是确定新类别的正确遮挡分布。最合理的做法涉及分析新类别的属性，并基于经验丰富的研究人员的专业知识推断出遮挡分布。然而，这种方法难以通用化，因为它需要针对每个新类别都有专家知识。因此，需要一种自动化的方法。

例如，在 COCO 数据集中，“人”类别经常被标注在如站在街上或坐在桌子旁的场景中。一个人可能被其他人在户外遮挡，或者在室内被桌子遮挡。有趣的是，即使在拥挤或远处的场景中，人的头部几乎总是可见的。如果一个人的头部不可见，这很可能意味着数据集组织者没有收集这样的图像，因为人类通常通过头部来识别他人。此外，标注人员可能不愿意对只能看到人下半身而看不到头部的图像进行标注。

一些在目标检测中关于遮挡的显著特征如下：1) 一个人的头部可能被雨伞遮挡。2) 头部可能出现侧面或背面，这在数据收集过程中需要考虑。在数据增强阶段，重点应放在模仿真实的遮挡关系，而不是不同的视角。3) 在极少数情况下，图像中可能只显示人体的小部分，如特写的手或脚，但仍被标注为人。这些特征可以推广到其他类别，包括动物，它们通常会在图片中展示头部以方便观察者识别。

COCO 数据集种类繁多，包括多种环境、光照条件、姿势和视角。例如，“熊”类别包含多个子类别，如北极熊、黑熊、棕熊和浣熊。通过添加一个新的仅包含少量图像的类别——比如几十张图像——就可以有效地训练检测网络，同时整合新旧类别。

对于较小的物体，如牙刷或遥控器，这些物体可能在近景和远景中都被捕获，导致图像中的物体大小存在显著差异。一个常见的问题是：将小物体的遮挡分布应用到较大物体上是否有效？我们的研究结果表明，这确实是有用的，并且通过将图像缩放作为数据增强技术，性能可以进一步提高。

我们还分析了 Open Images Dataset [6]，该数据集包含大量样本，并在各类别之间具有更大的多样性。该数据集提供了四种类型的注释：检测、分割、关系和局部叙述。检测注释使用边界框，而分割注释则由多边形表示。关系注释捕捉了人类与物体之间的各种互动或不同物体之间的互动，虚线边界框表示一个物体内含另一个物体。这些关系与类别遮挡分布紧密相关，并为我们的工作提供了有价值的见解。

C. 相机视角

模仿所有可能的视角，特别是在大型户外场景中，可能会很有挑战性。然而，我们发现了一个简单的方法，在现实世界设置中实现了高精度。我们采用了来自 NERFIES [11] 的方法，使用主摄像头视角以及几个稍微偏移的视角来捕捉物体图像，同时忽略罕见或极端的视角。

D. 复制粘贴增强

在为我们的新 SKU 收集了数十张图像后，我们采用了复制粘贴数据增强策略 [8] 来涵盖更多表示数据放置和遮挡分布的样本点，从而提高检测性能。复制粘贴策略涉及根据我们的数据遮挡分布将一个图像中的边界框区域随机转移到另一图像中，同时确保与现有物体的重叠最小。

E. 基于 *FairMOT* 的标注

边界框主要用于标注对象。在受控数据收集场景中，我们可以缓慢地将物体移动穿过帧，从而可以使用跟踪模型来标注每个连续运动的对象，从而减少人工注释者的工作量。

我们最初测试了单目标跟踪（SOT）。通过在一个片段中缓慢移动一个物体，我们可以在第一帧中标注出目标物体，同时保持其他物体静止。我们假设相机保持静止，尽管缓慢移动相机可以增强 SOT 性能。然而，SOT 存在两个问题：1) 当同一类别的多个物体放置得很近时，跟踪器可能会转移到附近的另一个物体上，这需要一个更准确的 SOT 模型来实现可扩展标注。2) 由于 SOT 仅支持单目标跟踪，因此需要许多片段来

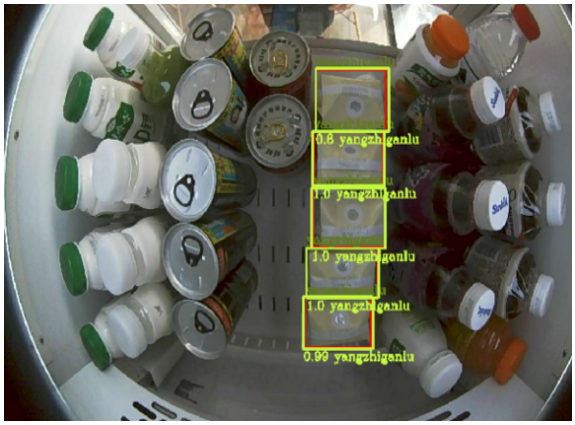

Fig. 3. 小盒装饮料的检测结果。

Fig. 4. 低高度零食的示例。

分别标注不同的物体。为了解决这些问题，我们采用了 FairMOT 进行多目标跟踪（MOT），从而实现了多个对象的同时跟踪，并进一步简化了标注过程。

III. 实验

实验表明，我们的方法只需要每个新类别几十张图像即可达到与使用成千上万张图像训练的模型相当的准确性。我们探索了两个实验方向：数据收集阶段的数据遮挡和数据增强阶段的数据遮挡。

我们的数据集，鱼眼视图货架 SKU（FVSS），用于数据收集阶段的验证。该数据集提供了货架层的鱼眼摄像机视角，如图 3和图 4所示，并为每个类别标注了边界框。在这些实验中，我们使用数百个类别作为基础数据集，并尝试添加一个新的类别。

对于数据增强阶段，我们使用 COCO 数据集作为测试平台。COCO 包含 80 个类别，从中随机选择一个新类别，并将剩余的 79 个用作基础数据集。我们分析新类别的遮挡分布，并选择 1%到 10%的图像作为训练的关键样本点。对于检测，我们使用 YOLOv5-small 模型并将所有标注转换为 YOLOv5 格式。

A. 数据收集

实验在一个货架环境中使用鱼眼摄像头进行，遵循 FVSS 数据集构建风格。我们的基础训练数据集包含 10,000 张图像，涵盖了 457 个类别。我们在基础数据集中添加了一个新的只有 10 张图像的类别，并在包含至少一个新类别的边界框的 1,000 张图像验证集上评估性能。我们数据集中两个与遮挡分布相关的类别的热图如图 5所示。

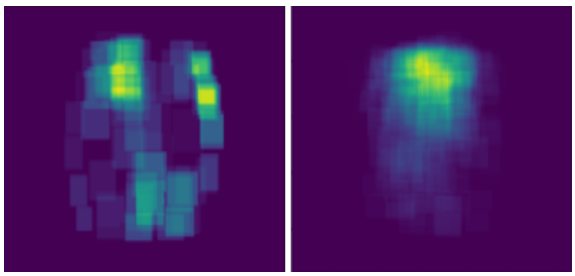

Fig. 5. 两类热图，“光拾波粒”（左）和“洋紫甘露”（右）。

例如，10 张“可乐罐”的图像被添加为一个新类别到包含 457 个类别的 10,000 张图像的训练数据集中，这些类别中没有包括“可乐罐”。这 10 张图像是从货架环境中“可乐罐”的数据遮挡分布的重要样本点中精心挑选出来的，结果得到了 58 个边界框。然后我们构建了一个包含 179 个类别的 1,000 张图像的验证数据集，每张图像至少包含一个“可乐罐”边界框（总共 3,939 个边界框）。我们使用三个指标来评估性能：1) 验证数据集中“可乐罐”的 AP@0.5 和 AP@0.5:0.95；2) 通过率衡量是否所有“可乐罐”实例都被正确检测到；以及 3) 错误分类率表明一个“可乐罐”是否被错误地分类为另一个类别，且置信度低于 95%。结果如表 I所示。

16 个不同的类别进行了测试，每个类别都作为新类别单独添加。这些类别都是零售领域的商品，比如零食、牛奶、饮料等。我们的研究结果表明，我们可以仅用少量图像训练一个新的类别，并保持超过 80%的准确率和平均超过 85%的错误分类率。这意味着验证数据集中只有 3%的图像是被高置信度错误分类或完全未被检测模型识别的。某些类别的结果显示在表 II中。错误分类率定义为置信度低于 90%的情况。我们还引入了一个新的指标，即“严重错误率”，专门衡量以超过 90%的

Table. I. 特定于“易拉罐”的结果。

AP@0.5	AP@0.5:0.95	pass rate	wrong-class rate	wrong-class rate@0.95
98.4%	83.6%	81.3%	77.0%	87.0%

Table. II. 将每个类别视为一个新的分类进行的实验。

name	image number	pass rate	wrong-class rate	severe error rate
xiandangao	6015	78%	91%	1.98%
yibaochunjingshui	2037	54%	92%	3.68%
jiaduobaoguan	1238	42%	78%	12.76%
cuiguoba	6359	91%	80%	1.8%
420meizhiyuanguolicheng	712	95%	95%	0.25%
feizixiaolizhi	1884	52%	95%	2.4%
heqingjiaotangbinggan	3569	84%	92%	1.28%
duoweixiaoxibing200	1799	90%	90%	1.0%
4wahahaadgainai	2807	55%	100%	0.0%
yizhongtaohuangtaoguantou	2520	78%	80%	4.4%
heqingjiaotangbinggan	3992	94%	86%	0.84%
4wahahaadgainai	3094	32%	91%	6.21%
enaakdianxinmian30g	488	62%	76%	9.12%
guowangshiguangguoba	867	90%	100%	0.0%
mailisu	531	95%	95%	0.25%
average	3194	72%	90.7%	4.2%

Table. III. 零样本 vs 少样本。

name	pass rate	wrong-class rate
w/o new category data	0.0%	79.0%
w new category data	72.0%	90.0%

置信度将边界框误识别为另一个类别或完全未被检测到的比例。结果呈现在表 II和表 III中。

此外，我们发现使用各种不同大小的类别来生成多样化的数据遮挡关系显著提高了模型的性能，几乎将平均准确率翻了一倍。

通过添加来自不同域的新类别图像，例如由手持智能手机拍摄的图像，并与货架鱼眼图像一起进行跨域实验。未应用任何领域适应方法来增强性能。结果显示，在 1000 张验证数据集图像中，新类别的通过率为 0%。对于新类别的错误分类率几乎与未添加新类别数据的情况相同。这表明在使用新类别进行训练时，领域适应仍然是一个挑战。结果如表 III所示。

接下来，评估了向训练数据集添加新类别的影响。在包含新类别数据的验证数据集中，我们发现如果新类别未被纳入训练数据集，则该类别无法在任何验证图像中正确检测到，导致通过率为 0%。此外，21%的边界框要么被漏检，要么以高置信度错误地检测为其他类别。然而，当我们仅向训练数据集添加了 10 张新类别的图片时，验证数据集中新类别的通过率提高到了 72%，且在 0.90 置信度下的错分率降低至 90%。我们还观察到，高宽比的类别倾向于在 0.90 置信度下显示出更大的错分率增加。添加这些高宽比类别的图片可能会减少误分为其他类别的机会。尽管如此，当仅使用少量图像进行训练时，高宽比的类别往往通过率较低。这表明即使是少量的图像也能显著提高新类别检测的效果，特别是当这些图像是很好的遮挡分布样本时。神经网络可以学习到最重要的特征。结果如表 IV所示。

在进一步的实验中，我们测试了向一个大型数据集添加仅包含少量图像的新类别。我们使用了一个包含超过 360,000 张图像的大尺寸鱼眼视图数据集，并添加了一个新类别 “sizhoushaokaoweixiatiao”，该类别仅有 15 张图像和 60 个边界框。经过 1.5 轮的训练，采用了常见的数据增强技术如图像翻转、色调调整和标准化后，我们在包含 500 张真实图像的测试数据集上评估了模型。只有大约 10 张图像是被错误分类的。这一结果展示了使用复制粘贴策略结合数据遮挡分布来利用边界框注释训练有效的检测模型的可能性。结果如 6和 7图所示。

Table. IV. 零样本结果。

name	image number	wrong-classed rate
xiandangao	6015	87.0%
yibaochunjingshui	2030	87.0%
jiaduobaoguan	2945	81.0%
cuiguoba	6992	90.0%
420meizhiyuanguolicheng	712	78.0%
feizixiaolizhi	1884	87.0%
heqingjiaotangbinggan	3569	52.0%
duoweixiaoxibing200	1799	93.0%
4wahahaadgainai	2805	44.0%
yizhongtaohuangtaoguantou	2520	82.0%
average	3127.1	79.3%

Table. V. 新类别训练的比较。

sku name	3000+ bboxes	60 bboxes (20 images)	370 bboxes (20 images + copy-paste)
guangshiboluopi	33.83%	12.7%	53.38%
yangzhiganlu	60.96%	34.76%	76.83%
zhiqingchunniunai	49.87%	3.56%	27.95%
tengyeyicunxiaoyuanbinggan	95.98%	38.16%	98.19%
aolangtangeweihuabinggan	37.50%	58.33%	97.22%
average	55.63%	29.52%	70.71%

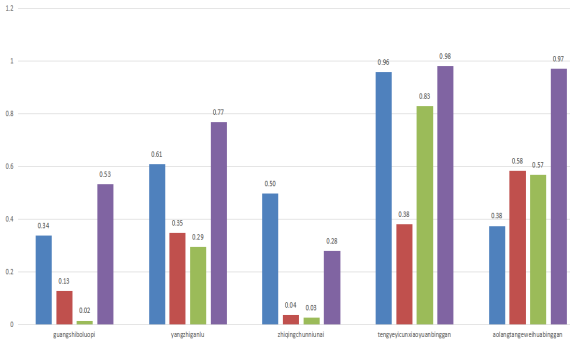

Fig. 6. 复制粘贴数据增强结果比较：蓝色代表使用目标类别的原始 3000 多个边界框训练的模型。红色显示仅使用 15 个随机采样边界框的性能。绿色表示使用从不同角度拍摄的 17 张近景智能手机图像进行训练的结果。紫色显示同时使用 17 张近景智能手机图像和 370 个随机采样边界框的综合结果。

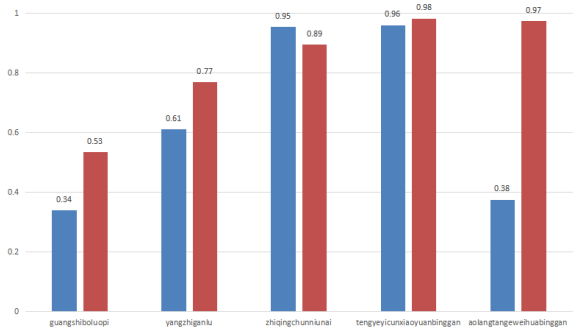

Fig. 7. 复制粘贴的”志清春牛奶”的结果。蓝色代表使用目标类别的原始 3000 多个边界框训练的模型。红色显示了结合使用 17 张近景智能手机图像和 370 个随机采样的边界框的结果。

B. 数据增强

比较使用了一个包含超过 3,000 个边界框的正常数据集与仅使用 20 张新类别图像进行训练。这 20 张图像是之前提到的大数据集的一个子集。我们对这 20 张新类别的图像进行了两种类型的实验。

在第一次实验中，我们仅使用这 20 张图像训练模型，并应用了数据增强技术，如图像翻转、HSV 变换和色相调整。在第二次实验中，我们采用了复制粘贴的数据增强策略，根据数据遮挡分布生成了另外 100 张图像来自原来的 20 张图像。这样新类别的图像总数达到了 120 张，其中 100 张是复制粘贴的。我们依次测试了五个新的类别，并将结果展示在表 V中。结果显示效果显著，使用少量图像结合复制粘贴增强的方法优于对原始大数据集进行训练。图 8展示了一张采用我们所使用的复制粘贴策略增强后的图像，而图 9则展示了网络使用我们的方法训练后的一个检测失败案例。

在某些情况下，我们可能会使用已经收集的数据集进行训练，并且无法控制数据收集阶段。然而，我们仍

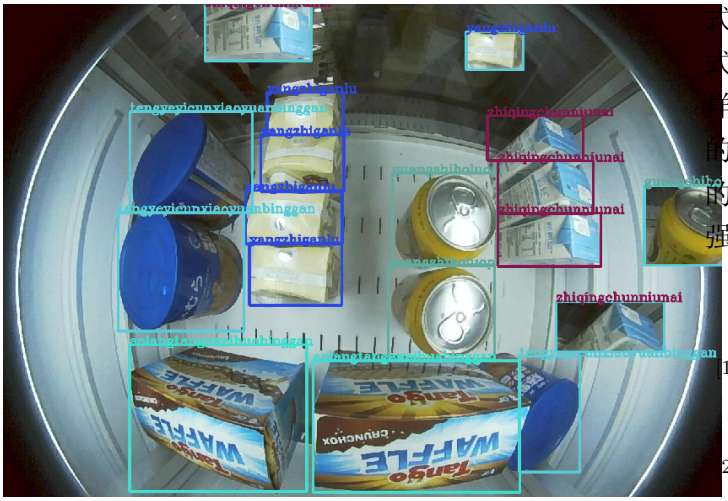

Fig. 8. 复制粘贴的边界框示例。

然希望向现有数据集中添加一些新类别的图像。为了证明即使是少量的新类别图像也能达到相对较高的准确率，我们设计了实验，在这些实验中，我们从测试数据集中新类别的数据遮挡分布的重要样本点中选择这些图像。

我们的实现灵感来源于 [8] 中的方法，他们利用简单的复制粘贴数据增强策略来显著提高准确率。我们认为这一结论是由复制粘贴操作生成许多新的遮挡关系所驱动的，捕捉到了数据遮挡分布的重要样本点。图 10 和 11展示了两类别的测试结果，说明了目标类别在测试数据集中的置信度分布。

IV. 结论

数据收集是将视觉系统应用于实际任务中的核心步骤。在本文中，我们提出了一种对象遮挡数据收集方法，该方法已被证明既有效又稳健。对象遮挡在多种实验设置下表现出色，并且即使使用少量数据也能带来显著改进。我们的实验基于 FVSS 数据集和 COCO 基准。

我们提出的对象遮挡数据收集和增强策略简单易行，可以集成到任何数据集中，无论是构建新数据集还是向现有数据集添加新类别。这种方法通过仅需少量图像即可减少训练成本。因此，我们可以使用具有适当数据遮挡策略的小模型——例如复制粘贴技术——来创建目标对象的合适遮挡关系。此方法在训练过程中还减少了内存使用。适当的对象遮挡数据收集和增强策略使小模型能够达到与更复杂模型相当的准确性。

我们的研究显示，网络可以从少量样本中学习新类别，类似于人类凭借其强大的推理能力进行学习的方式。另一方面，人类的学习也需要模仿网络的学习方式，这涉及较少的分析和推理能力但需要接触到更多的样本。这表明，通常通过展示更多示例而无需详细解释的网络学习过程，对于学习新概念或语言可能是有益的。未来的工作可以集中在改进物体遮挡数据收集和增强策略，以适用于更多类型的物体。

参考文献

[1] Z. Tian, C. Shen, H. Chen, and T. He, “Fcos: Fully convolutional one-stage object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[2] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[3] T. Rong, Y. Zhu, H. Cai, and Y. Xiong, “A solution to product detection in densely packed scenes,” *arXiv preprint arXiv:2007.11946*, 2020.

[4] Z. Zhang, T. He, H. Zhang, Z. Zhang, J. Xie, and M. Li, “Bag of freebies for training object detection neural networks,” *arXiv preprint arXiv:1902.04103*, 2019.

[5] K. He, R. Girshick, and P. Dollár, “Rethinking imagenet pre-training,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4918–4927.

[6] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, “The open images dataset v4,” *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

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

[8] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, “Simple copy-paste is a strong data augmentation method for instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2918–2928.

[9] K. Shmelkov, C. Schmid, and K. Alahari, “Incremental learning of object detectors without catastrophic forgetting,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3400–3409.

[10] S. Hinterstoisser, O. Pauly, H. Heibel, M. Martina, and M. Bokeloh, “An annotation saved is an annotation earned: Using fully synthetic training for object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[11] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, “Nerfies: Deformable neural radiance fields,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5865–5874.

Fig. 9. 检测失败案例示例。这两类表现出高度的视觉相似性。

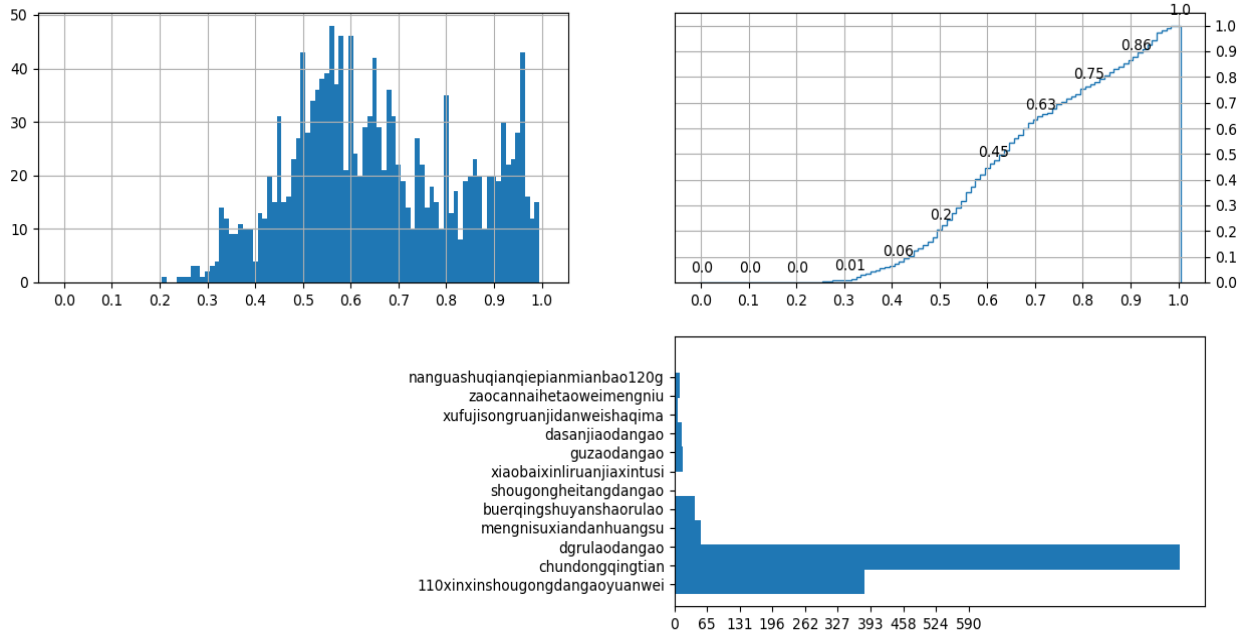

Fig. 10. 基于 30 张“现当高”图像的检测结果，通过率为 79%。该图包括三个图表：置信度分布（左上）、累计置信度（右上）和检测到的类别数量（右下）。

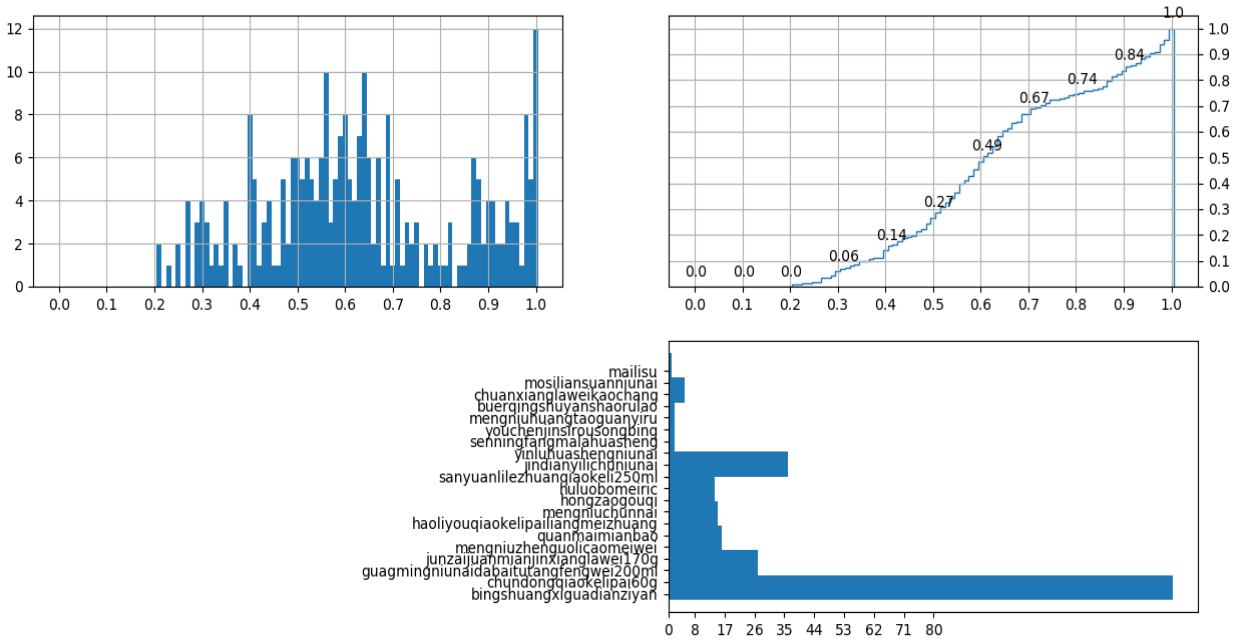

Fig. 11. 基于“和庆蕉堂病甘”30 张图像的检测结果，通过率为 93%。有三个子图，包括置信度分布（左上）、累积置信度（右上）和检测到的类别数量（右下）。