用于自监督检测预处理的实例定位

Ceyuan Yang[t] Zhirong Wu[f] Bolei Zhou[t,主] Stephen Lin[f]

[t]Chinese University of Hong Kong [f]Microsoft Research Asia

摘要

先前关于自监督学习的研究已经在图像分类方面取得了相当大的进展，但是在目标检测方面的传输性能通常会下降。本文的目的是提出专门用于目标检测的自监督预处理模型。基于分类和检测的内在区别，我们提出了一种新的自监督前置任务，称为实例定位。图像实例以不同位置和比例粘贴在背景图像上。前置任务是用来预测给定合成图像以及前景边界框的实例类别。我们发现，将边界框集成到预处理模型中可以使得迁移学习有更好的任务对齐和架构对齐（能力）。此外，我们还提出了一种基于边界框的加强方案来进一步增强特征对齐。因此，我们的模型在图像语义分类中比较弱，但在图像块定位中更强，具有用于对象检测的整体表现更强的预处理模型。实验结果证明，我们的方法为在 PASCAL VOC 和 MSCOCO 的数据集的目标检测上提供了最先进的迁移学习结果。

1. 介绍

在计算机视觉中训练深层网络的主要范例是预处理和微调[20，29]。典型地，优化后的预处理可以找到单一的通用表征，该表征随后被应用到各种下游应用。例如，使用图像级标签的监督预训练模型[26，25]和通过对比学习的自监督预训练模型[22]都非常好地适用于许多任务，例如，。图像分类、目标检测、语义分割和人体姿态估计。

尽管这种方法很受欢迎，但我们质疑这种通用的迁移学习表征是否存在。最近，有观察指出能够提高图像分类性能的自监督表征无法将其优势迁移到目标识别任务。

此外，我们还发现，在向检测和分割转移时，高级特征并不是真正重要的因素[46]。这表明当前的自监督模型可能对分类任务过拟合，而对其他关注的任务变得不那么有效。

我们发现了导致迁移学习中任务错位的两个问题。首先，需要将预处理后网络重新用于目标网络架构中，

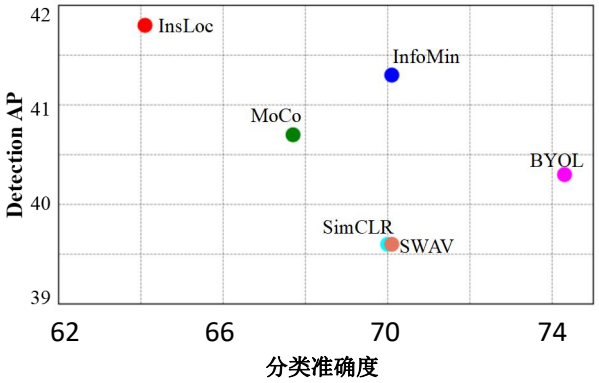

图1.对于视觉迁移学习，人们普遍认为ImageNet分类精度和目标检测性能是正相关的。通过研究最近的自监督模型，我们发现事实并非如此。我们提出了一种新的方法，称为实例定位(InsLoc)，它牺牲了ImageNet的分类精度，但对目标检测具有更好的泛化能力。

以进行微调。这通常涉及非平凡的架构变化，例如插入特征金字塔[27]或采用具有大膨胀的卷积核[4]。第二，对于典型的对比学习模型，预处理前置任务在实例辨别中整体考虑图像[41]，而没有对区域进行明确的空间建模。虽然它增强了分类的可转移性，但这种做法与空间推理任务(如对象检测)不太兼容。

本文提出了一种新的自监督预监督前置任务，称为实例定位，专门针对目标检测的下游任务。类似于实例识别，它为单个图像实例学习一个分类器，实例定位额外考虑了边框信息用于表示学习。我们创建我们的训练集，方法是将前景图像裁剪并以不同的纵横比（宽高比）和比例粘贴到背景图像的不同位置。自监督预处理通过使用边界框提取 RoI 特征并且使用实例标签进行对比学习来实现。这样，不仅网络架构在迁移过程中保持一致性，而且预处理任务还包括定位建模，这对目标检测至关重要。

在预处理中引入边框增强卷积特征和前景区域之间的显式对齐。因此，特征响应对图像域的翻译变得敏感，有利于检测[10]。我们还发现，特征对齐可以

通过在边界框坐标上引入增强来加强。具体地，从一组区域建议锚点中随机选择空间抖动的边界框。

我们在动量对比的框架内实施该方法[22]。该网络以合成图像和边界框为输入，提取区域嵌入进行对比学习。与考虑整体实例的基线方法相比，对最后一层特征的线性探测显示图像分类的性能降低，同时对边界框位置回归方面实现了改进。实验上，我们研究了两种流行的检测骨干网络，ResNet50-C4 和 ResNet50-FPN。对于这两种骨干网络，我们的实例定位方法大大提高了性能，超过了 PASCAL VOC [17]和 MSCOCO [28]的最先进的传输学习结果。值得注意的是，我们的模型对于小数据体制下的目标检测更加有利。

2. 相关工作

自监督学习。自监督学习的核心思想是从视觉数据中创建自由监督标签，并使用自由监督来获得可推广和可转移的表征。前置任务的最简单形式之一是使用生成模型重建输入图像。生成模型中的潜在表征被认为是捕捉输入分布的高级结构和语义流形。自动编码器[39]和玻尔兹曼机器[37]在手写数字上显示了这种能力，但在自然图像上无法工作。后来，GANs [47]的进展通过将潜在表征的神经反应分解成面部属性、姿势和光照条件，实现了对生成内容的操纵。最近关于 BigBiGAN [14]和 Image-GPT [6]的工作表明，超大生成模型可能会提供非常有前途的视觉识别表征。然而，仍然存在的一个基本问题是，如何将学习生成图像像素与高级的视觉理解相关联。

除了重建图像像素，另一种前置任务是保留一部分数据，然后从另一部分预测它。彩色化[44]保留了颜色信息，并试图从灰度值预测它。上下文预测[12]将空间内容分割成 3×3 的网格。然后训练网络来预测网格之间的空间关系。前置任务的制定方式对从数据中学到的知识有很大影响。当同一类别中的对象共享相同的颜色时，彩色化方法往往会起作用。上下文预测假设一个类别的对象共享相同的空间配置。既然不同的前置任务可以提取不同方面的视觉知识，（那么使用）一种多任务方法[13]结合了他们的个人知识，可以提高学习效果。

自监督学习的一个流行前置任务是对比学习，或者更具体地说是实例辨别[41]。训练数据集中的每个实例都被视为一个单独的类别。学习的目标只是将每个实例与其他实例进行分类。对比学习的关键部分是用于归纳不变量的数据扩充[41，22，7]。理想的数据增强应该反映类内变化，常用的增强包括裁剪、缩放、色彩抖动和模糊。最近关于对比学习的研究集中在开发更好的扩充方式[3]，设计项目头部结构[21]，甚至减轻负样本的要求[21]。尽管领先的对比学习方法 BYOL 和 SwAV 利用线性 readoff 分类器将 ImageNet 的性能提升到了令人惊讶的 74%，但它们对对象检测的迁移性能实际上低于 MoCo [22]。这表明这些自我监督的方法过度适应单一的下游分类任务，而牺牲了对其他任务的泛化能力。

我们提出了一个新的自我监督预处理前置任务，重点转移到目标检测。基于实例识别，我们在预处理阶段引入了边界框的使用。为了改进定位，我们的方法学习了一种表征方式，其中边界框和它们对应的前景特征之间是对齐的。明确处理网格级空间建模的先前工作包括 CPC [33]和上下文预测[12]。这些工作基于图像中的网格内容来解释空间排列。相反，我们的前置任务考虑合成在一起的两个不同图像之间的空间关系。

图像和视频上的自我监督学习还有一系列其他前置任务，例如图像修复- [35]、旋转预测[19]、拼图[32]、运动分割[34]，以及视频上的时间顺序[31]、时间速度[2]和同步[1]。对每个前置任务的详细调查和描述超出了本文的范围。

图像构图学习。通过将前景对象复制到背景上来创建合成图像是一种流行的数据增强技术。给定前景对象遮罩，先前的工作成功地将该技术应用于监督实例分割[16，18]和无监督学习[45]。我们的工作也合成图像组合，但不需要对象遮罩或干净的轮廓。

无标记的数据的自训练。除了迁移学习，自我训练[42，48]是一个有前途的方向，当标记数据有限时，可以利用未标记数据。其思想是通过在少数标记样本上使用监督学习来引导模型，从而在未标记样本上生成伪标签。通过对标签和伪标签的联合监督学习，进一步优化了模型。然而，当标签集稀缺时，自我训练可能会变得脆弱。正如 SimCLR-v2 [8]中所探讨的那样，迁移学习和自我训练可以整合在一起。

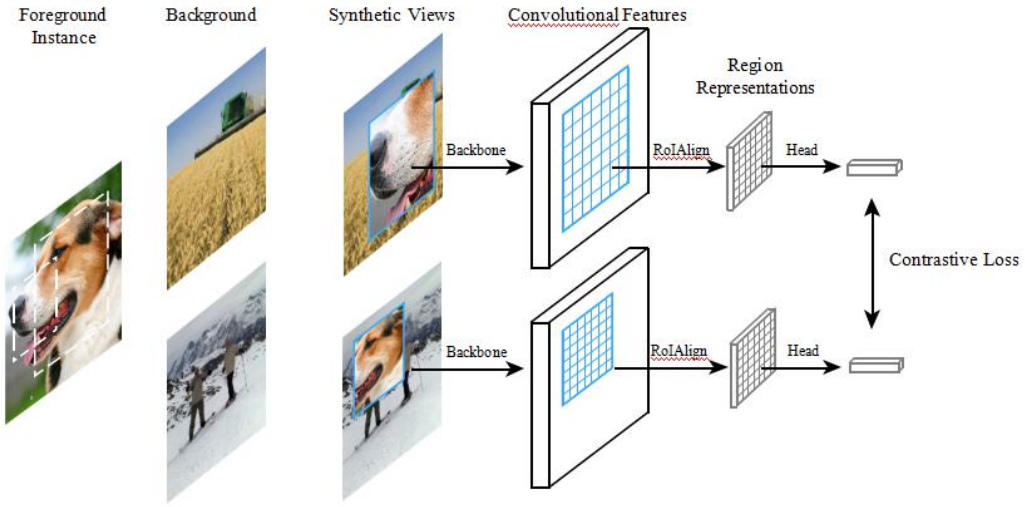

图 2.**实例定位概述。**给定一个前景图像实例，我们首先从图像库中随机抽取两幅背景图像。生成前景图像的两个视图，并将其复制粘贴到相应的背景图像上。卷积网络获取每个合成视图，而 RoIAlign 使用前景边界框坐标提取区域表示。对比学习遵循区域表征。为简洁起见，省略了负样本。

3. 前置任务–实例定位

图像分类支持平移和比例不变性，其中不同比例和位置的对象被简化为代表对象类别的单个离散变量。相比之下，目标检测需要平移和比例等变。用于对象检测的特征表示应该能够保留和反映关于对象大小和位置的信息。这两个任务之间的固有差异要求为每个任务专门建模。对比学习的最新研究集中在图像分类的设计技术上。平移和缩放不变性通过学习图像的两个随机视图之间的一致性来实现。因此，实例识别的前置任务过度偏向于整体分类，不能促进空间推理。

我们提出了一个新的前置任务，称为实例定位(InsLoc)，作为实例识别的扩展。如图 3 所示，我们通过将前景实例叠加到背景上来合成图像。目地是使用边界框信息区分前景和背景。为了实现这个任务，必须首先定位前景实例，然后提取前景特征。

将合成图像表示为$\mathbf{I}'$，前景图像**I**覆盖在边界框 b 上。任务 T 是为**I**预测实例标签 y，

$$\mathbf{y} \leftarrow \mathbf{T}(\mathbf{I}', \mathbf{b})。(1)$$

4. 学习方法

我们的目标是学习一种不仅语义强大，而且与迁移和规模相当的表示法。我们首先在 Sec.4.1 中描述了我们将边界框表示引入对比学习框架的方法。边界盒上的数据扩充是在 Sec.4.2 中提高定位能力的有效方法。最后，我们在 Sec.4.3 中给出了我们的方法在两个流行的检测主干(R50-C4 和 R50-FPN)上的架构细节。

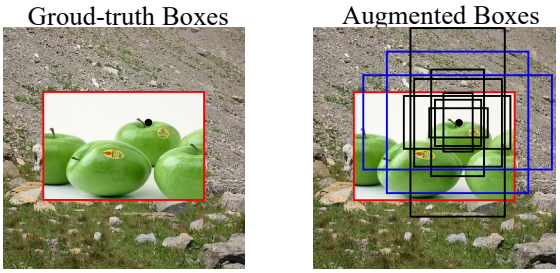

图 3.**空间模型的边界框。**红色框表示前景图像的真实边界框。在右侧，我们显示了一组以单个空间位置为中心的定位框。通过利用具有不同比例、位置和长宽比的多个锚，我们用 IoU 大于 0.5 的蓝色框来增强地面真实性。

4.1. 边界框的实例判别

实例判别。对比学习将两个随机的“视图”作为查询字$\boldsymbol{I_q}$和关键字$\boldsymbol{I_{k_+}}$，这两个图像是从同一个实例的随机扩充中导出的。相应的特征$\boldsymbol{v_q}$和$\boldsymbol{v_{k_+}}$首先由主干网 f 提取(例如。$\boldsymbol{v_q}$ = f($\boldsymbol{I_q}$))，然后通过头网络Φ投影到单位球面。对比损失，即。信息[33]计算如下

$$\mathcal{L} = -\log \frac{\exp(\phi(v_q)\cdot\phi(v_{k_+})/\tau)}{\sum_{i=0}^{N}\exp(\phi(v_q)\cdot\phi(v_{k_i})/\tau)} \quad (2)$$

其中τ和 N 分别是温度和负样本数。

使用边界框的空间模型。我们的目标是加强输入区域和卷积特征之间的空间对齐，以及区别实例的对比学习。为此，给定图像$\boldsymbol{I}$，我们首先对随机背景图像 B 进行采

样，该图像简单地作为训练集中的任何其他图像。然后我们定义合成操作 C，它以随机的位置和比例将图像$\boldsymbol{I}$的随机裁剪复制并粘贴到背景 B 上。该操作返回合成图像$\boldsymbol{I_0}$和边界框参数 b，

$$I'_q, b_q = C(I_q, B_q), \tag{3}$$
$$I'_{k_+}, b_{k_+} = C(I_{k_+}, B_{k_+}), \tag{4}$$

其中$\boldsymbol{I_q}$和$\boldsymbol{I_{k_+}}$是来自同一图像实例的裁剪，$\boldsymbol{B_q}$和$\boldsymbol{B_{k_+}}$是它们各自的背景图像。实际上，前景图像以 128 到 256 像素之间的随机纵横比和随机比例来调整大小。利用边界框参数 b，应用 RoIAlign[24]来提取卷积特征图上的前景特征，

$$v'_q = \text{RoIAlign}(f(I'_q), b_q), \tag{5}$$
$$v'_{k_+} = \text{RoIAlign}(f(I'_{k_+}), b_{k_+}). \tag{6}$$

有了查询字和关键字特征，对比学习遵循类似于 Eq.2。Figure.3 展示了我们的框架。

使检测复杂化的一个问题是图像区域与其空间上对应的深层特征之间的差异。由于汇集的深层特征的感受野通常在图像中延伸到汇集区域之外，因此汇集的特征受到其附近区域之外的图像内容的影响。因此，对于覆盖前景的边界框，其特征会受到周围背景的影响，使得定位变得更加困难。

我们使用边界框的实例识别以数据驱动的方式解决这个问题。通过增强相同实例但具有不同背景的混合前景特征之间的相似性，学习有效感受野以匹配边界框的空间范围。在卷积特征和它们的有效感受野之间建立这种明确的对应关系，有助于用所学的表征进行定位。

4.2. 边界框增强

图像增强在关于表征[15，7]的对比学习中起着关键作用。我们假设类似的增强策略也可能对边界框有效。具体来说，地面真实位置周围的抖动框可能包括背景上的区域。因此，可以进一步引导表示在空间上忽略背景并获得定位能力。

作为预定义锚的增强。我们不是直接在空间上移动边界框，而是利用区域建议网络(RPN) [36]中的锚来覆盖增强框的多样性。锚点是一组预定义的边界框方案，具有不同的比例、位置和纵横比。给定一个地面真值框，我们根据所有锚计算它的 IoU。具有高重叠(大于 0.5)的锚被过滤，并且随机的一个被选择作为增强框。由于基于锚的设计，我们能够获得各种各样具有动态 IoUs 范围的边界框提案。我们在查询编码器的 RoIAlign 模块上应用边界框拓展数据。

Methods	Epoch	AP	AP_{50}	AP_{75}
Random init	-	33.8	60.2	33.1
Supervised	90	53.5	81.3	58.8
Relative Loc.	200	50.6	76.9	55.2
MoCo-v2	200	57.0	82.4	63.6
MoCo-v2	800	57.4	82.5	64.0
InfoMin	200	57.6	82.7	64.6
InfoMin	800	57.5	82.7	64.3
SimCLR	200	51.5	79.4	55.6
BYOL	300	51.9	81.0	56.5
SwAV	400	45.1	77.4	46.5
InsLoc	200	57.9	82.9	64.9
InsLoc	400	**58.4**	**83.0**	**65.3**

表 1.PASCAL VOC 上的目标检测。模型在 trainval07+12 上进行了微调，并在 test2007 上进行了测试.我们自己评估 SimCLR、BYOL、SwAV 模型，同时报告他们原始论文的剩余结果。所有数字都是五次试验的平均值。

4.3. 架构对齐

导致迁移学习中任务错位的一个关键问题是不平凡的架构调整。需要通过附加区域操作和头部网络将预处理后的网络重新用于检测网络。我们引入的边界框表示允许最小化预处理和微调之间的架构差异。具体来说，预处理中的 RoIAlign 操作引入了区域式表示，它紧密模拟了微调中的检测行为。我们在预处理期间提供了检测架构 R50-C4 和 R50-FPN 的详细信息。

R50-C4。在标准 ResNet50 架构上，我们在第 4 个残差块的输出上插入 RoI 操作。然后使用边界框坐标提取区域特征。整个第 5 个残差块被视为用于对区域进行分类的头部网络。

R50-FPN。R50-FPN 使用横向连接在 ResNet50 之上形成 4 层功能层次结构.每一级特征负责以相应的比例建模对象。我们在 FPN 层级的所有级别插入 RoI 操作。实例定位任务在所有 4 个特征级别上同时执行[43]，其中每个级别都维护一个单独的负面示例内存队列，以避免跨级别欺骗。这样，不仅可以对 ResNet50 网络也可以对 FPN 层进行预处理。

5.实验结果

我们在主流的目标检测基准上评估了我们的迁移学习模型的泛化能力:PASCAL VOC [17]和 MSCOCO [28]。第 5.1 节给出了主要的实验结果和最先进的结果的比较.关于语义分类和定位之间权衡的消融研究和讨论在第 5.2 节中进行.在第 5.3 节中，我们在小型的 MSCOCO 上进行了一个实验，以证明我们的模型在少量标记数据下的快速泛化能力。

数据集。拥有 130 万张图像的 ImageNet 数据集[11]用于预处理，而 PASCAL VOC [17]和 MSCOCO [28]用于迁移学习。PASCAL VOC0712 包含大约 16.5K 个图像，带有 20 个对象类别的边界框注释。MSCOCO 包含大约 118K 图像，带有 80 个对象类别中的边界框和实例分割注释。

预处理。我们在很大程度上遵循 MoCo-v2 官方实现的超参数[9]。我们优化了模型，同步 SGD 超过 8 个图形处理器，权重衰减为 0.0001，动量为 0.9，每个图形处理器上的批处理大小为 32。优化需要 200-400 个时期，初始学习率为 0.03，余弦学习率时间表[30]。两层 MLP 头用于对比学习，温度参数在等式 2 中设置为 0.2. 我们还维护了 65536 个负样本的内存队列。动量系数设置为 0.999，用于更新关键编码器。

数据扩充。在预处理期间，前景内容的图像增强遵循 MoCo-v2 [9]。具体来说，我们应用随机调整大小的裁剪，颜色抖动，灰度缩放，高斯模糊和水平翻转。更强的增强可能会进一步提高转移性能[38，7]，但不在我们的工作重点之内。

微调。骨干网从预处理任务转移到下游任务。在 MoCov2 [9]之后，同步批处理标准化被用于所有层，包括新初始化的批处理标准化层。使用 detectron2 [40]实现并微调检测器。

5.1. 主要结果

我们提供了目标检测的实验结果，并与最先进的方法进行了性能比较。SimCLR [7]和 BYOL [21]的预处理权重是从第三方实现中借用的，而 MoCo [22]、InfoMin [38]和 SwAV [3]的权重是从它们的官方实现中收集的。

5.1.1 PASCAL VOC 物体检测

设置。我们使用 R50C4 主干架构的更快的 R-CNN 检测器[36]。优化总共需要 24k 次迭代。学习速率初始化为 0.02，经过 18k 和 22k 次迭代后衰减为原来的 10 倍。训练时图像比例在[480，800]像素内，推断时设置为 800。AP、AP_{50}和AP_{75}显示为评估指标。

结果。迁移结果总结在表 1 中.由于方差较大，所有数值均在五次独立试验中取平均值。我们报告了 200 个时期和 400 个时期的预处理结果。与我们的直接基线 MoCo-v2 [9]相比，我们的模型分别在 200 和 800 个时期获得了+0.9 和+1.0 的改进。它也优于所有以前的方法，不使用复杂和更强的数据增强，如随机增强或多作物。我们的预处理模型在这个基准上获得了最新的结果。

5.1.2 COCO目标检测和分割

设置。我们在 R50-C4 和 R50-FPN 主干网络中使用掩码 R-CNN [24]框架。由于以前的文献[23]建议，当训练时间表很长时，具有随机初始化的检测器可以匹配 COCO [28]上的监督对应检测器，因此我们在具有 180k 次优化迭代的 2×时间表上进行该转移实验。学习速率初始化为 0.02，经过 120k 和 160k 次迭代后衰减为 10 倍。训练时图像比例在[640，800]像素以内，测试时设置为 800。AP、AP_{50}和AP_{75}显示为边界框检测和实例分割的评估指标。

Methods	Epoch	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Random	-	35.6	54.6	38.2	31.4	51.5	33.5	38.4	57.5	42.0	34.7	54.8	37.2
Supervised	90	40.0	59.9	43.1	34.7	56.5	36.9	41.6	61.7	45.3	37.6	58.7	40.4
Rel. Loc.	200	38.0	57.4	41.0	33.3	54.1	35.4	39.4	58.7	42.7	35.6	55.9	38.1
MoCo-v2	200	40.7	60.5	44.1	35.6	57.4	37.1	41.7	61.6	45.6	37.6	58.7	40.5
MoCo-v2	800	41.2	60.9	44.6	35.8	57.7	38.2	42.5	62.3	46.8	38.2	59.6	41.1
InfoMin	200	41.3	61.2	45.0	36.0	57.9	38.3	42.5	62.7	46.8	38.4	59.7	41.4
InfoMin	800	41.2	61.2	44.8	35.9	57.9	38.4	42.1	62.3	46.2	38.0	59.5	40.8
SimCLR	200	39.6	59.1	42.9	34.6	55.9	37.1	40.8	60.6	44.4	36.9	57.8	39.8
BYOL	300	40.3	60.5	43.9	35.1	56.8	37.3	42.3	62.6	46.2	38.3	59.6	41.1
SWAV	400	39.6	60.1	42.9	34.7	56.6	36.6	42.3	62.8	46.3	38.2	60.0	41.0
InsLoc	200	41.4	60.9	45.0	35.9	57.6	38.4	43.2	63.5	**47.5**	38.7	60.5	**41.9**
InsLoc	400	**41.8**	**61.6**	**45.4**	**36.3**	**58.2**	**38.8**	**43.3**	**63.6**	47.3	**38.8**	**60.9**	41.7
(a) Mask R-CNN, R50-**C4**, **2**× schedule								(b) Mask R-CNN, R50-**FPN**, **2**× schedule					

Table 2. **Object detection and instance segmentation on COCO.** Models are fine-tuned on `train2017` and tested on `val2017`.

结果。表 2 显示了 R50-C4(表 2a)和 R50-FPN(表 2b)的结果。AP^{bb}和AP^{mk}分别表示边界框检测和实例掩码分割的 AP。经过 200 个时期的预训练，InsLoc 在 R50-C4 和 R50-FPN 主干上的性能优于直接基线 MoCo-v2[22]+0.7 和+1.5 AP。经过 400 个时代的预处理，InsLoc 达到了新的效率水平，超越了所有以前的自监督模型，可能有更强的图像增强。特别是，InsLoc 引入了对完全监督的 ImageNet 预处理的重大改进。R50-C4 和 R50-FPN 的 AP 分别为+1.8 和+1.7。值得注意的是，当模型预处理时间较长时，InfoMin 显示迁移性能下降。BYOL 和 SwAV 在 R50-FPN 主干方面具有竞争力，但在 R50-C4 主干方面相对较弱。我们的模式在各个方面都更加强大。

5.2. 消融研究

为了进一步了解实例定位的优势，我们进行了一系列消融研究，检查语义和定位的权衡、新前置任务的影响、用更长的时间表进行微调以及在更优化的时期进行预处理。

这种改善是否是由于更强的语义特征。最近的方法倾向于将线性对象分类作为学习表示的核心评估指标，基于具有更强语义的表示总是能很好地迁移成其他下游任务的假设。为了进一步研究和理解所提出的目标检测方法的改进，我们设计了一个新的读出任务来评估预处理模型的定位能力。

具体来说，给定一个输入图像，我们将整个图像分割成 M 个面片。任务是预测使用线性分类器的块的区域特征的位置。

Methods	Cls	Loc	AP^{bb}	AP^{mk}
SWAV	70.1	58.4	34.0	30.4
BYOL	74.3	67.6	37.5	32.8
MoCo-v2	67.7	71.9	38.9	34.1
InsLoc	61.7	74.2	39.5	34.5

(a) Semantic *vs*. Localization. The linear readout accuracy of linear classification (Cls) and localization (Loc) as well as overall fine-tuned detection AP are presented. Detector architecture is R50-C4.

RA	CP	BBA	AP^{bb}	AP^{mk}
			39.8	36.1
✓			40.2	36.4
✓	✓		41.1	36.9
✓	✓	✓	41.4	37.1

(b) RoiAlign (RA) is inserted into the baseline to reflect architectural changes. The instance localization task is then performed, *i.e.* foreground images are copied and pasted (CP) onto the background images to learn the spatial alignment. Bounding-box augmentation (BBA) is finally applied. The experiments are performed on the R50-FPN architecture.

Table 3. **Ablation Studies.** All numbers are reported with 1× schedule on the COCO `val2017` set.

图 4 说明了 M 等于 9 时的任务.虽然早期关于上下文预测的工作[12]预测了两个网格之间的相对空间位置，但是我们的评估任务考虑了网格相对于完整图像的空间排列。对于每个网格，我们通过主干网络转发、提取 RoI 特征并通过头部网络来提取其矢量表示。我们附加了一个线性分类器来预测网格索引。我们认为这个任务类似于检测流水线，反映了预训练模型的定位能力。

表 3a 显示了语义和定位准确性的比较。实例定位为线性定位任务带来了 2.3%的明显改进，而在线性分类中的 MoCo-v2 性能下降了 6.0%。这表明，目标检测的整体改进主要是由更好的空间定位带来的，而不是更强的语义。这些结果也与最近的发现[46]相匹配，即自我监督的预处理不会转移高级语义用于对象检测，而是更重要的低级和中级转移。我们还在表 3a 中加入了 BYOL 和 SWAV 的本地化评估条目。它们较差的定位能力限制了转移到目标检测的有效性。

实例定位前置任务的有效性。表 3b 展示了多个组件的消融研究:架构校准、实例定位任务和建议的边界框增强。我们首先集成了 RoiAlign 操作符来减轻对基准模

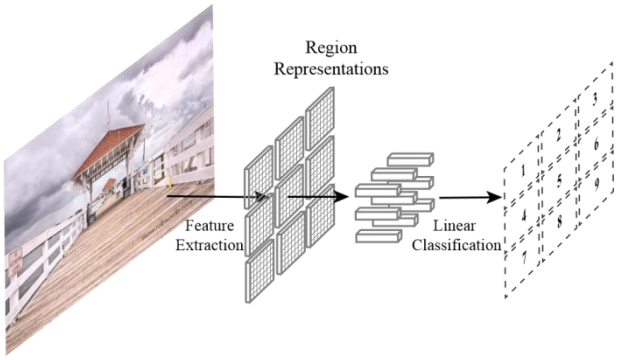

图 4. **线性定位评估。**我们将自然图像分割成区域面片网格。对于每个区域，我们提取其向量表示，并训练线性分类器来预测整个图像中的区域索引。

型结构的更改。具体来说，整体表征被汇集并从网络中提取，然后经过对比学习。此外，在 FPN 等级体系(第 4.3 节)上应用了多个对比损失，导致整体提高 +0.4 AP^{bb}。这些改进证明了调整检测转移架构的有效性。然后，我们使用复制粘贴操作和边界框表示对合成图像应用实例定位任务，性能达到 41.1 AP^{bb}，比 MoCov2 明显高出+1.3.最后，当对边界框应用空间抖动时，结果进一步提升到 41.4 AP^{bb}。这些结果有力地验证了新前置任务的有效性，比如实例定位和相应的扩充。

微调时间表效果。随着迭代次数的增加，对下游对象检测任务进行微调可以提高对象检测性能。我们研究

Methods	Epoch	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
MoCo-v2	200	38.9	58.6	41.9	34.1	55.5	36.0
MoCo-v2	800	39.3	58.9	42.5	34.3	55.7	36.5
InsLoc	200	39.5	59.1	42.7	34.5	56.0	36.8
InsLoc	400	**39.8**	**59.6**	**42.9**	**34.7**	**56.3**	**36.9**

(a) Mask R-CNN, R50-**C4**, **1**× schedule

Methods	Epoch	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
MoCo-v2	200	40.7	60.5	44.1	35.6	57.4	37.1
MoCo-v2	800	41.2	60.9	44.6	35.8	57.7	38.2
InsLoc	200	41.4	60.9	45.0	35.9	57.6	38.4
InsLoc	400	**41.8**	**61.6**	**45.4**	**36.3**	**58.2**	**38.8**

(b) Mask R-CNN, R50-**C4**, **2**× schedule

Methods	Epoch	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
MoCo-v2	200	39.8	59.4	43.6	36.1	56.5	38.9
MoCo-v2	800	40.4	60.2	44.2	36.4	57.2	38.9
InsLoc	200	41.4	61.7	45.0	37.1	58.5	39.6
InsLoc	400	**42.0**	**62.3**	**45.8**	**37.6**	**59.0**	**40.5**

(c) Mask R-CNN, R50-**FPN**, **1**× schedule

Methods	Epoch	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
MoCo-v2	200	41.7	61.6	45.6	37.6	58.7	40.5
MoCo-v2	800	42.5	62.3	46.8	38.2	59.6	41.1
InsLoc	200	43.2	63.5	**47.5**	38.7	60.5	**41.9**
InsLoc	400	**43.3**	**63.6**	47.3	**38.8**	**60.9**	41.7

(d) Mask R-CNN, R50-**FPN**, **2**× schedule

Table 4. **Baseline comparison with MoCo-v2 for object detection and instance segmentation on COCO**. R50-C4 and R50-FPN backbones are fine-tuned under 1× and 2× schedule.

了微调进度如何影响预训练模型的相对改进。在表 4 中，我们研究了在 1x 和 2x 微调计划下对象检测转移到 COCO 的情况。使用 R50-C4，1 倍计划的 0.6 $\boldsymbol{AP^{bb}}$ 改进相当于 2 倍计划的 0.7 $\boldsymbol{AP^{bb}}$ 改进。R50-FPN 也获得了类似的观测结果。这些结果表明，更长时间的微调可能不会显著削弱相对改善，证明了预训练模型对迁移学习的效用。

长时间预处理的影响。ImageNet 线性分类精度从更长的预处理中受益匪浅。例如，MoCo-v2 通过将预处理时期的数量从 200 个增加到 800 个，从 67.5%提高到 71.1%.然而，对于物体检测，较长的预处理可能是有害的，如 InfoMin [38]所示。在表 4 中，我们报告了 400 个优化时期的预处理模型在 COCO 上的传输性能。相比于进行 200 个时期的预处理模型，更长时间的预处理获得了持续的改进和新的最先进的性能。即使用 800 个时期进行更长时间的预处理在计算上是很昂贵的，我们把它留给未来的工作。

5.3. 小型 COCO的评估

由于数据集的规模，COCO 迁移学习的意义可能有限。先前的文献[5]也表明，在学习时间很长的 COCO 上从头开始训练可以提供一个强有力的基线。为了证明我们的预训练模型在少量标记数据下的泛化能力，我们在小型 COCO 数据集上进行了实验。

数据集。我们从最初的 train2017 设置为 Mini COCO 中随机选取 10%的训练数据(约 11.8K 图像)。总的训练数据与 PASCAL VOC 相似[17]。物体在比例和长宽比方面的巨大差异仍然特别具有挑战性。我们使用完整的验证集(即 val2017)，包含用于评估的 5K 注释图像。

微调。微调协议与整个 COCO 保持一致。我们使用 R50-C4 主干网，并为 12 个时期微调网络。在最后一个剩余块之后插入一个附加的批处理规范化层。

结果。表 5 总结了结果。相对于 MoCo-v2，我们获得了 3.3 $\boldsymbol{AP^{bb}}$和 2.4 $\boldsymbol{AP^{mk}}$的较大改进，相对于监督方法，我们获得了 3.1 $\boldsymbol{AP^{bb}}$和 2.3 $\boldsymbol{AP^{mk}}$的较大改进，证明了优越的泛化和迁移能力。注意，Mini COCO 的增益远大于原 COCO 的增益。这些结果清楚地表明，我们的预训练模型对于迁移学习来说更具数据效率。

Methods	Epoch	AP^{bb}	AP^{mk}
Supervised	90	22.9	21.2
Relative Loc.	200	17.2	16.1
MoCo-v2	200	22.7	21.1
InfoMin	200	23.6	21.7
SimCR	200	20.0	18.9
BYOL	300	20.6	19.6
SWAV	400	14.9	15.2
InsLoc	200	**26.0**	**23.5**

Table 5. **Object detection on Mini COCO.** Models are fine-tuned on 10% of COCO `train2017` for 12 epochs and evaluated on `val2017`.

6. 结论

我们提出了一种新的实例定位任务，并介绍了包围盒在自监督表示学习中的应用。预处理后的模型在整体图像分类中表现较弱，但在块定位中表现较强。当转移到目标检测时，它实现了相对于基本 MoCo 的显著改进，并在 VOC 和 COCO 上获得了的最先进的结果。我们还表明，当标记数据特别小时，我们的方法获得了更大的增益。实验结果表明，通过改进任务对齐可以增强目标检测的迁移性能。

致谢。该项目得到了创新和技术基金下的感知和交互智能中心(CPII)有限公司的部分支持。我们也感谢赵南轩、徐英浩和戴博的深刻讨论。