中图分类号：TP391.6　**文献标识码：**A　**文章编号：**1006-8961(
论文引用格式：（论文引用格式：）[DOI:]

概率图采样图像增广驱动的弱监督物体检测

李笑颜[1,2]，阚美娜[1,2]，梁浩[1,2]，山世光[1, 2, 3]

1.中国科学院智能信息处理重点实验室，北京市 100190；2.中国科学院大学 计算机科学与技术学院，北京市 100049；3.鹏城实验室, 广东省深圳市 518055

摘　要：弱监督物体检测是一种仅利用图像类别标签训练物体检测器的技术。近年来弱监督物体检测器的精度不断提高，但在如何提升检出物体的完整性、如何从多个同类物体中区分出单一个体的问题上仍然面临极大挑战。围绕上述问题，本文提出了基于物体布局后验概率图进行多物体图像增广的弱监督物体检测方法——ProMIS。该方法将检出物体存储到物体候选池，并将候选池中的物体插入到输入图像中，构造带有伪边界框标注的增广图像，进而利用增广后的图像训练弱监督物体检测器。该方法包含图像增广与弱监督物体检测两个相互作用的模块：图像增广模块将候选池中的物体插入一幅输入图像，该过程中通过后验概率的估计与采样对插入物体的类别、位置、尺度进行约束，以保证增广图像的合理性；弱监督物体检测模块利用增广后的多物体图像、对应的类别标签、物体伪边界框标签训练物体检测器，并将原始输入图像上检到的高置信度物体储存到物体候选池中。训练过程中，为了避免过拟合，本文在基线算法的基础上增加一个并行的检测分支，即基于增广边界框的检测分支，该分支利用增广得到的伪边界框标注进行训练，原有基线算法的检测分支仍使用图像标签进行训练。测试时，本方法仅使用基于增广边界框的检测分支产生检测结果。本文提出的增广策略和检测器的分支结构在不同弱监督物体检测器上均适用。在 Pascal VOC 2007、Pascal VOC 2012 数据集上，将该方法嵌入到多种现有的弱监督物体检测器中，mAP 平均获得了 2.9%、4.2%的提升。

关键词：弱监督物体检测；多物体数据增广；图像融合；概率图采样；后验概率估计

ProMIS: weakly supervised object detection with probability-based multi-object image synthesis

Xiaoyan Li[1,2]，Meina Kan[1,2]，Hao Liang[1,2]，Shiguang Shan[1, 2, 3]

1.Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China；2.University of Chinese Academy of Sciences, School of Computer Science and Technology, Beijing 100049, China；3. Peng Cheng Laboratory, Shenzhen, 518055, China

Abstract: The fully supervised object detectors based on neural networks improve the performance of object detection by a large margin and make them more reliable for real-world applications. However, they are over-reliant on the amount of annotated data. Considering that the workload for labeling a bounding box is non-negligible and the categories and application scenarios are countless, it is difficult to collect large-scale detection training datasets that can satisfy all real-world applications. Thus, the weakly supervised object detector is designed to reduce the annotation workload of object detection by only requiring image category annotations for training. Recently, weakly supervised object detectors show promising performances with the multiple instance learning (MIL) technique. In these methods, object proposals are classified and aggregated into an image classification result, and objects are detected by selecting the bounding box that contributes most to the aggregated image classification results among all object proposals. However, since weakly

收稿日期：；　修回日期：
基金项目：国家重点研发计划（No.2017YFA0700800），国家自然科学基金（Nos.62122074 和 62176251），北京市科技新星（Z191100001119123）Supported by National Key Research and Development Program of China (No.2017YFA0700800), the National Science Foundation of China (Nos.62122074 和 62176251), the Beijing Nova Program (Z191100001119123)

supervised object detection lacks instance-level annotations, how to differentiate an instance from a part of the instance or a cluster of multiple instances of the same category still remains challenging. We propose to learn the ability to distinguish instances by inserting detected objects with high confidence into an input image and generating the augmented image with pseudo bounding box annotations for training the object detector. However, it is found that the naive random augmentation method can not immediately improve the detection performance, owing to the following reasons: 1) the generated object annotations are sourced from the detection results of the detection head, and it easily leads to over-fitting when the generated data is used to train the detection head itself; 2) the spatial distribution of the generated objects are often quite distinct from the real data, since the hyper-parameters of the insertion are all sampled from uniform distributions, resulting in unreasonable augmented images (for example, a TV monitor is placed in the sky). To solve the above issues, the weakly supervised object detection with probability-based multi-object image synthesis (ProMIS) approach is proposed in this work with two iterative and interactive modules, namely the image augmentation module and the weakly supervised object detection module. In each training iteration, objects are detected in the original input image with the weakly supervised object detector (to ensure the accuracy during the initial training, the detector is pre-trained according to its baseline method), and the highly confident detected objects are stored in an object candidate pool for the latter image augmentation. The image augmentation module inserts one or more objects sampled from the object candidate pool to the input image for an augmented training image with pseudo bounding box annotations. In this process, to make the augmented image reasonable and realistic, the object category, position, and scale for the insertion are sampled from the estimated posterior probability maps with the detected objects in this image as references. Tree kinds of posterior probabilities are proposed in the ProMIS in charge of describing the category, spacial and scale relation of an object and another referenced object, respectively. These posterior probabilities can be estimated online according to the objects detected in the previous training iterations and the hyper-parameters of the newly inserted objects are assumed to obey these posterior probabilities. Then, the detection training module exploits the augmented image and its pseudo annotations to train the weakly supervised object detector. In the training process, to avoid over-fitting to the detected false positives, a new parallel detection branch is added to the baseline weakly supervised object detection head. The augmented bounding box annotations are only used to guide the newly added branch, while the original weakly supervised detection head is employed during the generation of the augmented data and is trained with only image-level labels. In the inference process, only the added branch trained with the augmented annotation is kept for generating the testing results, which keeps the efficiency of the weakly supervised object detector in inference. The above image augmentation module and the weakly supervised object detection module work iteratively and interactively, and steadily facilitate the weakly supervised object detector to learn the ability for distinguishing instances. The proposed ProMIS is an online augmentation method and does not require any additional images or annotations except the original weakly supervised detection training data. In addition, since the proposed approach is independent of the selection of the weakly supervised object detector, the proposed augmentation paradigm is generalized for all detector architectures. In the experiments, the effectiveness of the proposed parallel detection branch and the posterior probability maps is verified, and they improve the naive random augmentation method by 5.2% and 2.2%, respectively. The proposed ProMIS approach is applied to multiple previous weakly supervised object detectors (including OICR, SDCN, and OICR-DRN), and compared with these baseline methods, it achieves an average of 2.9%, 4.2% improvements on the Pascal VOC 2007 and the Pascal VOC 2012 datasets, respectively. Moreover, the error mode analysis is conducted and it is found that the proposed ProMIS can decrease the error mode of the ground-truth in the hypothesis and the hypothesis in the ground-truth, demonstrating that ProMIS make fewer mistakes when distinguishing instances from its parts or multiple instances of the same category.

Key words: weakly supervised object detection; multi-object data augmentation; image synthesis; probability map sampling; posterior probability estimation

0 引 言

近几年，基于大数据强监督的物体检测方法不断完善，在学术研究和实际应用中都取得了巨大成功。然而，现有的强监督物体检测模型均采用“标注-训练-推理”的流程，该流程建立在大规模数据标注的基础之上。标注大规模物体检测数据的金钱和时间成本都很高，而且真实世界中物体种类多、难以枚举。即使是同类物体，在不同自然环境、成像设备、运动状态下也存在较大的类内差异。为了获得鲁棒的检测器，需要标注大量具有不同类内差异的样本。与此同时，互联网及大型数据库中含有大量较为廉价、仅带有类别标注的图像。这些类别标注相较于检测框仅是一种弱标注，虽无法提供精准的物体定位，但却很容易获得。因此，弱监督物体检测技术目标是仅使用图像类别标注来训练物体检测器，以减轻检测模型对强标注数据的依赖。

弱监督物体检测的目标是仅利用图像类别标签定位出物体完整、精准的边界框。如何提升物体边界框的完整性、如何从多个同类物体中区分出单一个体是其核心挑战。近年来，研究者们对这些问题进行了许多卓有成效的研究和探索。例如，Blien等人（Bilen 和 Vedaldi，2016）提出弱监督深度检测模型（WSDDN），该模型使用两个全连接层分别对物体候选框进行分类和选择，并使用这两个结果聚合成整幅图像的分类结果从而对检测器进行训练。Tang等人（Tang 等，2017）发现WSDDN的结果常聚集在物体的显著部件上，为此提出叠加多个检测细化分支（OICR）来利用WSDDN产生的伪边界框标签进行学习，以提升检出物体的一致性并得到更好的检测结果。Li等人（Li 等，2019）进一步使用基于对抗训练的弱监督分割分支得到类别相关区域，并采用分割分支和检测分支互相监督的方式改进检出物体的完整性。然而，现有方法虽然探索了如何保证物体的完整性（Li 等，2019；Shen 等，2019；Ren 等，2020），即区分物体部件和物体整体，却并未显式保证同类物体之间的可区分性，即一张图像上多个同类物体可能被混淆为一个物体。

为了学习区分单个物体，一个直观想法是将检测到的物体随机融合到一幅输入图像上，生成具有多物体的图像和对应的伪标注框，从而强制检测器学习物体之间的区分信息。其中，最简单的方案是完全随机增广，即直接在均匀分布中采样控制插入过程的超参数（如：插入哪个物体、在输入图像上的插入的位置和尺度），并依此进行图像增广。本文实现了这种随机增广，并将其应用在OICR方法上，结果如表1所示。这种简单的随机增广方式并不能直接提升检测器性能，反而使检测性能由41.2下降到39.0。主要原因在于：1）生成的伪边界框标签源于检测器之前的检测结果，使用这些伪边界框标签继续训练的原始的检测器容易导致过拟合，并导致性能下降。2)生成的多物体布局分布和真实多物体布局分布差距大。这种差异导致在这些数据上学习到的检测器难以适应到真实图像的分布。如图1所示，将检测到的物体随机插入图像很容易产生不合理的增广图像，如电视机被插入到室外的天空上，轮船被插入到室内的椅子旁边。

表 1 在 Pascal VOC 2007 数据集上对 OICR 方法（Tang 等，2017）应用不同增广方法后的弱监督物体检测性能

Table 1 The performance of OICR detector with different data augmentation methods on the Pascal VOC 2007 dataset

增广方法	基于增广边界框的检测分支	mAP（%）
无		41.2
随机增广		39.0
随机增广	✓	44.2
本文增广	✓	**46.4**

注：加粗字体为该列最优值。

图 1 随机增广的多物体图像

Fig.1 Augmented images by uniformly random sampling insertion parameters

针对上述问题，本文提出了基于概率图采样增广的弱监督物体检测方法，简称为ProMIS方法。该方法是在线增广方法，无需额外的图像和数据。它将前几轮训练累积的高置信度检出物体存入物体候选池，从候选池中采样一个或多个物体插入到当前输入图像上，产生增广图像及对应边界框标签，进而用于本轮检测器的训练。该方法包含图像增广与弱监督物体检测两个相互作用的模块。图像增广

模块将候选池中的物体插入一幅输入图像得到增广图像。该过程中，图像增广模块利用弱监督物体检测器的历史检出情况估计插入物体的类别、位置、尺度的后验概率，并在这些后验概率中进行采样以得到较为合理的插入方式。弱监督物体检测模块基于增广图像和伪边界框训练物体检测器，并将原始输入图像上的高置信度检测结果存入物体候选池。在训练过程中，为了避免检测器过拟合到错误的检测结果上，本文在原始弱监督物体检测器基础上增加了基于增广边界框的检测分支。新增的分支使用增广得到的伪边界框训练，原有基线弱监督物体检测器仍使用图像标签进行训练。图像增广过程与弱监督物体检测过程相互迭代，不断提升弱监督检测器的性能。综上所述，本文主要有以下贡献：

1）提出了基于物体布局后验概率的多物体图像增广方法。该方法无需额外的数据，仅利用原始的弱监督物体检测训练集和检出物体，在线地进行多物体图像增广，并通过对后验概率图的估计和采样来产生真实且合理的增广图像。

2）提出了ProMIS方法，该方法迭代地进行图像增广和弱监督物体检测训练。该方法使用增广图像和对应的伪边界框标签，提升了弱监督物体检测器的性能，而且其增广策略在不同弱监督检测器上具有通用性。

3）在Pascal VOC数据集上，将ProMIS方法应用到不同弱监督物体检测器上均取得了显著的性能提升，在基线算法基础上mAP提升最高可达7%。

1 相关工作

1.1 弱监督物体检测

弱监督物体检测采用图像类别标签训练物体检测器。在深度学习算法得到广泛应用后，弱监督物体检测器的性能也大大提升。弱监督物体检测研究包含的子方向较多，主要包括检测器结构（Bilen和 Vedaldi，2016；Tang 等，2017）、物体定位的改进（Li 等，2019；Shen 等，2019；Chen 等，2020）、优化方法（Wan 等，2019）、学习策略（Gokberk 等，2014；Zhang 等，2018）、骨干网络（Shen 等，2020）、伪标签生成方法（Kosugi 等，2019）等子方向和子问题。

本文主要研究物体定位的改进，在此对该子问题相关的工作进行介绍。Blien等人（Bilen 和 Vedaldi，2016）提出弱监督深度检测模型（WSDDN），该模型对物体候选框的得分进行聚合得到图像的分类结果，根据图像分类标签对该结果进行约束，实现检测器的训练。Tang等人（Tang 等，2017）提出了OICR方法，该方法在WSDDN基础上叠加多个细化分支，用上一个检测分支产生的伪标签框，指导下一个检测分支的训练。该方法简单有效，在后续的工作中常被作为基线方法使用。然而，由于缺少精准的物体定位监督信息，弱监督物体检测器还存在难以区分物体部件与整体、多个同类物体与单一个体的问题。SDCN（Li 等，2019）、WS-JDS（Shen 等，2019）、SLV（Chen 等，2020）等采用了分割图或热力图与检测分支交互的方法，使检测结果尽可能覆盖弱监督分割找到的类别相关区域，来保证检出物体的完整性。Ren等（Ren 等，2020）采用了随机遮挡的方式，遮挡部分检出物体区域，但依旧要求检测器在图片上检测出该物体，从而保证检出物体的完整性。现有方法主要针对如何保证检到物体的完整性进行了探索，对于如何区分单个物体和多个物体构成的区域，目前弱监督物体检测方法并没有很好的解决方案。

1.2 多物体图像增广

虽然近年来基于对抗生成网络的图像生成方法取得了巨大的成功，然而这些方法主要用于生成限定场景、特定风格的图像（如人脸，风景画等）。对于生成检测任务中复杂多变的多物体结构化图像，目前基于对抗生成网络的图像生成方法仍然难以很好地实现。因此，本文中主要研究将检出物体区域直接和输入图像进行融合，以产生较为真实、多样的多物体图像。现有多物体图像增广方法主要用于全监督领域，即物体既具备类别标记又具备完整的分割图（或3D模型）标注。Dwibedi等人提出了一种离线数据增广的方法（Dwibedi 等，2017），该方法将带有3D信息的物体插入到背景图像集上，构建了一个合成的室内实例分割数据集。用该合成数据集训练检测器并在真实数据上进行测试，获得了与使用真实数据训练的模型相近的性能。Fang等人（Fang 等，2019）提出可以通过移动输入图像上的物体对数据集进行增广，以增强实例分割算法性能。该方法将前景物体移动到原图中与它的原始位置周边背景具有高相似性的位置，并使用图像修复方法对原有位置进行填充，形成新的输入图像。Kisantal等人（Kisantal 等，2019）提出将小物体在原图上复制粘贴多次的方法，以提高小物体检测的性能。Ghiasi等人（Ghiasi 等，2021）提出采用图

像间物体复制粘贴、尺寸缩放等方式来得到较好的　　实

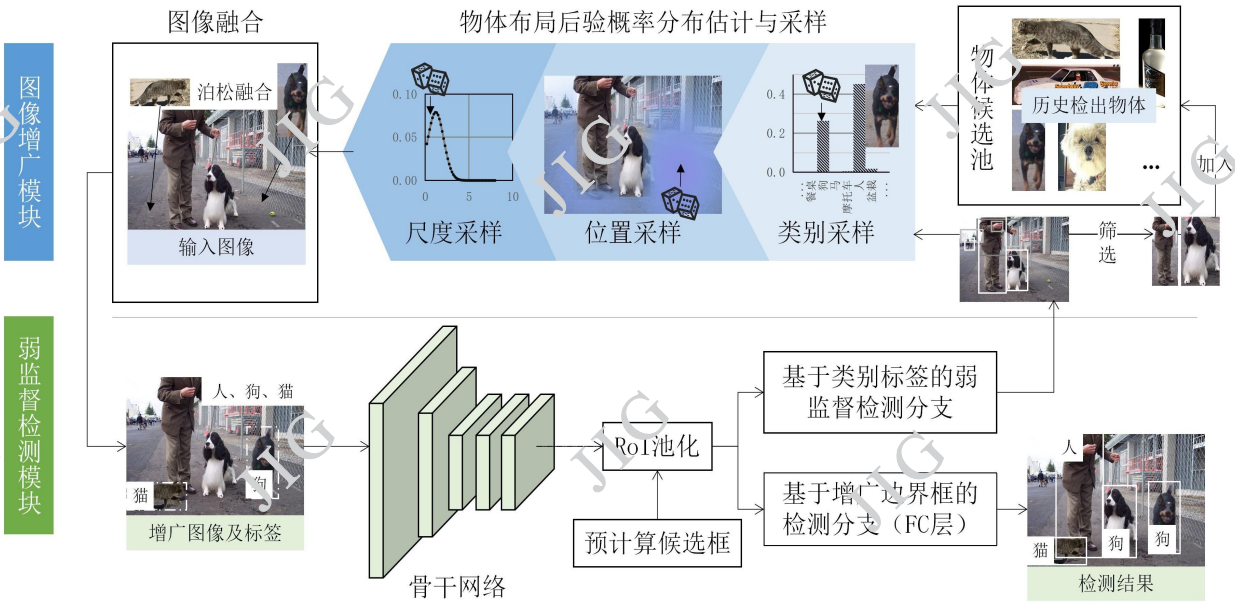


图 2　基于概率图采样增广的弱监督物体检测方法（ProMIS）框架图

Fig.2 The overview of weakly supervised object detection with probability-based multi-object image synthesis (ProMIS)

例分割数据增广效果。总的来说，多物体图像增广方法的有效性在全监督实例分割上已经得到了证明。但在缺乏准确定位信息的弱监督数据上如何生成有效的多物体增广图像尚无相关工作。

2　基于概率图采样增广的弱监督物体检测（ProMIS）

由于弱监督物体检测仅使用类别标签进行监督，缺乏对单个物体判别能力的约束，弱监督物体检测器只要检测出部分或整个类别相关区域即可满足分类损失较小的约束。由此导致在同一区域存在多个同类实例时容易将其检测为一个物体，或物体判别性强的部件被检测为整个物体。

为了辨别单一物体，直接的方法是构造含有单一物体、多物体、物体部件的图像，并标出物体的边界框作为检测器的监督信号，强制检测器学习它们之间的区别。然而弱监督物体检测问题没有提供边界框标注，一个可行的替代方案是将检测到的物体 $\boldsymbol{o}$ 插入到输入图像 $\boldsymbol{I}$ 上合成多物体图像 $\boldsymbol{I}'$，并记录该物体插入的位置构造出伪边界框 $\mathrm{box}_{\boldsymbol{o}}$。根据得到的带伪边界框的增广图像，指导检测器学习单一物体的区分性信息，来提升检测框定位的准确性。如图2所示，本文提出基于概率图采样增广的弱监督物体检测方法（ProMIS）。该方法主要包括图像增广和弱监督物体检测两个模块。图像增广模块将弱监督检测模块检出的高置信度物体（如“狗”、“人”）加入物体候选池，然后基于物体布局后验概率从候选池中采样出一个待插入物体及其在输入图像上的位置、尺度，进而通过融合算法将物体样本插入到输入图像上得到多物体增广的图像。弱监督物体检测模块使用增广后的类别标签和伪边界框标签训练弱监督物体检测器的两个并行检测分支。其中，图像增广模块主要解决已知输入图像 $\boldsymbol{I}$ 和一些弱监督检测器检出的高置信度物体 $O^{\boldsymbol{I}}=\{\boldsymbol{o}_j^{\boldsymbol{I}}\}$（式中，$j=1,\mathrm{L},N_d$，$N_d$ 为图像 $\boldsymbol{I}$ 上检出物体的个数），如何在输入图像上插入新物体的问题。弱监督物体检测模块主要解决已知增广图像和它对应的插入边界框 $\mathrm{box}_{\boldsymbol{o}}$，如何进行检测器训练的问题。下面将在2.1节和2.2节分别对上述两个模块进行详细介绍。

2.1　基于物体布局后验概率的图像增广模块

图像增广模块的目的是依据训练数据集，在输入图像某位置上、以某尺度、插入某类别的物体，以合成较为真实且符合数据分布的增广图像。该问题可形式化为：已知数据集上所有输入图像及其类别标签，根据一张输入图像 $\boldsymbol{I}$，估计新插入物体在输入图像 $\boldsymbol{I}$ 上的类别 c、插入位置 $\boldsymbol{p}$ 和尺度 s，进而依据它们合成图像。定义物体布局 $\boldsymbol{o}$ 为上述插入物体描述参数组成的向量 $\boldsymbol{o}\triangleq\{\boldsymbol{p},s,c\}$。给定一张输入图像，其物体布局 $\boldsymbol{o}$ 可能有多个合理的取值。

为保持这种多样性和随机性，本文提出将物体布局 $\boldsymbol{o}$ 的求解过程建模为“后验概率分布估计与采样”的过程，即已知输入图像 $\boldsymbol{I}$ 和其类别标签 $\boldsymbol{c}_{fg}$，输入图像上已有的物体满足分布 $P(\boldsymbol{o}|\boldsymbol{I},\boldsymbol{c}_{fg})$，利用检出物体可以估计物体布局 $\boldsymbol{o}$ 的后验概率分布 $P(\boldsymbol{o}|\boldsymbol{I},\boldsymbol{c}_{fg})$，并按照估计出的后验概率分布对 $\boldsymbol{o}$ 进行采样。值得注意的是，该方法不使用任何额外训练数据和神经网络模块，而是利用训练集上图像的检测结果不断地在线统计并估计后验概率 $P(\boldsymbol{o}|\boldsymbol{I},\boldsymbol{c}_{fg})$。

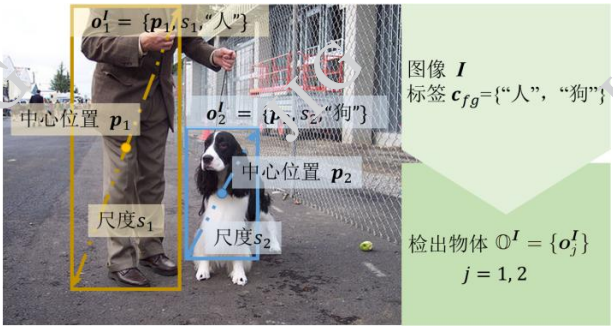


图 3 文中数学符号含义示意图

Fig.3 The notations in an example image

考虑到输入图像 $\boldsymbol{I}$、类别标签 $\boldsymbol{c}_{fg}$、以及物体布局 $\boldsymbol{o}$ 之间的关系难以直接描述，本文将输入图像 $\boldsymbol{I}$ 和类别标签 $\boldsymbol{c}_{fg}$ 抽象为图像上所有物体的集合，即输入图像 $\boldsymbol{I}$ 可用检出物体 $0^{\boldsymbol{I}} = \{\boldsymbol{o}_j^{\boldsymbol{I}}\}$ 进行描述，如图3所示。物体布局的集合可以通过弱监督物体检测器检出的物体进行近似估计。进而，欲估计的物体布局 $\boldsymbol{o}$ 后验概率分布 $P(\boldsymbol{o}|\boldsymbol{I},\boldsymbol{c}_{fg})$ 可写为 $P(\boldsymbol{o}|0^{\boldsymbol{I}})$，由贝叶斯公式进一步拆解为，

$$\begin{aligned} P(\boldsymbol{o}|0^{\boldsymbol{I}}) &= P(\boldsymbol{p},s,c|0^{\boldsymbol{I}}) \\ &= P(c|0^{\boldsymbol{I}})P(\boldsymbol{p},s|0^{\boldsymbol{I}},c) \\ &= P(c|0^{\boldsymbol{I}})P(\boldsymbol{p}|0^{\boldsymbol{I}},c)P(s|0^{\boldsymbol{I}},c,\boldsymbol{p}), \end{aligned} \quad (1)$$

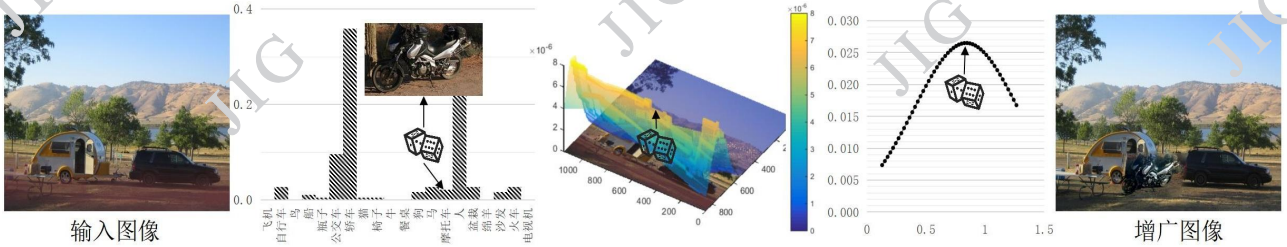
由式（1），物体布局后验概率 $P(\boldsymbol{o}|0^{\boldsymbol{I}})$ 的估计可转化为分别估计下面三种后验概率：

1）插入物体的类别后验概率 $P(c|0^{\boldsymbol{I}})$。该后验概率刻画了已知输入图像的物体布局集合 $0^{\boldsymbol{I}}$，在该图像上插入不同物体类别的合理程度。图4(a)通过举例的方式对该项的作用进行解释。已知图像已包含“轿车”，图像上出现不同类别物体的概率是不同的，如：可能出现“摩托”，但不太可能出现“电视”。这种可能性可以通过插入物体的类别后验概率 $P(c|0^{\boldsymbol{I}})$ 进行描述。

2）物体中心在图像上插入位置的后验概率 $P(\boldsymbol{p}|0^{\boldsymbol{I}},c)$。已知输入图像物体布局集合 $0^{\boldsymbol{I}}$ 和待插入物体的类别 c，该项描述了物体在不同位置出现的可能性。如图4(b)所示，已知输入图像“轿车”及其大概位置，欲插入“摩托”，摩托车在不同位置的概率也是不同的，如：通常摩托车在轿车的两侧可能性较高，而摩托车在汽车上方的可能性较低。

3）插入物体在图像上尺度的后验概率 $P(s|0^{\boldsymbol{I}},c,\boldsymbol{p})$。即已知输入图像物体布局集合 $0^{\boldsymbol{I}}$、待插入物体的类别 c 和物体中心在图像上的插入位置 $\boldsymbol{p}$，该项描述不同物体尺度对应的合理程度。如图4(c)所示，已知输入图像“轿车”，欲插入“摩托”，通常图像上摩托车的大小是轿车的0.5、1倍，而摩托车的大小是轿车的10或0.1倍是不常见的，这种“常识”可由插入物体尺度的后验概率 $P(s|0^{\boldsymbol{I}},c,\boldsymbol{p})$ 进行建模。

前三个小节将对上述三个后验概率的估计和采样方法进行具体介绍。具体地，本方法根据不同后验概率的特性用不同分布去描述它们，每一轮训练时根据统计到的历史检出物体在线地对分布进行估计。进而，插入新物体时，本方法根据估计出的后验概率采样出一组新的合理布局参数。图4举例展示了一组物体布局参数的采样过程。最后一个小节将介绍如何根据物体布局参数将物体插入到输入图像上，即图像融合方法。



(a) 类别后验概率 $P(c|\mathbb{O}^{\boldsymbol{I}})$　(b) 中心位置后验概率 $P(\boldsymbol{p}|\mathbb{O}^{\boldsymbol{I}},c)$　(c) 尺度后验概率 $P(s|\mathbb{O}^{\boldsymbol{I}},c,\boldsymbol{p})$

图 4　物体布局（类别、中心位置、尺度）后验概率采样过程示例（骰子和箭头标明了采样值）

Fig.4 Sampling from the posterior probabilities

2.1.1 插入物体的类别后验概率 $P(c|0^{I})$

已知当前图像中出现的前景类别标签 $\boldsymbol{c}_{fg}=\{c_j\}$，类别 c 物体出现的频率可以用来近似估计 c 类物体出现的后验概率 $P(c|0^{I})$。因而，本文统计频率直方图作为类别后验概率 $P(c|0^{I})$ 的估计。举例来说，图4(a)所示直方图展示了一组估计到的 $P(c|0^{I})$，输入图像中出现了“轿车”，它与“人”、“公交”、“摩托”等一起出现的频率相对较高，但与“餐桌”、“电视”一同出现的频率较低。该直方图比较符合人的认知。

具体地，为简化计算，本文假设已检出物体之间相互独立，类别后验概率 $P(c|0^{I})$ 可拆解如下：

$$P(c|0^{I})=\prod_{j=1}^{N_d}P(c|\boldsymbol{o}_j^{I})\approx\prod_{j=1}^{N_d}P(c|c_j),\quad(2)$$

式中，$\boldsymbol{o}_j^{I}$ 对应的类别记为 c_j。由式（2），只需估计后验概率 $P(c|c_j)$ 即可计算出 $P(c|0^{I})$。后验概率 $P(c|c_j)$ 可用出现频率进行近似，如下：

$$P(c|c_j)=\frac{M_{c,c_j}}{M_{c_j}},$$

式中，分子 M_{c,c_j} 为 c_j 类别出现时，c 类物体同时出现的频次；分母 M_{c_j} 为 c_j 类别出现的频次。

由此，给定一张输入图像 $\boldsymbol{I}$，在该图像上检出的物体为 $0^{I}=\{\boldsymbol{o}_j^{I}\}$（$j=1,\cdots,N_d$），则在该图像上插入类别 c 的概率 $P(c|0^{I})$ 可计算为：

$$P(c|0^{I})=\prod_{j=1}^{N_d}\frac{M_{c,c_j}}{M_{c_j}}.\quad(3)$$

式（3）计算出了图像 $\boldsymbol{I}$ 上插入物体类别的后验概率 $P(c|0^{I})$，该概率由训练集上历史检出物体统计得到。为使合成图像仍满足原始数据分布，新插入物体类别也应满足该后验概率分布：$P(c|0^{I})$。基于此，本文在该后验概率分布中采样出 N_s 个物体的类别。其中，为了简便起见，如果插入多个物体，本文不再考虑新插入物体之间的关系，而是直接独立地采样出了多个物体对应的超参数，后续2.1.2和2.1.3小节均采用了这种设置。对于每个类别，本文首先将每轮检出的高置信度物体（每个前景类别得分最高的一个物体）记录下来构成候选池，然后从候选池中随机选择一个该类别的实例，作为待插入物体实例。为了保证物体候选池中的实例具有较高的正确性，本文首先进行基线弱监督物体检测器的训练，再以此为初始值，加入图像增广模块进行训练。在图4(a)的例子中，图像中包含了“轿车”这一类别，依据式（3）计算出后验概率直方图，并依据该直方图采样，采到了“摩托”这一类别，在候选池中随机选出了一个“摩托”的实例作为待插入的物体。

2.1.2 物体中心在图像上插入位置的后验概率 $P(\boldsymbol{p}|0^{I},c)$

已知待插入物体，下面需要采样该物体中心在输入图像 $\boldsymbol{I}$ 上的插入位置，即估计位置后验概率分布 $P(\boldsymbol{p}|0^{I},c)$ 并进行采样。

本文希望新插入的物体不要影响图像上的原有前景物体，因此约束新插入物体尽量插入到图像的背景区域 B 上，即 $\boldsymbol{p}\in\mathrm{B}$。用 $P(\boldsymbol{p},\boldsymbol{p}\in\mathrm{B}|0^{I},c)$ 替代 $P(\boldsymbol{p}|0^{I},c)$，并进一步写为，

$$\begin{aligned}P(\boldsymbol{p},\boldsymbol{p}\in\mathrm{B}|0^{I},c)&=P(\boldsymbol{p}\in\mathrm{B}|0^{I},c)P(\boldsymbol{p}|0^{I},c,\boldsymbol{p}\in\mathrm{B})\\&=P(\boldsymbol{p}\in\mathrm{B}|0^{I})P(\boldsymbol{p}|0^{I},c,\boldsymbol{p}\in\mathrm{B}),\end{aligned}\quad(4)$$

考虑到输入图像背景区域 B 与插入物体所属的类别 c 无关，因此可以将 $P(\boldsymbol{p}\in\mathrm{B}|0^{I},c)$ 进一步化简为 $P(\boldsymbol{p}\in\mathrm{B}|0^{I})$。

式（4）中，第一项表示当前图像上位置 $\boldsymbol{p}$ 属于背景的概率 $P(\boldsymbol{p}\in\mathrm{B}|0^{I})$，该项可以用检测结果生成的热力图进行表示。本文采用如下方式生成热力图：检测器输出的第 i 个候选框对于类别 c 的打分为 $s_{i,c}$，如果类别 c 没有出现在该图像上（$c\notin\boldsymbol{c}_{fg}$），使 $s_{i,c}=0$；将上述修正后的候选框得分对应到图像像素上，进而生成热力图 $\boldsymbol{H}$ 如下，

$$\boldsymbol{H}(\boldsymbol{p},c)=\frac{1}{F_1}\sum_{i\ \mathrm{s.t.}\ \boldsymbol{p}\in\mathrm{box}_i}s_{i,c},$$

式中，F_1 是归一化因子，F_1 使 $\sum_c\boldsymbol{H}(\boldsymbol{p},c)=1$。$\boldsymbol{p}\in\mathrm{box}_i$ 表示第 i 个候选框覆盖了位置 $\boldsymbol{p}$。上述公式将检测框得分 $s_{i,c}$ 累计到对应像素 $\boldsymbol{p}$ 上并进行归一化，计算出该像素属于类别 c 的概率 $\boldsymbol{H}(\boldsymbol{p},c)$。由于前景类别得分较小的像素属于背景类别的概率较高，估计 $P(\boldsymbol{p}\in\mathrm{B}|0^{I})$ 如下，

$$P(\boldsymbol{p}\in\mathrm{B}|0^{I})=1-\max_{c\in\boldsymbol{c}_{fg}}\boldsymbol{H}(\boldsymbol{p},c).$$

该公式将该概率图 $\boldsymbol{H}$ 在所有前景类别 $\boldsymbol{c}_{fg}$ 上取最大值作为 $\boldsymbol{p}$ 属于前景的概率，1减去该概率即为 $\boldsymbol{p}$ 属于背景的概率。

式（4）中，第二项 $P(\boldsymbol{p}|0^{I},c,\boldsymbol{p}\in\mathrm{B})$ 表示物

体中心在图像上插入位置的后验概率。类似于式（2），后验概率 $P(\boldsymbol{p}|0^{I},c,\boldsymbol{p}\in B)$ 可以分解为 $\sum_{j=1}^{N_d}P(\boldsymbol{p}|\boldsymbol{o}_j^{I},c,\boldsymbol{p}\in B)$。那么，类似地，概率分布估计过程只需考虑两个类别之间的关系。两个物体间的相对位置变化范围通常较大，但仍存在一些规律，如：沙发上常坐着人，瓶子常放在桌子上。因此，采用二维（$x-y$）混合高斯分布建模这种弱相关性。随协方差矩阵变化，该二维混合高斯分布既可以近似均匀分布进而建模空间上低相关的两类物体；又可以近似锐利的多峰概率分布，进而建模空间上高相关性的两类物体。已知一个参考（检出）物体 $\boldsymbol{o}_j^{I}$，假设 c 类物体中心相对 $\boldsymbol{o}_j^{I}$ 的位置后验概率满足双峰的混合高斯分布。由于自然图像做左右翻转后仍然真实，两物体相对位置概率分布通常具有以参考（检出）物体为轴左右对称的特性。因此，本文进一步约束该混合高斯分布关于物体 $\boldsymbol{o}_j^{I}$ 左右对称，即：

$$P(\boldsymbol{p}|\boldsymbol{o}_j^{I},c,\boldsymbol{p}\in B):\frac{1}{F_2}\left(N(\boldsymbol{\mu}_{c|c_j}^{-},\ \Sigma_{c|c_j})+N(\boldsymbol{\mu}_{c|c_j}^{+},\ \Sigma_{c|c_j})\right). \quad (5)$$

上式中的分布描述了以 c_j 类别物体为原点，相对偏移量（$\hat{x}$，$\hat{y}$）变化时，c 类别物体中心存在的概率。式中 F_2 为归一化因子，$\boldsymbol{\mu}_{c|c_j}^{-}$ 和 $\boldsymbol{\mu}_{c|c_j}^{+}$ 为左右对称的两个高斯成分的均值，$\Sigma_{c|c_j}$ 为对应的协方差矩阵。根据对称性，上述均值可以写为，

$$\boldsymbol{\mu}_{c|c_j}^{-}=[-m_{c|c_j}^{\hat{x}},m_{c|c_j}^{\hat{y}}]^T,$$

$$\boldsymbol{\mu}_{c|c_j}^{+}=[+m_{c|c_j}^{\hat{x}},m_{c|c_j}^{\hat{y}}]^T,$$

$m_{c|c_j}^{\hat{x}}$ 和 $m_{c|c_j}^{\hat{y}}$ 分别为 c 类物体相对于参考 c_j 类物体的相对偏移量 $\hat{x}$ 和 $\hat{y}$ 的均值。其中，在 x 轴上统计 $m_{c|c_j}^{\hat{x}}$ 时只考虑距离的绝对值，不考虑方向，即 $m_{c|c_j}^{\hat{x}}>0$。

根据弱监督物体检测结果，以 c_j 为参考类别，在线地统计 c 类别对应的均值 $m_{c|c_j}^{\hat{x}}$，$m_{c|c_j}^{\hat{y}}$ 和协方差矩阵 $\Sigma_{c|c_j}$。具体地，当 c_j 和 c 两类物体同时出现在输入图像上时，以类别为 c_j 的检出物体 $\boldsymbol{o}_j^{I}$ 为参考，在线记录类别为 的检出物体 $\boldsymbol{o}^{I}$ 的相对偏移量，

$$\hat{x}=|x_{\boldsymbol{o}^{I}}-x_{\boldsymbol{o}_j^{I}}|/w_{\boldsymbol{o}_j^{I}},$$

$$\hat{y}=|y_{\boldsymbol{o}^{I}}-y_{\boldsymbol{o}_j^{I}}|/h_{\boldsymbol{o}_j^{I}}.$$

使用最近记录的200个 c 类别相对于 c_j 类别的偏移量 $\hat{x}$、$\hat{y}$ 估计均值 $m_{c|c_j}^{\hat{x}}$，$m_{c|c_j}^{\hat{y}}$ 和协方差矩阵 $\Sigma_{c|c_j}$。

由此，本文可以根据当前图像检出物体 0^{I}、待插入类别 c 和历史检出物体估计出在输入图像上位置的后验概率图 $P(\boldsymbol{p}|0^{I},c)$。对该概率图采样，即可得到待插入物体 $\boldsymbol{o}$ 中心所在的位置 $\boldsymbol{p}$。在图4(b)的例子中，三维曲面图为待插入实例“摩托”的中心在输入图像上的位置后验概率分布 $P(\boldsymbol{p}|0^{I},c)$，依据该分布进行采样，得到图示插入位置。可观察到图中高概率区域对应的位置较符合常理，即已知图上检出实例“轿车”，“摩托”与“轿车”同时出现时一般出现在同一水平面上，因而“轿车”的两侧应概率较高。

2.1.3 插入物体在图像上尺度的后验概率 $P(s|0^{I},c,\boldsymbol{p})$

待插入物体尺度的后验概率分布也可根据检出物体拆解为两类物体之间的尺度关系，类似式（2）尺度的后验概率可分解为 $\sum_{j=1}^{N_d}P(\hat{s}\times s_{\boldsymbol{o}_j^{I}}|\boldsymbol{o}_j^{I},c,\boldsymbol{p})$。这里 $\hat{s}$ 为同一幅图两类物体间的相对尺度 $\hat{s}=l_{\boldsymbol{o}}/l_{\boldsymbol{o}_j^{I}}$，式中 $l_{\boldsymbol{o}}$ 表示当前（待插入）物体 $\boldsymbol{o}$ 的对角线长度，$l_{\boldsymbol{o}_j^{I}}$ 表示参考（检出）物体 $\boldsymbol{o}_j^{I}$ 的对角线长度，$\hat{s}\times s_{\boldsymbol{o}_j^{I}}$ 即为物体 $\boldsymbol{o}$ 在输入图像上的绝对尺度 s。

由于两类物体的相对尺度关系比较固定，且它们的相对尺寸通常在某值附近取值概率较高，与该值相比相对尺寸值越大（或越小）取值的概率越低。因此，假设 c 类物体相对于 c_j 类物体的相对尺度 $\hat{s}$ 的后验概率分布满足高斯分布，

$$P(\hat{s}|\boldsymbol{o}_j^{I},c,\boldsymbol{p}):\ N(m_{c|c_j}^{\hat{s}},s_{c|c_j}^{2}). \quad (6)$$

根据弱监督物体检测器历史检出物体，可以统计该高斯分布中的样本，进而估计该高斯分布的参数。以 c_j 为参考类别，为 c 类别在线地统计一组高斯分布的参数 $m_{c|c_j}^{\hat{s}}$，$s_{c|c_j}^{2}$：当两类 c_j 和 c 物体同时在输入图像上检出时，在线记录以 c_j 类物体为参考 c 类物体的相对尺度 $\hat{s}$；使用最近记录的200个相对尺度样本估计均值 $m_{c|c_j}^{\hat{s}}$ 和方差 $s_{c|c_j}^{2}$。

根据上述公式可得到插入物体的尺度后验概率 $P(s|0^{I},c,\boldsymbol{p})$，根据该分布采样即可得到待插入物体的尺度 s。在图4(c)的例子中，图示高斯分布为待插入物体“摩托”相对于图上检出物体“轿车”的相对尺度后验概率分布 $P(\hat{s}|\boldsymbol{o}_j^{I},c,\boldsymbol{p})$。依据该分布求解绝对尺度概率分布，再进行采样，采到图示待插入物体的尺度。可观察到图中相对尺度分布基本符合常识中“摩托”对于“轿车”的尺寸

关系。

2.1.4 图像融合

综上，根据式（3）、（5）、（6）可统计历史检出物体样本，并根据当前给定图像 $\boldsymbol{I}$ 上的弱监督物体检测器检出的物体 $0^{\prime}$，估计待插入物体的类别、位置、尺度的后验概率分布 $P(c|0^{\prime})$、$P(\boldsymbol{p}|0^{\prime},c)$、$P(s|0^{\prime},c,\boldsymbol{p})$。进而，基于估计出的概率模型采样得到一组待插入物体的类别、位置、尺度的参数。具体而言，根据给定的图像 $\boldsymbol{I}$，首先采样出物体类别 c_o，并从该类别候选池中随机采样出该类别的一个实例作为待插入物体，该物体对应的图像区域为 $\boldsymbol{I}_o$。然后，采样出在图像 $\boldsymbol{I}$ 上的插入位置 $\boldsymbol{p}_o$ 和尺度 s_o。上述过程只依据当前图像和历史图像上的检出物体，是一种概率模型估计与采样的过程。

接下来，考虑如何将待插入物体区域 $\boldsymbol{I}_o$ 融合到输入图像 $\boldsymbol{I}$ 上合成最终的多物体增广图像 $\boldsymbol{I}^{*}$。对物体和图片进行融合有多种已有方法，其中较为简单、常用的包括：直接覆盖、高斯模糊、泊松融合和图像叠加。

本文对上述四种融合方法进行了实验，发现泊松融合能够在保持物体轮廓信息和图像真实性之间取得较好的平衡，在本文的实验设定下取得了最优的效果。泊松融合要求在边界处图像梯度变化最小，较好地缓解了融合算法边界过渡不自然的现象。在Pascal VOC 2007数据集上，采用OICR方法做基线算法，泊松融合的检测mAP较直接覆盖、高斯模糊、图像叠加和随机任意融合方法相比，分别提升2.3%、2.2%、0.1%和1.0%。因此，本文采用泊松融合（Perez 等，2003）对物体和图像 $\boldsymbol{I}$ 进行融合得到最终的增广图像。

值得注意的是，这一结论与全监督任务中有所不同。在基于多物体图像增广的全监督方法中（Dwibedi 等，2017；Kisantal 等，2019；Ghiasi 等，2021），使用直接覆盖或者多种融合方式随机的方法效果要好于泊松融合的效果。导致这一差异的主要原因是：上述全监督方法中，直接给出了精确的插入物体的像素级分割图，只插入分割出的物体区域，直接覆盖或者随机融合可以尽可能地保持插入物体轮廓信息。而弱监督物体检测任务，使用物体外边界框得到的物体实例包含一定的背景信息，如果直接覆盖会导致融合边界处有较大的表观差异，进而导致检测器去拟合这些人为添加的差异，而不是真正的物体特征。而泊松融合能够在融合边界处过渡地较为自然，同时尽可能保留插入物体的表观特征，较好地平衡保留物体轮廓和保持增广图像真实性两个问题。

2.2 弱监督物体检测器训练

通常，弱监督物体检测器只使用输入图像及其对应的类别标签 $\boldsymbol{c}_{fg}$ 进行训练。如果对一幅输入图像进行增广，在该图像上插入物体 c_o，输入图像的标签可以扩充为 $(\boldsymbol{c}_{fg},c_o,\mathrm{box}_o)$，即：1）原有的类别标签 $\boldsymbol{c}_{fg}$，2）新增物体类别 c_o 和3）新增物体伪边界框标签 box_o。

使用扩充的标签进行训练，一种简单的方案是直接使用增广图像和增广边界框标签作为强监督信号训练原始检测分支。但是同一检测支路既产生伪边界框标签又使用该伪标签进行训练等同于指导检测器不断拟合之前检出的物体。这种情况下，检测器非常容易过拟合，进而导致性能下降。因此，本文在现有弱监督物体检测器基础上对检测结构进行微调，加入一个并行的检测分支，即基于增广边界框的检测分支，用于学习增广得到的伪边界框标签 box_o，如图2所示。

下面对每个分支的监督情况进行详细的介绍。如前文所述，基于类别标签的弱监督检测分支，即基线检测器的检测头部，通常包括两种检测子分支：一种是多示例检测子分支，它将候选框的分类结果利用注意力的形式聚合成图像分类结果，然后从分类标签中学习检测知识；另外一种是实例分类细化子分支，它依据多示例检测子分支（或上级实例分类细化子分支）的结果产生伪边界框标签，并采用全监督的方式进行训练。而基于增广边界框的检测分支，在模型结构和损失函数形式上与实例分类细化子分支是完全一样的，它们的不同点仅仅是伪边界框标签来源不同。具体地，原始弱监督物体检测分支使用类别标签 $(\boldsymbol{c}_{fg},c_o)$ 进行训练，如果该分支包含实例分类细化子分支，它的伪边界框标签生成方式均与基线弱监督物体检测算法保持一致，即该伪标签生成过程仅考虑了类别标签 $(\boldsymbol{c}_{fg},c_o)$，而完全忽略增广产生的物体伪边界框标签。基于增广边界框的检测分支采用两种伪边界框组成的集合 $(\mathrm{box}_o,\mathrm{box}_{oj})$ 进行训练。其中，box_o 为新增物体伪边界框标签，box_{oj} 为原始弱监督物体检测分支产生的关于类别 $\boldsymbol{c}_{fg}$ 的伪边界框标签。测试时，不同于OICR方法需要对多个检测实例分类细化子分支进行平均得到检测结果，本文直接采用基于增广

边界框检测分支的输出作为最终的检测结果。

3 实验

上述章节详述了本文提出的ProMIS方法。下面，本文将测试算法中各部分的作用，并和现有算法进行性能对比。

3.1 实验设置

数据集 本文在两个数据集上进行弱监督物体检测的验证，即PASCAL VOC 2007（Everingham 等，2010）和PASCAL VOC 2012（Everingham 等，2015）。PASCAL VOC 2007数据集包括9, 963幅图像、24, 640个物体和20个类别。该数据集划分为5, 011幅图像组成的训练验证集和另外4, 952幅图像构成的测试集。PASCAL VOC 2012数据集规模更大、难度更高，它也划分为两个子集：训练验证集包括11, 540幅图像和27, 450个物体，测试集包括10, 991幅图像。实验中，本文使用训练验证集进行训练，主要评测检测器的两个性能指标：1）在训练验证集上的正确定位指标（CorLoc）（Deselaers 等，2012），2）在测试集上的平均精度（AP和mAP）。消融实验部分在PASCAL VOC 2007数据集上进行，和现有方法的对比在多个数据集上均进行了实验。图像增广方面，本文未采用任何额外数据集学习增广模块，图像增广过程均在训练的同时在线进行，后验概率也在每轮训练时根据最新的200个历史样本进行更新。

实现细节 本文使用两块NVIDIA RTX 2080Ti GPU在mmdetection（Chen 等，2019）检测平台上进行训练和测试。本文方法均先训练好的基线弱监督检测器，再以此作为预训练权重，进行图像增广和检测训练。训练超参数方面，本文采用随机梯度下降（SGD）学习器（ ，1998）以10^{-3}的学习率训练40k轮，再以10^{-4}的学习率训练30k轮。伪标签生成和尺度增广方面，采用与OICR（Tang 等，2017）一致的方式，即只使用前景类别置信度最高的框作为伪边界框标签。新增的基于增广边界框的检测分支随机初始化，然后使用增广图像和增广得到的伪边界框标签进行训练。消融实验部分在OICR-VGG结构（Tang 等，2017；Simonyan 和 Zisserman，2014）上进行验证，和现有算法的对比在多种弱监督物体检测器上均进行了实验。对于实验中涉及的弱监督物体检测器OICR-VGG（Tang 等，2017）、SDCN-VGG（Li 等，2019）和OICR-DRN-ResNet50（Shen 等，2020；He 等，2016），如没有额外说明均采用和原文一致的设置。对于图像增广模块，每张训练图像上的每个前景类别，只选置信度最高的一个检出物体加入物体候选池。这些物体在候选池中生存200轮后，就会被删除，以避免训练前期不准确的检出物体对后续训练带来不良影响。

3.2 消融实验

3.2.1 设置基于增广边界框的检测分支的必要性

基于OICR结构，对基于增广边界框的检测分支进行实验验证。如表2所示，如果不添加基于增广边界框的检测分支，而是直接使用增广图像和标签训练原始弱监督物体检测分支，检测指标mAP仅有39.1%，较原始OICR方法反而会下降2.1%，表明增广图像中的噪声样本对检测器产生了不良的影响，使检测器过拟合，导致了检测性能下降。在原始OICR方法上增加基于增广边界框的检测分支并进一步采用增广方法进行训练，检测器性能较OICR有明显的提升（+5.2%），mAP高达46.4%，表明过上述拟合现象得到了极大缓解。

为了验证上述提升的主要原因不是网络参数增加，本文增加基于增广边界框的检测分支的同时减少OICR细化分支的数目到2个，这种结构与原始OICR结构参数量完全一致。如表2所示，使用图像增广方法进行训练后，mAP可达45.7%，与具有3个细化分支的OICR比仍然有高达4.5%的提升，且与在原始OICR上加入基于增广边界框的检测分支的mAP性能可比。

由于并不是所有弱监督检测网络都具有如OICR一样的多个细化分支，能够直接改造出一个基于增广边界框的检测分支。后文中在测试其他方法时为了不失通用性均直接加入一个额外的分支。

表 2 设置基于增广边界框的检测分支的必要性

Table 2 The effect of the detection branch based on the augmented bounding boxes

图像增广	基于增广边界框的检测分支	OICR 细化分支数	mAP（%）
		3	41.2（OICR）
✓		3	39.1
✓	✓	3	**46.4（本文）**
✓	✓	2	45.7

注：加粗字体为该列最优值。

3.2.2 物体布局后验概率分布估计和采样的有效性

为了证明物体布局概率分布解耦、估计和采样过程的有效性，本文对上述位置、类别、尺度后验

概率的建模方法分别进行了验证，如表3所示。该表中“随机”表示在上文提到的加入基于增广边界框的模型基础上，插入物体位置、类别、尺度都在均匀分布中随机采样得到。由该表可以看出，与完全随机采样相比，本文提出的位置、类别、尺度后验概率的建模方式均可提升最终检测性能的效果，并且它们共同作用效果最好，检测指标mAP可以在基线增广算法的基础上提升2.2%。本文认为，物体布局后验概率分布估计和采样对物体之间的相关性、物体空间相对位置等知识进行了概率化的建模。通过这些知识约束使插入物体布局采样过程更接近数据集上真实物体的布局，从而帮助弱监督检测器更好地学习了真实场景中物体之间的相对位置信息，进而提升了检测性能。

表 3 物体布局后验概率分布估计与采样的合理性

Table 3 The effects of estimating and sampling the posterior probabilities

类别 $P(c\|0^I)$	位置 $P(\boldsymbol{p}\|0^I,c)$	尺度 $P(s\|0^I,c,\boldsymbol{p})$	mAP（%）
–	--	--	41.2（OICR）
			44.2（随机）
✓			44.4
✓	✓		45.7
✓	✓	✓	**46.4（本文）**

注：加粗字体为该列最优值。

3.2.3 多个插入物体之间的关系

新插入物体与输入图像上原有物体之间的关系已通过 $P(\boldsymbol{p}|0^I,c)$ 表示。现考虑增广后图像上多个插入物体之间的关系对检测结果的影响。本小节希望通过控制变量的方式，观察多个物体怎样插入增广图像对检测器的训练更有利。主要考虑两个因素：新插入物体的个数和新插入物体之间的最大重合程度。其中，重合度定义为先插入物体被后插入物体遮挡的百分比。实验中，只约束重合度的最大值（设为0，0.3，1），最小值均不做限制。注意，在本文提出的增广算法中这两个参数均可通过在均匀分布中采样产生，不属于算法的超参数。

弱监督物体检测性能随插入物体个数和最大重合度变化的情况如图5所示，这两个因素是互相耦合在一起的，因为它们对图上物体的遮挡程度和物体空间关系的多样性都会造成影响，进而影响物体检测器的性能。

具体地，固定物体不重合，添加不同个数的物体，随着物体个数从1增加到5，检测性能增加。物体个数为5时，mAP达到最高46.4%。但同一幅图像新增物体个数过多达到7时时，对原始输入图像的遮挡过于严重，导致检测性能下降。添加物体个数较少时（如3个），检测器的性能随着最大重合程度增大（至0.3）先增大至46.1%。重合程度过大时，插入物体之间允许完全遮挡。部分增广图像中，后插入物体遮挡了先插入物体并使之不可见，然而这时产生的伪边界框标签却没有改变，因此检测性能反而随之下降。添加物体个数较多时（7个），添加物体总会互相遮挡或挡住原图像上的物体，从而导致伪边界框标签错误。这种情况下，允许新插入物体之间有较大的重合度反而一定程度地避免了挡住原图上的物体而导致的图像类别标签变化，检测器的性能随着最大重合程度而增大。

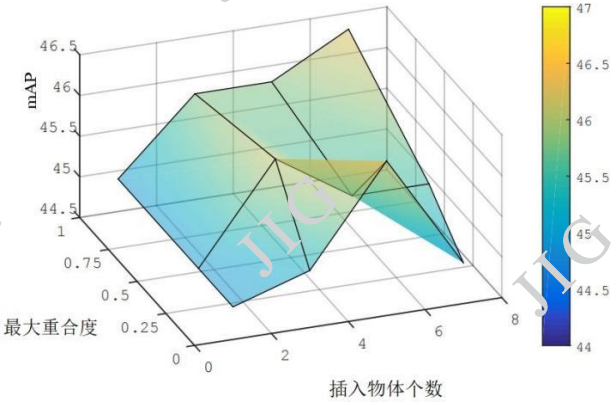


图 5 插入物体个数和最大重合度对弱监督物体检测性能的影响

Fig.5 The influence of the number of inserted objects and the overlapping ratio tolerance on the detection performance (mAP)

综上所述，在保证较少遮挡（原有或插入）物体的前提下，物体个数插入较多对检测器的训练最有利。物体之间遮挡无法避免时，就需要平衡遮挡和插入物体个数才能获得较好的效果。

3.3 与现有方法对比

上文验证了基于概率图采样增广的弱监督物体检测方法（ProMIS）中各模块的作用。下面将本文提出的ProMIS方法应用到现有方法上，并在Pascal VOC 2007和Pascal VOC 2012数据集上进行性能对比，结果见表4、表5、表6。本文在一些代表性方法的基础上应用了ProMIS方法，以观察ProMIS方法的有效性。这些基线算法包括：广为应用的OICR方法，针对定位问题的SDCN方法，和采用更好骨干网络的OICR-DRN方法。值得注意的是，ProMIS算法可插入到其他任意弱监督检测算法中，

并不仅限于这里提到的基线算法。由表4、表5、表

表 4 Pascal VOC 2007 测试集上本文和已有方法的 mAP（%）性能对比

Table 4 Comparisons with the state-of-the-art on the Pascal VOC 2007 test set

方法	骨干网络	飞机	单车	鸟	船	瓶	公交	轿车	猫	椅	牛	桌	狗	马	摩托	人	盆栽	绵羊	沙发	火车	电视	mAP
WSDDN（Bilen 和 Vedaldi，2016）	VGG16	ss	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
MELM（Wan 等，2019）	VGG16	55.6	66.9	34.2	29.1	16.4	68.8	68.1	43.0	25.0	65.6	45.3	53.2	49.6	68.6	2.0	25.4	52.5	**56.8**	62.1	57.1	47.3
ZLDN（Zhang 等，2018）	VGG16	55.4	68.5	50.1	16.8	20.8	62.7	66.8	56.5	2.1	57.8	47.5	40.1	**69.7**	68.2	21.6	27.2	53.4	56.1	52.5	58.2	47.6
GAL-fWSD512（Shen 等，2018）	VGG16	58.4	63.8	45.8	24.0	22.7	67.7	65.7	58.9	15.0	58.1	47.0	53.7	23.8	64.3	**36.2**	22.3	46.7	50.3	**70.8**	55.1	47.5
（Wei 等，2018）	VGG16	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
WSRPN（Tang 等，2018）	VGG16	57.9	**70.5**	37.8	5.7	21.0	66.1	**69.2**	59.4	3.4	57.1	**57.3**	35.2	64.2	68.6	32.8	**28.6**	50.8	49.5	41.1	30.0	45.3
ML-LocNet（Zhang 等，2018）	VGG16	59.3	68.9	45.7	29.0	24.5	64.8	68.1	59.3	18.6	49.1	50.2	43.1	65.8	**70.2**	19.9	24.3	48.1	54.2	62.8	41.8	48.4
WS-JDS（Shen 等，2019）	VGG16	52.0	64.5	45.5	26.7	27.9	60.5	47.8	59.7	13.0	50.4	46.4	56.3	49.6	70.7	25.4	28.2	50.0	51.4	66.5	29.7	45.6
CMIL（Wan 等，2019）	VGG16	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
OICR W-RPN（Kumar 和 Lee，2019）	VGG16	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	46.9
Kosugi et al.（Kosugi 等，2019）	VGG16	61.5	64.8	43.7	26.4	17.1	67.4	62.4	67.8	25.4	51.0	33.7	47.6	51.2	65.2	19.3	24.4	44.6	54.1	65.6	59.5	47.6
CSC C5（Shen 等，2020）	VGG16	51.4	62.0	35.2	18.7	27.9	66.7	53.5	51.4	16.2	43.6	43.0	46.7	20.0	58.4	31.1	23.8	43.6	48.8	65.4	53.5	43.0
OICR（Tang 等，2017）	VGG16	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
OICR-ProMIS	VGG16	**64.7**	60.7	48.3	28.3	23.7	67.3	67.2	33.7	25.9	61.7	43.3	30.0	46.2	69.6	6.9	25.4	**56.9**	50.7	51.5	65.5	46.4
SDCN（Li 等，2019）	VGG16	59.8	67.1	32.0	**34.7**	22.8	67.1	67.8	67.9	22.5	48.9	47.8	60.5	51.7	65.2	11.8	20.6	42.1	54.7	60.8	64.3	48.3
SDCN-ProMIS	VGG16	54.9	70.0	48.7	22.0	26.4	70.1	66.6	71.0	**28.0**	59.2	40.0	56.3	52.1	67.3	19.7	27.8	49.0	53.2	59.9	**66.9**	50.5
OICR-DRN（Shen 等，2020）	ResNet50	61.2	50.9	55.0	33.2	**36.2**	68.6	65.7	**79.2**	17.3	58.1	19.3	69.1	65.7	64.8	15.1	18.9	50.1	55.1	69.8	64.4	50.9
OICR-DRN-ProMIS	ResNet50	61.7	52.9	**59.7**	31.2	27.4	**73.7**	68.4	73.0	21.5	**68.6**	20.2	**72.4**	58.4	65.0	19.2	22.6	55.4	48.7	61.9	**66.9**	**52.1**

注：加粗字体为该列最优值。

表 5 Pascal VOC 2007 训练验证集上本文和已有方法的 CorLoc（%）性能对比

Table 5 Comparisons with the state-of-the-art on the Pascal VOC 2007 trainval set

方法	骨干网络	飞机	单车	鸟	船	瓶	公交	轿车	猫	椅	牛	桌	狗	马	摩托	人	盆栽	绵羊	沙发	火车	电视	CorLoc
WSDDN（Bilen 和 Vedaldi，2016）	VGG16	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
MELM（Wan 等，2019）	VGG16	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	61.4
ZLDN（Zhang 等，2018）	VGG16	74.0	77.8	65.2	37.0	46.7	75.8	83.7	58.8	17.5	73.1	49.0	51.3	76.7	87.4	30.6	47.8	75.0	62.5	64.8	68.8	61.2
GAL-fWSD512（Shen 等，2018）	VGG16	78.6	81.9	63.6	40.3	48.8	80.7	85.3	76.3	30.3	78.0	54.5	65.3	48.4	86.5	56.3	46.9	76.0	68.1	**83.9**	73.1	66.1
（Wei 等，2018）	VGG16	84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	42.3	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	61.0
WSRPN（Tang 等，2018）	VGG16	77.5	81.2	55.3	19.7	44.3	80.2	86.6	69.5	10.1	**87.7**	**68.4**	52.1	**84.4**	**91.6**	**57.4**	**63.4**	77.3	58.1	57.0	53.8	63.8
ML-LocNet（Zhang 等，2018）	VGG16	78.6	82.3	68.2	42.0	53.3	78.5	**88.5**	70.3	36.4	70.2	60.5	58.0	80.5	88.2	38.8	59.2	75.0	69.0	78.2	64.5	67.0
WS-JDS（Shen 等，2019）	VGG16	82.9	74.0	73.4	47.1	**60.9**	80.4	77.5	78.8	18.6	70.0	56.7	67.0	64.5	84.6	47.0	50.1	71.9	57.6	83.3	43.5	64.5
CMIL（Wan 等，2019）	VGG16	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	65.0
OICR W-RPN（Kumar 和 Lee，2019）	VGG16	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	66.5
Kosugi et al.（Kosugi 等，2019）	VGG16	85.5	79.6	68.1	55.1	33.6	83.5	83.1	78.5	42.7	79.8	37.8	61.5	74.4	88.6	32.6	55.7	77.9	63.7	78.4	74.1	66.7
CSC C5（Shen 等，2020）	VGG16	76.1	75.3	61.8	42.0	54.1	74.7	78.8	67.4	32.8	73.1	46.5	59.9	37.6	78.0	56.0	42.5	71.9	67.3	82.4	65.6	62.2
OICR（Tang 等，2017）	VGG16	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
OICR-ProMIS	VGG16	**89.3**	72.2	65.5	51.6	45.4	79.7	86.5	44.5	37.9	83.6	45.6	41.2	74.8	89.2	8.4	59.7	83.5	58.9	65.4	81.7	63.3
SDCN（Li 等，2019）	VGG16	85.8	83.1	56.2	**58.5**	44.7	80.2	85.0	77.9	29.6	78.8	53.6	74.2	73.1	88.4	18.2	57.5	74.2	60.8	76.1	79.2	66.8
SDCN-ProMIS	VGG16	80.4	**84.7**	72.4	49.5	48.1	**82.8**	85.0	83.7	**44.2**	79.5	47.5	72.1	71.1	83.2	42.3	54.9	70.1	57.5	75.3	83.2	**68.7**
OICR-DRN（Shen 等，2020）	ResNet50	75.9	65.6	70.9	56.9	50.0	81.6	86.8	**83.8**	33.0	79.5	27.3	79.9	81.7	81.0	30.4	45.0	**85.5**	**72.2**	79.1	81.2	67.4

方法	骨干网络	飞机	单车	鸟	船	瓶	公交	轿车	猫	椅	牛	桌	狗	马	摩托	人	盆栽	绵羊	沙发	火车	电视	CorLoc
OICR-DRN-ProMIS	ResNet50	85.0	68.1	**75.3**	53.1	44.7	83.2	88.0	78.9	37.1	85.2	26.3	**80.1**	81.6	87.6	32.4	46.2	84.2	65.6	74.0	**83.5**	68.0

注：加粗字体为该列最优值。

表 6 Pascal VOC 2012 数据集上本文和已有方法的 mAP（%）、CorLoc（%）性能对比

Table 6 Comparisons with the state-of-the-art on the Pascal VOC 2012 dataset

方法	骨干网络	mAP	CorLoc
MELM（Wan 等，2019）	VGG16	42.4	--
ZLDN（Zhang 等，2018）	VGG16	42.9	61.5
GAL-fWSD300（Shen 等，2018）	VGG16	43.1	67.2
（Wei 等，2018）	VGG16	40.0	64.4
WSRPN（Tang 等，2018）	VGG16	40.8	64.9
ML-LocNet（Zhang 等，2018）	VGG16	42.2	66.3
WS-JDS（Shen 等，2019）	VGG16	39.1	63.5
CMIL（Wan 等，2019）	VGG16	46.7	67.4
OICR W-RPN（Kumar 和 Lee，2019）	VGG16	43.2	67.5
Kosugi et al.（Kosugi 等，2019）	VGG16	43.4	66.7
CSC（Shen 等，2020）	VGG16	37.1	61.4
OICR（Tang 等，2017）	VGG16	37.9	62.1
OICR-ProMIS	VGG16	42.3	64.3
SDCN（Li 等，2019）	VGG16	43.5	67.9
SDCN-ProMIS	VGG16	50.5	**72.6**
OICR-DRN(Shen 等,2020)	ResNet50	49.7	71.2
OICR-DRN-ProMIS	ResNet50	**50.9**	72.5

注：加粗字体为该列最优值

6可见，在Pascal VOC 2007、Pascal VOC 2012数据集上分别对3种代表性检测器使用本方法，mAP平均获得了2.9%、4.2%的提升。对OICR、SDCN和OICR-DRN方法分别应用本文的增广训练方法，在Pascal VOC 2007数据集上，mAP指标分别比基线算法提升5.2%、2.2%和1.2%，CorLoc指标分别比基线算法提升2.7%、1.9%和0.6%。在Pascal VOC 2012数据集上，ProMIS方法在OICR、SDCN和OICR-DRN方法的基础上，mAP指标分别提升4.4%、7%和1.2%，CorLoc指标分别比基线算法提升2.7%、4.7%和1.3%。上述实验结果，证明了本文提出的ProMIS方法在多种基线算法上的通用性和有效性。

从类别上来看，本文提出的多物体图像增广方法在大部分类别上均有所提升，以VGG16为骨干网络在Pascal VOC 2007数据集上的结果为例，其中OICR-ProMIS和SDCN-ProMIS提升较为明显的是：对于“鸟”类分别获得了+17.2%、+16.7%的提升，“牛”类分别获得了+17.0%、+10.3%的提升，“绵羊”类分别获得了+15.2%、+6.9%的提升，等等。这些类别都是多个同类物体聚集现象较为常见的类别。由此可见，ProMIS有助于提升同一类别多个实例之间的可区分性。此外，OICR-ProMIS对于“人”、SDCN-ProMIS对于“船”的性能有所下降，主要原因是对应的弱监督检测算法对这些类别的检测有明显且一致的偏差，如集中在物体关键部件上。如图6所示，OICR检出的“人”大部分集中在脸部区域，使用这些带偏差的样本产生增广图像后，增广

图 6 OICR 检出的“人”边界框

Fig.6 The detected “person” bounding box from the OICR detector

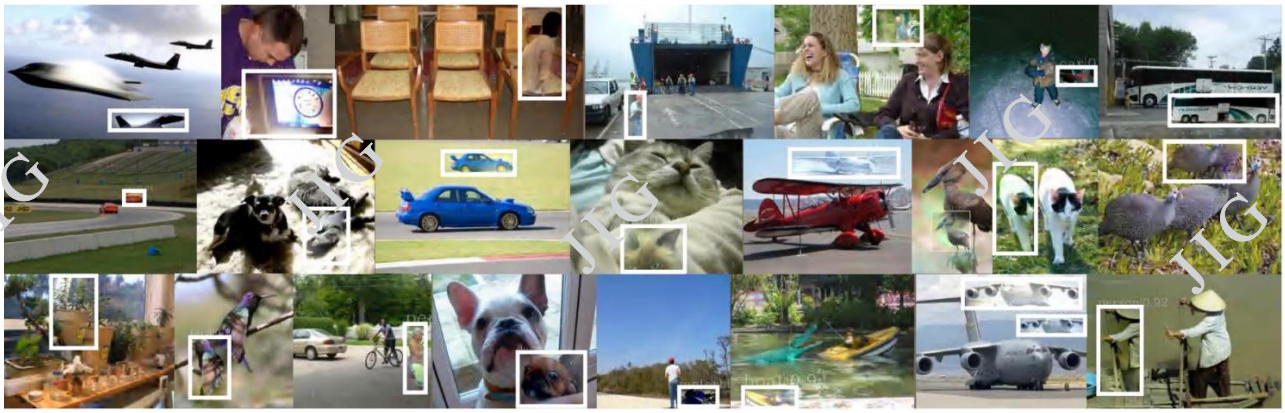
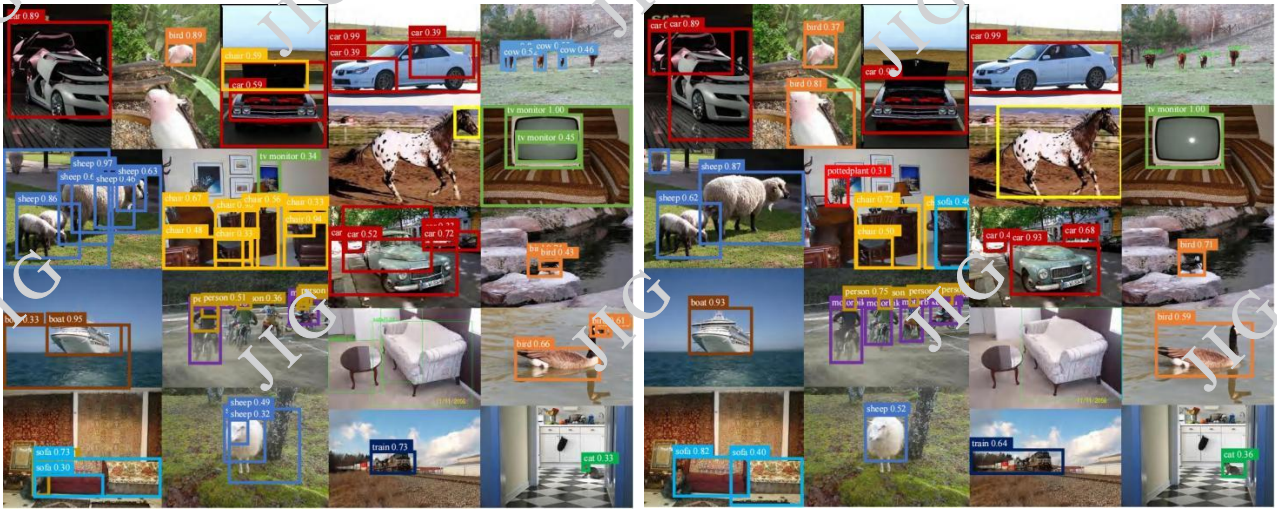


图 7 增广图像及插入物体的伪边界框标签（白色框）

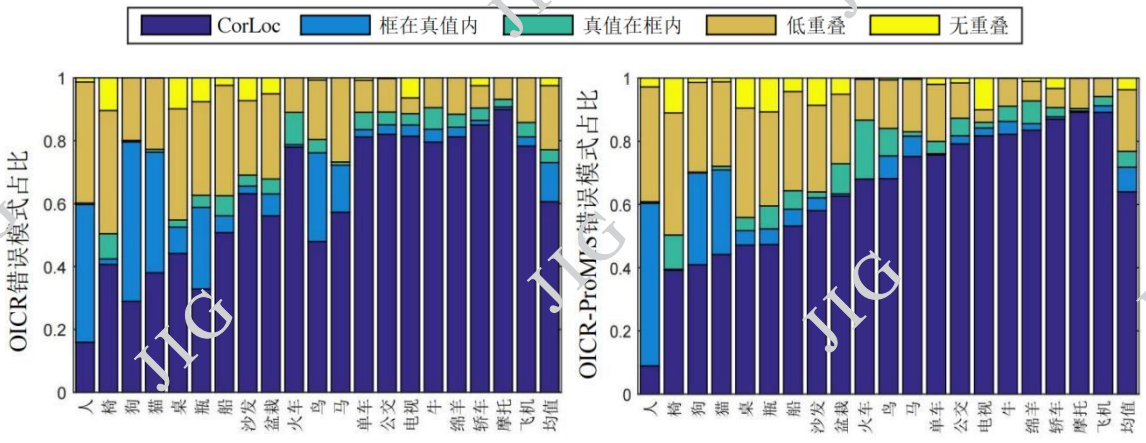
Fig.7 The augmented images and the pseudo-labels of the inserted objects



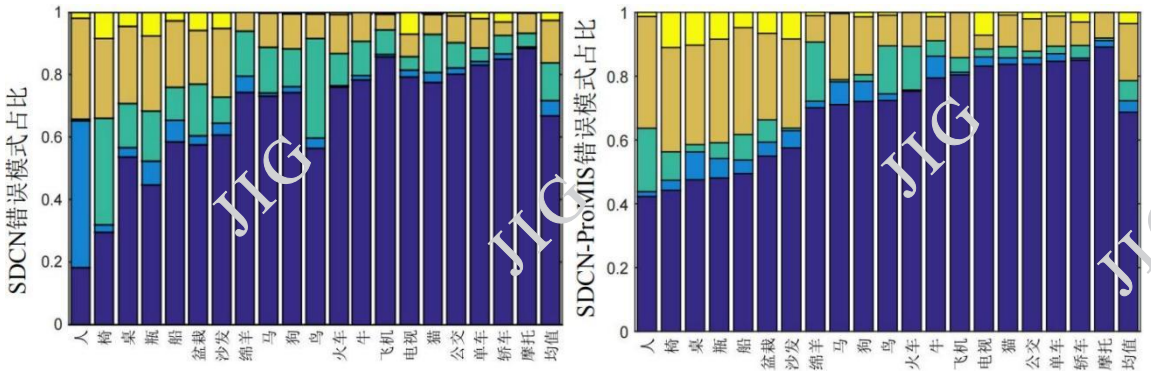
(a) OICR (b) OICR-ProMIS

图 8 (a) OICR 方法和本文提出的(b) OICR-ProMIS 方法检测结果的可视化

Fig.8 The visualization of the detection results of the (a) OICR and (b) OICR-ProMIS



(a) OICR (b) OICR-ProMIS



(c) SDCN　　(d) SDCN-ProMIs

图 9 (a) OICR、(b) OICR-ProMIS、(c) SDCN、(d) SDCN-ProMIS 错误模式占比分析

Fig.9 The error mode analysis of the (a) OICR, (b) OICR-ProMIS, (c) SDCN, and (d) SDCN-ProMIS

伪边界框标签大部分带有同样的错误模式，因此反而加深了对错误定位的拟合，导致了在对应类别上的性能下降。

本文对产生的增广样本和检测结果进行了可视化，如图7和图8所示。由图7，可以观察到使用物体布局后验概率采样后，生成的增广图像真实性和合理性大大提升。使用这些增广图像训练检测器，产生的检测结果对于单个物体实例的判别能力也有所增强，多个物体混淆为一个物体和物体部件视为整个物体的情况都相对减少，如图8所示。

3.4　错误模式分析

本文对OICR、OICR-ProMIS、SDCN和SDCN-ProMIS检测器的错误模式及占比情况进行了可视化，结果如图9所示。图中，检测到的框被分为5类：1）正确定位（CorLoc），即检测框和真值的交并比大于等于0.5，2）框在真值内，即检测框完全落在真值框内部；3）真值在框内，即真值框完全落在检测框内部；4）重叠低，即以上情况均不是，但检测框和真值框交并比不为零；5）无重叠，即检测框和真值框交并比为零。

对于OICR检测器来说，它的错误模式主要为框在真值内如图9(a)所示，即常将部件和物体混淆。使用本文提出的OICR-ProMIS方法，该错误模式明显下降，其他错误模式稍有变化，正确定位占比明显提升，如图9(b)所示。该实验表明ProMIS方法可明显改善检测器将部件混淆为物体的现象。

对于SDCN检测器来说，它的错误模式主要为真值在框内如图9(c)所示，即常将上下文、多个物体和单个物体混淆。应用本文提出的ProMIS方法后，该错误模式明显下降，同时正确定位占比明显提升，如图9(d)所示。由此可见ProMIS方法可改善检测器将上下文或多个物体混淆为单个物体的现象。

ProMIS方法在低重叠和无重叠这两种错误模式上，较基线算法没有明显下降。主要是由于检测结果本身存在完全分类错误、漏检物体某部件或误检带有固定模式的背景区域，当某种错误模式出现频率较高时，增广得到的伪边界框标签也会携带该错误模式，这种情况下图像增广的方式有时不能从根本上改善检测效果。

4　结论

本文提出的ProMIS方法通过检出物体估计物体布局的概率密度函数，在估计出的概率图中采样得到较为合理的物体类别、位置、尺度，再依此将检出物体插入到输入图像上获得具有部分监督信号的增广图像，并进一步用于训练弱监督检测器。虽然弱监督物体检测任务只能产生带大量噪声的伪边界框标签，但是在给予适当约束时用这些伪边界框标签产生的增广图像仍有丰富信息，能够促使弱监督检测器学习物体部件、整体、多物体簇之间的区别。然而，如果弱监督物体检测器本身对实例的判别带有一致的偏差，图像增广算法会加重这种偏差。该问题的解决需要弱监督物体检测器和增广算法的深度交互，判断样本的可靠性，有待进一步的探索。

参考文献（References）

[1] Bilen H and Vedaldi A.2016.Weakly supervised deep detection networks//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Las Vegas: IEEE:2846-2854 [DOI: 10.1109/CVPR.2016.311]

[2] Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J, Zhang Z, Cheng D, Zhu C, Cheng T, Zhao Q, Li B,

Lu X, Zhu R, Wu Y, Dai J, Wang J, Shi J, Ouyang W, Loy C C and Lin D.2019.MMDetection: Open mmlab detection toolbox and benchmark [EB/OL].[2019-06-17].https://arxiv.org/pdf/1906.07155v1.pdf

[3] Chen Z, Fu Z, Jiang R, Chen Y, Hua X.2020.SLV: spatial likelihood voting for weakly supervised object detection//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Virtual: IEEE/CVF:12992-13001 [DOI: 10.1109/CVPR42600.2020.0130

[4] Deselaers T, Alexe B and Ferrari V.2012.Weakly Supervised Localization and Learning with Generic Knowledge.International Journal of Computer Vision (IJCV), 100(3): 275--293 [DOI: 10.1007/s11263-012-0538-3] 12

[5] Dwibedi D, Misra I and Hebert M.2017.Cut, paste and learn: surprisingly easy synthesis for instance detection//In Proceedings of the IEEE International Conference on Computer Vision (ICCV).Venice: IEEE:1310-1319 [DOI: 10.1109/ICCV.2017.146]

[6] Everingham M, Eslami SM A, Van G L, Williams C KI, Winn J and Zisserman A.2015.The pascal visual object classes challenge: A retrospective.International Journal of Computer Vision (IJCV), 111(1): 98--136 [DOI: 10.1007/s11263-014-0733-5].

[7] Everingham M, Van G L, Williams C KI, Winn J and Zisserman A.2010.The pascal visual object classes (voc) challenge.International Journal of Computer Vision (IJCV), 88(2): 303--338 [DOI: 10.1007/s11263-009-0275-4]

[8] Fang H S, Sun J, Wang R, Gou M, Li Y L, Lu C.2019.InstaBoost: boosting instance segmentation via probability map guided copy-pasting//In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).Seuol: IEEE:682-691 [DOI: 10.1109/ICCV.2019.00077]

[9] Gokberk C R, Verbeek J and Schmid C.2014.Multi-fold mil training for weakly supervised object localization//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Columbus: IEEE:2409-2416 [DOI: 10.1109/CVPR.2014.309]

[10] He K, Zhang X, Ren S, Sun J.2016.Deep Residual Learning for Image Recognition//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Las Vegas: IEEE:770-778 [DOI: 10.1109/CVPR.2016.90]

[11] Ghiasi G, Cui Y and Srinivas A.2021.Simple Copy-Paste is a Strong Data Augmentation Method//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Virtuall: IEEE/CVF:2917-2927 [DOI: 10.1109/CVPR46437.2021.00294]

[12] Kisantal M, Wojna Z, Murawski J, Naruniec J and Cho K.2019.Augmentation for small object detection//In Proceedings of the 9th International Conference on Advances in Computing and Information Technology(ACITY 2019).Chennai, : Aircc Publishing Corporation: 119-133 [DOI:10.5121/csit.2019.91713]

[13] Kosugi S, Yamasaki T and Aizawa K.2019.Object-Aware Instance Labeling for Weakly Supervised Object Detection//In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).Seuol: IEEE/CVF:6063-6071 [DOI: 10.1109/ICCV.2019.00616]

[14] Kumar K S and Lee Y J.2019.You reap what you sow: Using Videos to Generate High Precision Object Proposals for Weakly-supervised Object Detection//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Long Beach: IEEE/CVF:9406-9414 [DOI: 10.1109/CVPR.2019.00964]

[15] B.Online Algorithms and Stochastic Approximations.1998.Online Learning and Neural Networks, Cambridge University Press: 9-42 [DOI: 10.1017/CBO9780511569920.003] 29

[16] Li X, Kan M, Shan S and Chen X.2019.Weakly supervised object detection with segmentation collaboration//In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).Seuol: IEEE/CVF:9734-9743 [DOI: 10.1109/ICCV.2019.00983]

[17] Perez P, Gangnet M and Blake A.2003.Poisson Image Editing//ACM Transactions on Graphics, 22(3): 313-318 [DOI: 10.1145/882262.882269]

[18] Ren Z, Yu Z, Yang X and Liu M Y.2020.Instance-aware, context-focused, and memory-efficient weakly supervised object detection//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Virtuall: IEEE/CVF:10595-10604 [DOI: 10.1109/CVPR42600.2020.01061]

[19] Shen Y, Ji R, Wang Y, Chen Z, Zheng F, Huang F and Wu Y.2020.Enabling deep residual networks for weakly supervised object detection//In Proceedings of the European Conference on Computer Vision (ECCV).Online: Springer:118--136 [DOI: 10.1007/978-3-030-58598-3_8]

[20] Shen Y, Ji R, Wang Y, Wu Y, Cao L.2019.Cyclic guidance for weakly supervised joint detection and segmentation//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Long Beach: IEEE/CVF:697-707 [DOI: 10.1109/CVPR.2019.00079]

[21] Shen Y, Ji R, Yang K, Deng C and Wang C.2020.Category-Aware Spatial Constraint for Weakly Supervised Detection.IEEE

Transactions on Image Processing (TIP), 29: 843-858 [DOI: 10.1109/TIP.2019.2933735]

[22] Shen Y, Ji R, Zhang S, Zuo W and Wang Y.2018.Generative Adversarial Learning Towards Fast Weakly Supervised Detection//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Salt Lake City: IEEE:5764-5773 [DOI: 10.1109/CVPR.2018.00604] 23

[23] Simonyan K and Zisserman A.2014.Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL].[2014-09-04].https://arxiv.org/abs/1409.1556.pdf

[24] Tang P, Wang X, Bai X and Liu W.2017.Multiple instance detection network with online instance classifier refinement//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).HonoluluLas: IEEE:3059-3067 [DOI: 10.1109/CVPR.2017.326]

[25] Tang P, Wang X, Wang A, Yan Y, Liu W, Huang J and Yuille A.2018.Weakly Supervised Region Proposal Network and Object Detection//In Proceedings of the European Conference on Computer Vision (ECCV).Munich: Springer:370-386 [DOI: 10.1007/978-3-030-01252-6_22]

[26] Wan F, Liu C, Ke W, Ji X, Jiao J, Ye Q.2019.C-MIL: continuation multiple instance learning for weakly supervised object detection//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Long Beach IEEE:2194-2203 [DOI: 10.1109/CVPR.2019.00230] 14

[27] Wan F, Wei P, Han Z, Jiao J and Ye Q.2019.Min-Entropy Latent Model for Weakly Supervised Object Detection.IEEE Transactions on Pattern Analysis and Machine Intelligence.41(10): 2395-2409 [10.1109/TPAMI.2019.2898858] 22

[28] Wei Y, Shen Z, Cheng B, Shi H, Xiong J, Feng J and Huang T.2018. : Tight Box Mining with Surrounding Segmentation Context for Weakly Supervised Object Detection//In Proceedings of the European Conference on Computer Vision (ECCV).Munich: Springer:454-470 [DOI: 10.1007/978-3-030-01252-6_27]

[29] Zhang X, Feng J, Xiong H, Tian Q.2018.Zigzag learning for weakly supervised object detection//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Salt Lake City: IEEE:4262-4270 [DOI: 10.1109/CVPR.2018.00448] 16

[30] Zhang X, Yang Y and Feng J.ML-LocNet: Improving Object Localization with Multi-view Learning Network//In Proceedings of the European Conference on Computer Vision (ECCV).Munich: Springer: 248-263 [DOI: 10.1007/978-3-030-01219-9_15]

作者简介

李笑颜，1991 年生，女，博士研究生，主要研究方向为计算机视觉、物体检测与弱监督物体检测。

E-mail：xiaoyan.li@vipl.ict.ac.cn

山世光，通信作者，男，研究员，主要研究方向为计算机视觉、模式识别和机器学习。

E-mail：sgshan@ict.ac.cn

阚美娜，女，副研究员，主要研究方向为计算机视觉、模式识别、迁移学习与弱少监督学习。

E-mail：kanmeina@ict.ac.cn

梁浩，男，硕士研究生，主要研究领域为计算机视觉、物体检测和弱监督物体检测。

E-mail：lianghao211@mails.ucas.ac.cn