基于实例、图像、数据集信息的弱监督实例分割

刘云 *, 吴宇寰 *, 温佩松, 施宇钧, 邱宇, 程明明

摘要—仅依靠图像级监督的弱监督语义实例分割，而不是依赖昂贵的像素级蒙版或边界框标注，是缓解深度学习的数据耗费性的重要问题。在本文中，我们通过将所有训练图像的图像级别信息聚合到一个大的知识图中并利用该图中的语义关系来解决这一难题。具体来说，我们的工作从一些类别无关的、通用的、基于分割的拟物性采样 (Segment-based Object Proposal, SOP) 开始。我们提出了一个多实例学习 (Multiple Instance Learning, MIL) 框架，该框架可以使用带有图像级标签的训练图像以端到端的方式进行训练。对于每个 SOP，此 MIL 框架可以同时计算概率分布和类别感知的语义特征，利用这些信息我们可以构造一个大型无向图。此图中还包括背景类别，以删除 SOP 中的大量噪声。因此，该图的最佳多路割可以为每个 SOP 分配可靠的类别标签。这些带有指定类别标签的去噪后的 SOP 可以视为训练图像的伪实例分割，用于训练全监督的模型。所提出的方法在弱监督实例分割和语义分割方面都达到了最新的性能。该工作的代码已经开源：https://github.com/yun-liu/LIID.

关键词—弱监督学习，实例分割，语义分割，多实例学习，多路割。

1 引言

实例感知语义分割（简称实例分割）专注于同时检测和分割图像中的所有对象实例。由于其巨大的学术和工业价值，使得它成为计算机视觉中最重要的任务之一。实例分割的最新进展是由功能强大的基准系统驱动的，例如 Fast/Faster/Mask R-CNN [2]–[4] 和全卷积神经网络（Fully Convolutional Network, FCN）[5]。但是，这些深度模型的性能在很大程度上依赖于大量训练数据以及昂贵的逐像素标记。标注此类训练数据一直是将实例分割应用于实际应用中的一个严重瓶颈，因为对大量图像进行逐像素标注特别耗时。例如，在 Cityscapes 数据集中逐像素标注一张图像需要“平均超过 1.5 小时”[6]。

为了减轻对昂贵的逐像素标注的需求，一些研究使用边界框 [7]–[10] 来放宽监督，其中训练数据可以只是用于物体检测的数据。尽管标记边界框要比标记像素便宜，但由于边界框标记仍然是一件劳动密集型的事情，弱监督物体检测实际上早已是一个经过充分研究的领域 [11]–[13]。我们在本文中的工作遵循 [14]–[18] 等进一步放宽监督，即**仅使用图像级监督来进行弱监督实例分割**。由于图像级标签的标注成本较低，因此该类方法将使许多实际应用受益。

在弱监督实例分割中，主要挑战之一是将图像的标签分配给每个语义实例，例如，拟物性采样的物体推荐 [19]。Zhou 等人 [14] 试图通过计算从图像分类器 [20], [21] 获得的类激活图（Class Activation Map, CAM）[22] 中的类峰值响应来解决这一难题。这些峰值响应可用于查询与类别无关的对象建议，以预测实例遮罩。与 [14] 类似，许多其他弱监督实例分割方法 [15]–[18] 和弱监督语义分割方法 [23]–[31] 在很大程度上也依赖于 CAM 进行物体识别。但是，CAM 倾向集中于目标对象的小部分具有区分性的区域，并且 CAM 也很难从包含小对象、多个对象和复杂背景的复杂场景中准确定位对象。尽管已经引入了各种技术 [28], [32]–[34] 来改善 CAM，但 CAM 的天然局限性仍然阻碍了弱监督学习的发展 [17]。

基于以上观察，我们提出了一种可以克服这些局限性的新方法。与以前的基于 CAM 的弱监督分割方法不同，这些方法直接使用 CAM 或 CAM 的改进版本进行物体识别 [14]–[18], [23]–[32]，我们的方法使用 CAM 作为多实例学习（Multiple Instance Learning, MIL）框架中的**监督源之一**来学习训练过程中每张图像的语义信息。因此，CAM 通过提供近似的粗略信息来帮助训练我们的系统，但是我们的系统性能并不完全依赖于 CAM，因为我们还有其他设计来确保对 MIL 框架进行训练，这在实验中得到了证明。此外，我们提出将所有训练图像的有用信息集成到一个大型的知识图中，并探索该图中的信息以桥接图像级标签和相应的语义实例。这样，我们的方法不仅考虑了每个图像的固有属性，还考虑了训练数据库的整体数据分布，从而打破了 CAM 在弱监督分割方面的局限性。

具体来说，我们的工作从一些通用的基于分割的拟物性采样开始（Segment-based Object Proposal, SOP），例如 selective search [35]，LPO [36]，和 MCG [19]。因为这些方法与类别无关，所以它们不依赖任何语义标签。因此，我们的系统可以仅使用图像级信息将其推广到任何类别。给定一张图像的标签和它的 SOP，我们旨在为每个 SOP 分配正确的类别标签并过滤出噪声采样。为了实现此目标，我们建立了一个 MIL 框架，用于使用图像标签作为监督的图像分类。在此框架中，如果一个 SOP 包含特定类别的对象，则我们的模型将学习使该 SOP 为相应类别的最终分类概率做出更多贡献。如果一个

- 所有作者均来自南开大学。
- 刘云与吴宇寰为并列第一作者。程明明为通讯作者 (cmm@nankai.edu.cn)。

- 本文是 IEEE TPAMI 论文 [1] 的中译版。

SOP 不包含目标类别中的任何物体，则我们的模型将忽略该 SOP。最后，该 MIL 框架可以为每个 SOP 分配所有目标类别的概率分布并且计算语义特征向量。

通过将训练数据库中的所有图像的 SOP 视为*非终端结点*，并将所有目标类别（包括背景）视为*终端结点*，我们可以使用产生的概率分布和语义特征向量构造一个无向图。这个大图可以很好地表示训练数据库中每个 SOP 的属性以及所有 SOP 之间的关系。该无向图的最佳*多路割*可以将每个 SOP 与适当的类别标签相关联。删除 SOP 中的噪声后，带有自动分配的标签的其余 SOP 可以用作伪实例分割，以用于训练全监督的模型。由于我们的方法利用了实例、图像和数据集级别的信息，因此我们将其称为 LIID，即英文 "Leverages Instance-, Image- and Dataset-level information" 的缩写。

我们在 PASCAL VOC2012 [37] 和 MS-COCO [38] 数据集上进行了广泛的实验，以在各种实验设置下评测所提出的方法。评测结果表明，该方法在弱监督实例分割和语义分割方面都达到了最新的性能。综上所述，本文的主要贡献有三点：

- 我们提出一种新的*多实例学习*（MIL）框架，从而为每个 SOP 计算概率分布并提取语义特征向量。
- 我们使用产生的概率分布和语义特征构造一个大型无向图，其中目标类别（包括背景）被视为*终端结点*。我们进一步提出了一种有效的近似优化算法，可以对该图进行*多路割*以获得伪实例分割。
- 大量实验表明，对于弱监督实例分割和语义分割，所提出的 LIID 始终能够达到最新的性能。

2 相关工作

实例分割. 实例分割是用于场景理解的一个活跃的研究领域。长期的努力集中在全监督的条件下。大多数效果好的方法都是基于物体检测网络来输出排列好的物体分割，而不是边界框 [4], [39]–[42]。在这些方法中，Mask R-CNN [4] 及其派生方法 [41], [42] 主导了最新技术。一些研究人员还基于初始的语义分割网络提供了一些方法来生成实例蒙版 [43]–[45]。尽管全监督的方法可以实现高精度，但它们通常需要带有昂贵的逐像素标注的大规模训练数据，这使其不适用于实际应用。

弱监督实例分割. 对于弱监督实例分割，Khoreva 等人 [7] 首先提出使用边界框标注作为监督，而不是逐像素蒙版。具体来说，他们使用了 GrabCut [46] 的修改版，从其边界框估计物体的分割。通过 MCG [19] 生成的 SOP 进一步改善所获得的物体的分割。Li 等人 [8] 通过迭代改善伪真值来扩 [7]。他们使用训练集上的网络输出作为新的伪真值。Hsu 等人 [9] 通过基于每个边界框的扫掠线生成正负袋来将此问题表示为 MIL 任务。该 MIL 表示可以集成到端到端网络中，以学习实例分割模型。Hu 等人 [10] 引入了使用迁移学习的半监督实例分割模型，训练数据中的一些类具有逐像素标注，而其他类则仅有边界框标注。

Zhou 等人 [14] 提出了一个更具有挑战性的问题，即在图像级弱监督下训练神经网络进行实例分割。他们引入了一个非常新颖的*类峰值响应*概念，该响应反映了驻留在每个语义实例内部的强烈视觉信息。所学习的类峰值响应图可用于查询和排序 SOP。他们的方法明显优于各种基准方法。在 [15] 之后，Zhu 等人 [15] 提出了一种实例范围填充方法，以选择性地从有噪声的 SOP 中收集伪监督。伪监督用于学习一个可区分的填充模块，该模块可为每个实例预测一张与类无关的激活图。Cholakkal 等人 [16] 通过构造物体类别密度图，引入了一种图像级别的监督方法，以用于普通物体计数和图像级监督的实例分割。Ahn 等人 [17] 扩展了图像分类模型的 CAM，以发现被视为伪真值的整个实例区域，以训练一个全监督的模型。Ge 等人 [18] 提出的 Label-PEnet 通过交替训练四个顺序级联的模块（包括多标签分类、物体检测、实例细化和实例分割）来逐步将图像级标签转换为逐像素标签。我们遵循 [14]–[18] 仅将图像级监督用于实例分割。我们尝试同时使用每个 SOP 的固有属性和整个训练数据库的总体数据分布来确定每个 SOP 的语义类别，而不是使用基于 CAM 的模型。

弱监督语义分割. 语义分割与实例分割高度相关，因为语义分割仅识别每个像素的类别，而不会区分不同的物体实例。通过简单地消除物体实例的区分，可以将弱监督实例分割应用于语义分割。本文还提供了语义分割的评测结果，因此在这里我们概括地综述了弱监督语义分割的相关工作。

使用提供位置信息的标注，例如点 [47]、涂鸦 [48] 或边界框 [49]，最近的方法已经取得了良好的性能。使用图像级标注的弱监督语义分割仍然是一个具有挑战性的问题。给定图像级标注，CAM [22] 是发现粗略物体位置的一个很好的起点。但是，CAM [22] 倾向于将注意力集中在目标物体的小的具有区分性的区域上，这不适合于训练语义分割网络。当前大多数方法旨在改进 CAM 以仅使用图像标签提取完整的物体。这些方法要么采用图像遮挡和擦除操作，以防止分类器仅关注物体的具有区分性的部分 [24], [32], [50]，要么使用特征层面的处理 [28], [30], [51]–[54] 和区域增长技术 [31], [33], [55], [56]。这些方法经常使用各种辅助信息，例如显著性图 [57]–[61]，图像边缘 [62]–[64] 和拟物性采样 [19], [65] 以提高准确性 [23], [24], [30], [31], [56], [66]–[69]。

除上述方法外，Saleh 等人 [70] 和 Pinheiro 等人 [71] 提出了用于弱监督语义分割的 MIL 方法，但是它们的方法仅限于按像素分类，无法学习实例感知信息。相反，本文介绍的 MIL 框架着重于学习用于区分物体实例的实例感知信息。最近，Fan 等人 [29] 提出了一种基于图的弱监督语义分割模型，该模型与我们的模型相关。尽管我们专注于与 [29] 不同的任务，但我们分析了我们的方法与 [29] 之间的差异，可以将其总结如下：

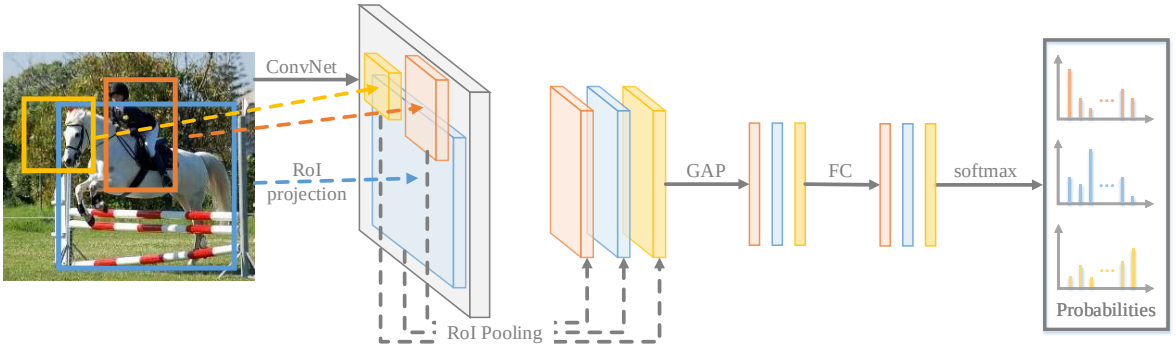

图 1. 我们针对基于 MIL 的多标签图像分类提出的网络架构。该网络旨在为每个 SOP 同时计算概率分布并提取语义特征。

1) 我们的方法在对 SOP 的信息提取方面不同于 [29]。对于每个 SOP，Fan 等人 [29] 直接使用 CAM [22] 来估计概率分布，然后采用预先训练的 ImageNet [72] 模型来提取语义特征。与此不同，我们提出了一个端到端的 MIL 框架，以便从给定图像中同时学习概率分布和语义特征。
2) 我们的方法在图建模方面不同于 [29]。范等人 [29] 通过将所有 SOP 视为图结点将类别标签分配表示为一个普通的图割问题，并且初始概率仅在优化公式中用作平衡项。与此不同，我们通过将所有 SOP 视为图的普通结点（非终端结点），将目标类别标签视为终端结点来构建无向图。SOP 的概率分布和语义特征用于计算不同类型边缘的权重。我们将类别分配表示为一个多路割问题，然后提出一种有效的近似优化算法来解决该问题。

尽管我们的方法和 [29] 都使用图来利用数据集级别的信息，但是我们提出的方法在模型训练、概率预测、特征提取、图构造和图割方面更合理、更直观，这导致了实验证明，所提出的方法的性能明显更好。

3 问题表述

假设我们有一个训练图像集 $\mathcal{I} = \{I_1, I_2, \cdots, I_N\}$ 和相应的图像级别标签 $\mathcal{Y} = \{Y_1, Y_2, \cdots, Y_N\}$，其中 N 是训练图像的数量。假设 $\mathcal{K} = \{0, 1, 2, \cdots, K\}$ 是类别集合，其中 0 表示背景，K 是目标语义类别的数量。在每个图像都有背景区域的温和假设下，我们有 $0 \in Y_i$ 和 $Y_i \subseteq \mathcal{K}$ $(i = 1, 2, \cdots, N)$。为方便起见，我们定义 $\mathcal{K}' = \{1, 2, \cdots, K\}$，不包括背景类别[1]。我们可以将图像 $\mathcal{I}$ 输入到任何自下而上的 SOP 生成方法 [19], [35], [36], [73]–[76] （此处为 MCG [19]）中去，以获得 SOP $\mathcal{S} = \{S_1, S_2, \cdots, S_N\}$。假设 $S_i = \{s_i^1, s_i^2, \cdots, s_i^{|S_i|}\}$ $(i = 1, 2, \cdots, N)$，并且 s_i^j $(j = 1, 2, \cdots, |S_i|)$ 是二进制分割蒙版。注意 $|\cdot|$ 表示一个集合中元素的数量。我们可以轻松地获得这些基于分割的 SOP 的相应边界框，它们可以表示为 $\mathcal{B} = \{B_1, B_2, \cdots, B_N\}$，其中 $B_i = \{b_i^1, b_i^2, \cdots, b_i^{|S_i|}\}$。

1. 为清楚起见，我们用 $k \in \mathcal{K}$ 中使用 $k' \in \mathcal{K}'$ 分别代表包括背景和不包括背景的类别。

这些与类别无关的 SOP 可能不包含任何语义物体、多个或一个语义物体。不包含完整语义物体和多个物体的 SOP 在本文中被认为是噪声。为了进行实例分割，本文的主要目的是删除 SOP 中的噪声，并为紧密包含一个完整物体的 SOP 分配正确的类别标签。因此，我们的目标可以表述为

$$F(s_i^j) = \begin{cases} 0 & \text{如果 } s_i^j \text{ 是一个噪声采样} \\ k' & \text{如果 } s_i^j \text{ 属于类别 } k' \end{cases}, \quad (1)$$

其中 $k' \in \mathcal{K}'$，s_i^j 表示第 i 张图像中的第 j 个 SOP。具有 $F(s_i^j) > 0$ 的采样 s_i^j 将作为我们的伪实例分割。图2中展示了所提出的用于计算 $F(s_i^j)$ 的解决方案的概述。

4 基于 SOP 的 MIL 框架

给定具有图像级别标签的图像，以前的研究 [14]–[16] 通常会训练用于为目标定位计算 CAM 的多标签图像分类器。然后，他们将 CAM 和 SOP 结合起来以产生伪分割。由于 CAM 的天然局限性，如上所述，训练数据没有得到充分利用。与此不同，我们考虑将 SOP 纳入训练过程，使得每个 SOP 都能学习有用的信息。给定带有图像标签 Y_i 的输入图像 I_i，我们将知道相应的物体推荐 S_i/B_i 包含类别 Y_i，但每个物体推荐分别对应于哪个类别未知。这实际上是多实例学习（Multiple Instance Learning, MIL）的一种情况。因此，我们建立了一个 MIL 框架，该框架以图像和 SOP 作为输入，并以图像级别的标签作为监督。通过训练，该模型有望学习为每个 SOP 生成类概率分布和语义特征向量，并将其用于后续的多路割。在本节中，我们首先介绍所提出的网络架构，然后介绍基于 SOP 的 MIL 框架的几种损失函数。

4.1 网络架构

在这一部分中，我们介绍为基于 MIL 的多标签图像分类而设计的卷积神经网络。所提出的网络架构展示在图1中。在这里，与类别无关的物体推荐是由 MCG 算法 [19] 生成的。输入图像 I_i 首先通过骨干网络，即 ResNet50 [21]。我们使用 SOPS_i 的边界框 B_i 对生成的特征图执行 ROI 池化 [2]。ROI 池化之后跟随一个全局平均池化（GAP）层来将每一个 SOP 对应的

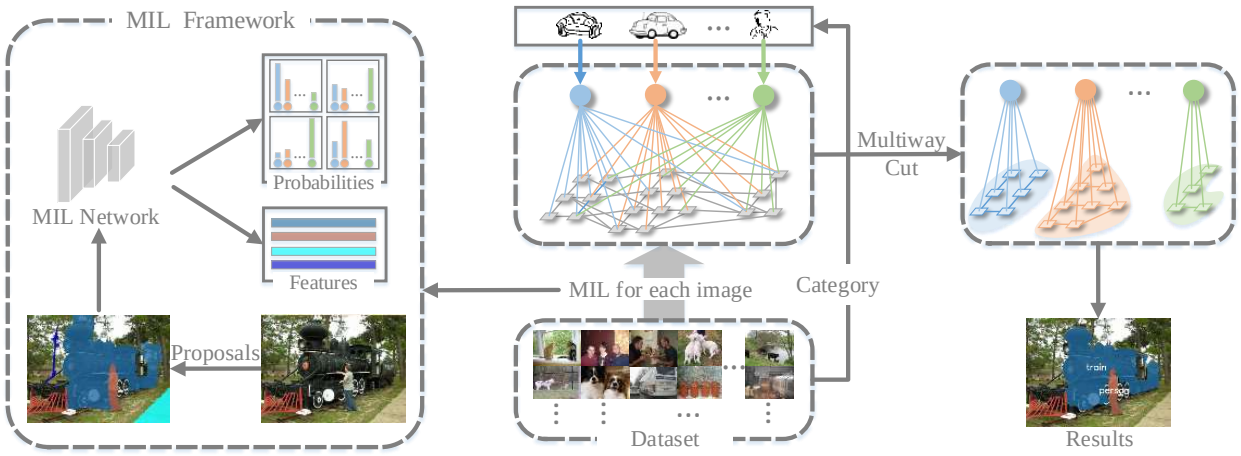

图 2. 本文所提出的方法的概述。带有图像级别标签的训练图像用于训练我们基于 MIL 的多标签图像分类网络，如第4节所示。所有训练图像以及相应的 SOP 都被输入到 MIL 网络中，以计算概率分布和语义特征。然后，使用所有训练图像构造一个大型知识图。伪实例分割可以使用改进的多路割算法来获得。

特征图转换为一个 2048 维的特征向量 f_i^j $(j = 1, 2, \cdots, |S_i|)$。然后，我们连接一个全连接层，该全连接层有 $(K+1)$ 个输出 a_i^j $(|\mathrm{a}_i^j| = K+1)$，代表 K 个目标类别以及背景的预测分数。最后，令 $(\mathrm{p}_i^j)_k$ 为在一个 softmax 层之后获得的类别 k 的概率，因而我们有

$$(\mathrm{p}_i^j)_k = \frac{exp((\mathrm{a}_i^j)_k)}{\sum_{m=0}^{K} exp((\mathrm{a}_i^j)_m)}, \tag{2}$$

其中 $k \in \mathcal{K}$。通过这种设计模式，我们可以为每个 SOP 计算特征向量 f_i^j 和概率分布 p_i^j。通过使用适当的损失函数，所提出的神经网络将会为每个 SOP 学习类别感知信息。

4.2 基于 SOP 的 MIL 损失函数

对于 MIL 框架的训练，我们提出了几个损失函数以同时推断概率分布并提取 SOP 的语义特征。考虑到 SOP 的标签未知，我们设计了一个基于 CAM 的损失函数来估计每个 SOP 的伪标签，并且我们还设计了一个基于 MIL 的图像分类损失函数来计算每个图像的汇总的概率，以便我们可以采用图像标签作为监督。通过施加这些损失函数来监督概率分布 p_i^j，从而使网络收敛。此外，我们设计了一种基于 MIL 的中心损失函数，以将语义特征向量 f_i^j 集中在同一类别上，这样，属于同一类别的 SOP 的特征向量 f_i^j 将具有较小的特征距离。

4.2.1 基于 CAM 的损失函数

我们不再像之前的方法 [14]–[16] 那样依赖 CAM 来定位物体，而是通过用 CAM 为每个 SOP 估计伪标签，来将 CAM 用作训练的监督源之一。具体来说，使用具有 K 个独立的交叉熵损失函数的标准 ResNet50 [21] 网络，我们可以训练一个多标签图像分类模型。然后，我们可以使用著名的 CAM 算法 [22] 计算图像 I_i 的 CAM $A_i^{k'}$ $(k' \in \mathcal{K}')$。$A_i^{k'}$ 被标准化为 $[0, 1]$ 的范围。令 $\widetilde{y}_i^j$ 表示第 j 个 SOP 的估计类别标签 $(j = 1, 2, \cdots, |S_i|)$。假设我们有 $(R_i^j)_{k'} = mean(A_i^{k'}[b_i^j]) + max(A_i^{k'}[b_i^j])$ 且 $(R_i^j)_{k'} \in [0, 2]$，其中 $A_i^{k'}[b_i^j]$ 表示 SOP b_i^j 在 $A_i^{k'} \in [0, 1]$ 中的对应区域。我们使用计算的 CAM 来估计 $\widetilde{y}_i^j$，如下所示

$$\widetilde{y}_i^j = \begin{cases} 0 & if \ \ \forall k', (R_i^j)_{k'} < \eta \\ \arg\max\limits_{k'} \ (R_i^j)_{k'} & otherwise \end{cases}, \tag{3}$$

其中 η 是一个阈值。因此，可以将 $\widetilde{y}_i^j$ 视为第 j 个 SOP b_i^j 的伪标签，而 $\mathcal{K} \backslash \{\widetilde{y}_i^j\}$ 是除了 $\widetilde{y}_i^j$ 以外的类别集合。我们将基于 CAM 的损失函数定义为

$$\begin{aligned} L_{\mathrm{Att}}^{(i)} = -\frac{1}{|S_i|} \sum_{j=1}^{|S_i|} \Bigg[& log(\mathrm{p}_i^j)_{\widetilde{y}_i^j} \\ & + \frac{1}{K} \sum_{k \in \mathcal{K} \backslash \{\widetilde{y}_i^j\}} log(1 - (\mathrm{p}_i^j)_k) \Bigg]. \end{aligned} \tag{4}$$

通过这种方式，预训练的多标签图像分类模型可以通过 CAM 来帮助所提出的 MIL 框架的训练。对于 $(R_i^j)_{k'}$ 的计算，我们在 $A_i^{k'}[b_i^j]$ 中使用边界框池化而不是蒙版池化，因为基于边界框的物体推荐往往比基于分割的蒙版（即 SOP）更可靠。如相关研究 [19], [35], [36], [73] 中所描述的那样，边界框的生成比蒙版生成要容易得多，因此可以实现更高的准确性。自下而上的方法很难准确地分割物体，而且不准确的蒙版将会对 MIL 的训练有害。我们将通过消融实验在第6.2节中进一步证明这种设计的合理性。

4.2.2 基于 MIL 的图像分类损失函数

尽管 SOP 的标签是未知的，但是图像中所有 SOP 的学习到的概率分布 p_i^j 的聚合可以反映网络的分类能力。换句话说，虽然我们不能直接监督每个 SOP 的概率 p_i^j，但是我们可以监督一张图像的总体的概率聚合。假设图像 I_i 中每个类的聚合概率为 $(\mathrm{Z}_i)_k$ $(k \in \mathcal{K})$，可以从 $(\mathrm{p}_i^j)_k$ 推断得到。我们使用 Log-Sum-Exp（LSE）函数 [77] 来计算 $(\mathrm{p}_i^j)_k$ $(j = 1, 2, \cdots, |S_i|)$

的平滑的最大值的估计，而不是 $(\mathrm{p}_i^j)_k$ 的简单的最大值或平均值，这可以表示为

$$(\mathrm{Z}_i)_k = \frac{1}{r} log\Big[\frac{1}{|S_i|}\sum_{j=1}^{|S_i|} exp(r\ (\mathrm{p}_i^j)_k)\Big], \tag{5}$$

其中 r 是一个使得 LSE 函数的表现介于最大值和平均值之间的参数。在本文中，我们根据经验将 r 设置为 5 [71]。与简单的最大值相比，LSE 函数不仅可以估计其最大值，而且可以考虑到 $(\mathrm{p}_i^j)_k$ 的所有元素。通过估计的 $(\mathrm{Z}_i)_k$，我们将基于 MIL 的图像分类损失函数定义为

$$L_{\mathrm{MIL}}^{(i)} = -\frac{1}{|Y_i|}\sum_{k\in Y_i} log((\mathrm{Z}_i)_k) - \frac{1}{|\overline{Y_i}|}\sum_{k\in\overline{Y_i}} log(1-(\mathrm{Z}_i)_k), \tag{6}$$

其中 $\overline{Y_i}$ 是 Y_i 的补集。符合直觉的是，当前类别应出现在 SOP 中，而对缺席类别具有高概率的 SOP 应受到惩罚。

如第3节中所述，我们假定每个图像都有背景区域，即 $0 \in Y_i$ $(i = 1, 2, \cdots, N)$。这个温和的假设对于公式 (6) 至关重要。一方面，自下而上算法生成的 SOP 通常包含许多噪声，这些噪声 SOP 并不在目标物体类别中，涵盖其他物体类别甚至非物体区域，因此我们必须为每张图像包括背景类，以确保神经网络的训练。另一方面，我们的目标要求按照公式 (1) 识别并过滤掉这些噪声 SOP，因此我们必须将背景类别纳入训练中，以学习有关噪声 SOP 的适当信息，然后使用第5节中的技术过滤掉它们。

4.2.3 基于 MIL 的中心损失函数

下一个损失函数被设计用于语义特征提取。我们期望该训练能使具有相同类别的 SOP 的语义特征的相似性最大化，并且使具有不同类别的 SOP 的相似性最小化。为此，我们引入了基于 MIL 的中心损失函数，以集中具有相似语义的语义特征：

$$\begin{aligned} \widehat{y}_i^j &= \arg\max_k\ (\mathrm{p}_i^j)_k, \\ L_{\mathrm{Cent}}^{(i)} &= \frac{1}{|S_i|}\sum_{j=1}^{|S_i|}\Big[1 - \frac{\mathrm{f}_i^j \cdot \mathrm{c}_{\widehat{y}_i^j}}{\|\mathrm{f}_i^j\|_2 \|\mathrm{c}_{\widehat{y}_i^j}\|_2}\Big], \end{aligned} \tag{7}$$

其中 c_k 是学习到的第 k 类的输入样本的中心，而 $\|\cdot\|_2$ 表示向量的 ℓ^2 范数。该损失度量了特征向量 f_i^j 与学习到的类别中 c_k 之间的 cosine 相似度。在每次训练迭代中，根据语义特征向量 f_i^j 将 c_k 更新为

$$\begin{aligned} \mathrm{c}_{\widehat{y}_i^j}^{new} &= \mathrm{c}_{\widehat{y}_i^j}^{old} + \theta \cdot (\mathrm{f}_i^j - \mathrm{c}_{\widehat{y}_i^j}^{old}), \\ &\text{for } j = 1, 2, \cdots, |S_i|, \end{aligned} \tag{8}$$

其中 θ 是更新速率。因此，任意两个 SOP 之间的相似距离可以通过它们学习到的特征向量 f_i^j 来计算。

通过以上定义，可以通过以下公式来表示基于 MIL 的多标签图像分类问题的总体损失函数：

$$L^{(i)} = \alpha L_{\mathrm{Att}}^{(i)} + \beta L_{\mathrm{MIL}}^{(i)} + \gamma L_{\mathrm{Cent}}^{(i)}. \tag{9}$$

实际上，我们根据经验将 α，β 和 γ 分别设置为 0.5、0.5 和 0.1。我们所提出的 $L_{\mathrm{Att}}^{(i)}$ 可以利用预先训练的多标签图像分类模型来帮助 MIL 的训练，而 $L_{\mathrm{MIL}}^{(i)}$ 自然地适合此处的 MIL 训练。因此，$L_{\mathrm{Att}}^{(i)}$ 和 $L_{\mathrm{MIL}}^{(i)}$ 的系数被均设置为 0.5。对于损失 $L_{\mathrm{Cent}}^{(i)}$，其目的是最大程度地减少类内差异，这与图像分类无关，因此我们为其设置一个小的系数 0.1 以避免它对分类结果的影响。

5 基于多路割的标签分配

直观地，考虑所有训练样本的数据分布的预测将比仅考虑单个样本的预测更好。这是因为单个样本可能有偏差或随机误差，但总体数据分布更为可靠。尽管 MIL 框架的训练过程已经利用了所有训练数据，但这只是整体数据分布的间接使用。在这里，我们考虑一种直接的方法。具体来说，我们利用一个庞大的知识图，其中包括所有训练图像中的 SOP，以提供一种全局的解决方案。

5.1 多路割问题的回顾

在介绍我们用于 SOP 标签分配的方法之前，我们将在本部分中简要回顾多路割问题。让我们首先描述传统的图割。假设我们有一个连通的无向图 $G = (V, E)$，其中结点集为 V，边集为 E。该图 G 的权重函数可以表示为 $w : E \to \mathbb{R}^+$，其中 $\mathbb{R}^+$ 表示非负实数的集合。交换属性适用于任何结点对 $u \in V$，$u \in V$，即 $w(u, v) = w(v, u)$。通过将 V 划分为不相交的子集 V_1 和 V_2 来定义图割，从而得到边集的子集 $E' \subseteq E$，其中每条边在 V_1 中有一个顶点，另一个顶点在 V_2 中。因此，边的子集 E' 可用于表示该图割。该图割的代价定义为 $\sum_{(u,v)\in E'} w(u, v)$。典型的最小割问题是找到将两个给定结点 $\dot{u}$ 和 $\dot{v}$（我们称这些结点为 终端结点）分开的具有最小代价的切割，即 $\dot{u} \in V_1$ 和 $\dot{v} \in V_2$。这个最小割问题是最大流问题的对偶，可以在多项式时间内解决。

多路割问题是最小割问题的一种泛化，也被称为多终端切割问题 [78]–[80]。给定一组终端结点 $\hat{E} \subseteq E$，多路割是找到具有最小代价的边的子集 $E' \subseteq E$，删除 E' 将使得终端结点相互隔绝。换句话说，图 $(V, E - E')$ 的任何连通子图都不可能包含 $\hat{E}$ 中的两个终端结点。当只有两个终端，即 $|\hat{E}| = 2$ 时，此问题等效于上述在多项式时间内可解决的最小割问题。当存在三个或更多终端，即 $|\hat{E}| \geq 3$ 时，多路割成为 NP 难问题，需要近似算法来解决。在下面的小节中，我们将 SOP 标签分配表示为多路割问题，并提出了解决该问题的一种简单方案。

5.2 知识图的构造

为了计算公式 (1) 中的 $F(s_i^j)$，我们构造了一个大知识图，该图不仅包含每个 SOP 的内在属性，而且还包含整个训练数据库中不同 SOP 之间的关系。我们使用所有训练图像来构造该图。利用该知识图将为每个 SOP 分配一个可靠的类别标签。我们将标签分配过程表示为一个多路割问题，并为该问题引

入了一个有效的近似解决方案。训练图像的图割结果将作为我们的伪实例分割，可用于训练全监督的模型。

如第5.1节中所述，我们构造了一个连通的无向图 $G = (V, E)$。具体来说，我们将所有 SOP s_i^j $(i = 1, 2, \cdots, N; j = 1, 2, \cdots, |S_i|)$ 和目标类别 $\mathcal{K}$ $(\mathcal{K} = \{0, 1, 2, \cdots, K\})$ 看作图的结点，因此我们有 $V = \mathcal{K} \cup S_1 \cup S_2 \cup \cdots \cup S_N$。此外，令 $\mathcal{K}$ 为终端结点的集合，即 $\hat{E} = \mathcal{K}$。每条边 $(u, v) \in E$ 有一个非负权重

$$w(u,v) = \begin{cases} (\mathrm{p}_i^j)_k & if\ \exists i,j\ \ u = s_i^j; v \in \mathcal{K} \\ 0 & if\ u \in \mathcal{K}, v \in \mathcal{K} \\ \delta \cdot \dfrac{|\mathrm{f}_i^j \cdot \mathrm{f}_{i'}^{j'}|}{\|\mathrm{f}_i^j\|_2 \|\mathrm{f}_{i'}^{j'}\|_2} & if\ \exists i,j\ \ u = s_i^j;\ \exists i', j'\ \ v = s_{i'}^{j'} \end{cases}, \tag{10}$$

其中 δ 是一个平衡因子。因此，终端结点之间的边缘权重为 0，这是知识图 G 中边的最小权重。SOP 结点和终端结点之间的边的权重就是该 SOP 属于相应类别的预测概率。两个 SOP 结点之间的边的权重是其特征向量的 cosine 相似性 [29], [81]，因此具有相似语义内容的 SOP 对将具有较大的 cosine 相似性。通过这种方式，图 G 通过合并在第4节中学习的所有训练图像的概率分布和语义特征，包含了整个训练数据库的知识。

5.3 知识图上的多路割

给定带有一组终端结点 $\mathcal{K}$ 的知识图 $G = (V, E)$，我们的目标是找到一种多路割方法，以断开每个终端结点与其余终端结点的连接。也就是说，我们的主要目标是找到具有最小代价的边的子集 $E' \subseteq E$，以便在新图 $(V, E - E')$ 中，任何两个终端结点之间没有连通的路径。经过多路割后，具有相似语义信息的 SOP 的对应结点将落入同一子图中，因为上述多路割通过使图割代价最小化，已使每个子图内的相似度最大化，并使不同子图之间的相似度最小化。每个子图中只有一个终端结点 $k \in \mathcal{K}$，每个 SOP 的伪类别标签就是其对应子图中的终端结点 k。此处，类别 $k = 0$ 表示背景或噪声 SOP，因为它不属于目标物体类别。

常用数据集，例如 PASCAL VOC2012 [37] 和 MS-COCO [38] 通常具有 $|\mathcal{K}| \geq 3$，即存在三个或更多的物体类别。如第5.1节中所讨论的，我们需要一种近似算法来解决上述多路割问题。假设 Δ_K 表示 K-单纯形，因此 $\mathbb{R}^{K+1}$ 中的 K 维凸多面体可以表示为 $\{x \in \mathbb{R}^{K+1} | (x \geq 0) \wedge \sum_k x_k = 1\}$。对于 $k, \dot{k} \in \mathcal{K}$，$e^k \in \mathbb{R}^{K+1}$ 表示单位向量，即 $(e^k)_k = 1$ 且 $(e^k)_{\dot{k}} = 0$ $(\forall k \neq \dot{k})$。根据 [80]，我们可以制定以下优化函数来解决多路割问题

$$\begin{aligned} \min_x \frac{1}{2} \sum_{(u,v) \in E} w(u,v) \cdot \|x^u - x^v\|_1 \quad s.t. \\ x^u \in \Delta_K,\ \ \forall u \in V; \\ x^k = e^k,\ \ \forall k \in \mathcal{K}, \end{aligned} \tag{11}$$

其中 $\|\cdot\|_1$ 表示 ℓ^1 范数。但是，由于指数级数量的约束 [80]，直接求解 公式 (11) 中的线性规划是不切实际的，尤其是在我们整个训练数据库上的知识图非常大的情况下。直接的解决方案所需的 CPU 内存和运行时间对于现有设备来说是不可行的。具体来说，直接的解决方案的空间复杂度为 $O(|E||V|^2)$，PASCAL VOC2012 [37] 训练集所需的 CPU 内存约为 $10^3 \sim 10^4$ GB，比现有计算机的存储容量大得多。

为了解决这个问题，我们将每个结点 $u \in (V - \mathcal{K})$ 连接到最多三个具有最大的边的权重的其他结点 v $(v \in (V - \mathcal{K})$ 且 $v \neq u)$，而不是将每个结点 $u \in V$ 连接到所有其他节点。我们观察到，在获得的稀疏图中，整个大型知识图将自动划分为许多小的相互不连通的子图，每个子图可以记作 $G_t = (V_t, E_t)$：

$$\begin{aligned} \cup_t V_t = V, \\ \cup_t E_t = E. \end{aligned} \tag{12}$$

在多路割问题中，每个子图彼此独立。可以通过将公式 (11) 分解为许多项来轻松证明这一点，所分解的每个项代表子图的多路割的代价。这些分解项中的公共图结点仅是终端结点，这不会影响最终的图割结果，因为这些终端结点最终必须落入不同的图割中去。因此，我们可以单独处理每个子图，以计算其多路割 E'_t。为了解决此线性规划问题，我们首先使用单纯形方法来求解公式 (11)，其结果将使用 IBM-CPLEX [82] 的分支定界法进一步转换为多路割的结果。原始大图的多路割 E' 可以通过下式求得

$$\cup_t E'_t = E'. \tag{13}$$

通过这种方式，我们可以通过计算许多小图来成功地估计大图的多路割。在这里，我们选择为每个节点连接三条边，是因为为每个节点连接四条边将会导致子图过大，如上所述，子图也将很难求解。有了多路割结果，如果 SOP s_i^j 与终端结点 k $(k \in \mathcal{K})$ 属于同一子集，我们可以轻松地将公式 (1) 中的 $F(s_i^j)$ 分配给类别 k。如果 $F(s_i^j) = 0$，那么 SOP s_i^j 将是噪声，因此将被丢弃。对于 $F(s_i^j) \neq 0$ 的其余 SOP，我们使用对应的边界框 b_i^j 应用非极大值抑制（Non-Maximum Suppression, NMS），非极大值抑制的重叠率（Intersection-over-Union, IoU）阈值为 0.4，就像物体检测领域中常做的那样 [2]–[4]。此 NMS 操作解决了多个 SOP 代表同一个物体的情况。最后，我们将其余的 SOP 和相应的类别标签 $F(s_i^j)$ 作为训练图像的伪真值，以便我们可以训练 Mask R-CNN 模型 [4]（使用 ResNet50 [21] 作为骨干网络）用于弱监督实例分割，或训练 DeepLab 模型 [83]（使用 ResNet101 [21] 作为骨干网络）用于弱监督语义分割。

6 实验

6.1 实验设置

数据集. 我们在 PASCAL VOC2012 数据集 [37] 和 MS-COCO 数据集 [38] 上对提出的方法进行了评测。请注意，仅图像级标签被用于训练。VOC2012 数据集 [37] 包含 20 个语义类别以及背景类别。我们遵循 [14]–[16] 来利用 VOC2012 main trainval

表 1
在 VOC2012 segmentation train 数据集 [37] 上对不同的 θ 值（在公式 (8) 中）和 γ 值（公式 (9) 中）的评测结果。每一对结果 w_1/w_2 分别表示没有（w_1）和有（w_2）知识图的结果。

编号	θ	γ	AP_{50}	AP_{75}	ABO
1	0.01	0.05	30.3/33.8	14.9/16.3	36.7/38.7
2	0.01	0.1	32.5/34.8	15.5/16.7	38.2/39.4
3	0.01	0.5	32.4/33.7	15.1/16.1	37.9/39.2
4	0.03	0.1	32.0/33.7	14.9/16.0	38.0/39.3
5	0.03	0.3	31.9/33.7	14.8/16.3	37.6/39.3
6	0.05	0.1	32.3/33.6	15.1/16.3	37.8/39.2
7	0.005	0.5	29.1/32.3	13.6/15.5	36.3/38.5
8	0.05	0.5	31.5/33.2	14.9/16.2	37.7/39.0
9	-	0	31.3/-	15.0/-	37.8/-

子集（不包括 segmentation val 中的图像）来训练我们的 MIL 框架（大约 10K 图像）。我们使用 1449 张 segmentation val 中的图片来评测我们的方法和基准模型。对于消融实验，我们采用 VOC2012 main trainval 的子集（不包括 segmentation train+val 中的图像）进行训练，并采用 segmentation train 进行验证。MS-COCO 数据集 [38] 包含 80 个语义类别。我们遵循 [29] 在标准的 trainval 集上进行训练，并在 test-dev 集上进行评测。

实现细节. 在训练中，我们采用自下而上的 MCG [19] 算法为每张图像生成 500 个 SOP，然后使用 [84] 中的简单过滤方法从中为 VOC2012/MS-COCO 选择 20/40 个 SOP。我们使用 PyTorch 框架来实现基于 MIL 的多标签图像分类模型。我们将 SGD 优化算法与 step 学习率策略一起应用。对于 VOC2012 和 MS-COCO 数据集，初始学习率均为 5×10^{-4}，在 5 个纪元后将其除以 10。我们使用一张图像的小批量运行 SGD，总共运行 10 个纪元。权重衰减率和动量分别设置为 10^{-4} 和 0.9。在建图过程中，我们遵循 [29] 来计算显著性实例 [76] 来作为 SOP。Mask R-CNN [4] 和 DeepLab [83] 的训练遵循默认设置。

评测指标. 对于实例分割的评测指标，我们遵循 [14] 来采用在 IoU 阈值 0.5（AP_{50}）和 0.75（AP_{75}）下的基于蒙版的平均精度（Average Precision, AP）指标（详情请参见 [38]），以及另一个视角下的平均最佳重叠（Average Best Overlap, ABO）指标（详情请参见 [35]）。

6.2 消融实验

在与其他方法进行比较之前，我们进行了一些消融实验，以评测不同设计选择和参数设置的有效性。如上所述，所有消融实验都是针对 VOC2012 segmentation train 数据集 [37] 上的弱监督实例分割进行的。在这里，如果没有提及，我们不会训练 Mask R-CNN [4] 以节省时间。调整每组超参数时，其他参数将保持为默认值。

表 2
在 VOC2012 segmentation train 数据集 [37] 上对不同的 α 值和 β 值（在公式 (9) 中）的评测结果。每一对结果 w_1/w_2 分别表示没有（w_1）和有（w_2）知识图的结果。

编号	α	β	AP_{50}	AP_{75}	ABO
1	1.0	0.0	28.7/31.8	13.9/15.7	35.5/37.9
2	0.8	0.2	31.5/34.0	14.7/16.6	37.4/39.2
3	0.5	0.5	32.5/34.8	15.5/16.7	38.2/39.4
4	0.2	0.8	31.3/32.8	14.8/16.1	36.2/37.6
5	0.0	1.0	18.7/19.3	8.8/9.2	22.9/23.0

表 3
在 VOC2012 segmentation train 数据集 [37] 上对 $(R_i^j)_{k'}$ 中 $mean$ 和 max 的存在与否（在公式 (3) 中）的评测。每一对结果 w_1/w_2 分别表示没有（w_1）和有（w_2）知识图的结果。

编号	mean	max	AP_{50}	AP_{75}	ABO
1	✔	✗	28.7/32.6	13.1/15.5	34.3/37.7
2	✗	✔	32.4/33.6	15.1/16.1	37.7/38.7
3	✔	✔	32.5/34.8	15.5/16.7	38.2/39.4

表 4
在 VOC2012 segmentation train 数据集 [37] 上对公式 (3) 中计算 $(R_i^j)_{k'}$ 时是用边界框池化还是蒙版池化的评测。每一对结果 w_1/w_2 分别表示没有（w_1）和有（w_2）知识图的结果。

编号	Proposal types	AP_{50}	AP_{75}	ABO
1	Box	32.5/34.8	15.5/16.7	38.2/39.4
2	Mask	30.7/32.4	14.5/15.7	36.8/38.0

表 5
在 VOC2012 segmentation train 数据集 [37] 上对不同 η 值（在公式 (3) 中）的评测。每一对结果 w_1/w_2 分别表示没有（w_1）和有（w_2）知识图的结果。

编号	η	AP_{50}	AP_{75}	ABO
1	0.25	28.3/30.8	13.7/15.3	34.6/36.5
2	0.50	30.0/33.4	14.2/16.0	36.5/38.7
3	0.75	32.5/34.8	15.5/16.7	38.2/39.4
4	1.00	30.1/32.5	14.5/16.0	36.4/38.3

中心损失函数 $L_{\text{Cent}}^{(i)}$ 的超参设置. 中心损失函数旨在聚合特征向量 $\mathbf{f}_i^j$。超参数 θ（在公式 (8) 中）控制每个类别的中心特征向量的更新速度，而参数 γ（在公式 (9) 中）控制其对骨干网络的影响。表1中展示了 θ 和 γ 的不同设置和相应结果。当我们有 $\gamma = 0$ 时，将省略参数 θ 和相应的知识图（表1中的第 9 号）。我们可以看到，此设置的结果比没有知识图的最佳设置要差，这表明基于 MIL 的中心损失函数（第4.2.3节）不仅对于知识图的构建是必要的，而且对 MIL 框架的训练很有帮助。当我们有 $\gamma \neq 0$ 时，θ 和 γ 似乎对不同的值不敏感。$\theta = 0.01$ 和 $\gamma = 0.1$ 的设置可获得更好的性能。因此，我们分别使用 0.01 和 0.1 作为 θ 和 γ 的默认值。

表 6
在 VOC2012 segmentation train 数据集 [37] 上对不同 δ 值（在公式 (10) 中）的评测结果。

编号	δ	AP_{50}	AP_{75}	ABO
1	1	30.3	14.9	37.0
2	2	34.6	16.3	39.3
3	3	34.8	16.7	39.4
4	5	34.8	16.7	39.4
5	10	34.8	16.6	39.4

表 7
在 VOC2012 segmentation train 数据集 [37] 上对 LIID 的上界的评测。LIID 的上界使用标注的边界框来过滤和标记 SOP。

编号	GT boxes (Oracle)	AP_{50}	AP_{75}	ABO
1	✗	34.8	16.7	39.4
2	✔	44.9	23.0	39.1

损失函数 $L^{(i)}_{\text{Att}}$ 和 $L^{(i)}_{\text{MIL}}$ 的平衡因子. 我们还评测了公式 (9) 中损失函数 $L^{(i)}_{\text{Att}}$ 和 $L^{(i)}_{\text{MIL}}$ 的平衡因子 α 和 β 的效果。结果显示在表2中。我们可以看到，$L^{(i)}_{\text{Att}}$ 和 $L^{(i)}_{\text{MIL}}$ 对最后的实例分割的贡献都很大。当 $\alpha = 0.5$ 且 $\beta = 0.5$ 时，所提出的方法效果最佳，因此我们将此设置用作默认设置。

$(R^j_i)_{k'}$ 的 $mean$ 和 max 项. 在公式 (3) 中，我们定义了一个辅助项 $(R^j_i)_{k'} = mean(A^{k'}_i[b^j_i]) + max(A^{k'}_i[b^j_i])$ 以估计类别标签 $\widetilde{y}^j_i$，它将在公式 (4) 中用于计算 $L^{(i)}_{\text{Att}}$。在表3中，我们仅使用 $(R^j_i)_{k'}$ 的 $mean$ 项，仅 max 项、以及同时使用 $mean$ 和 max 项进行 MIL 训练。第三个实验明显胜过其他两个。

$(R^j_i)_{k'}$ 的边界框或蒙版池化. 在第4.2.1节中，我们直观地分析了为什么在公式 (3) 中使用边界框池化而不是蒙版池化来计算 $(R^j_i)_{k'}$ 的原因。在这里，我们在 VOC2012 segmentation train/val 数据集 [37] 上进行实验以验证边界框池化相对于蒙版池化的优越性。结果显示在表4和表8中（第 6 号）。我们可以观察到，蒙版级别的池化会导致性能显著下降，这可能是因为不准确的 SOP 损害了 MIL 框架的训练。

$(R^j_i)_{k'}$ 的阈值 η. 在表5中，我们对公式 (3) 中的 $(R^j_i)_{k'}$ 应用不同的阈值 η。尽管我们有 $\eta \in [0, 2]$，但是我们仅测试 $\eta \leq 1.00$，因为 $\eta \geq 0.75$ 会导致性能显著下降。阈值 0.75 表现最佳，因此我们将其用作默认设置。

知识图的有效性. 在第5节中，我们使用 MIL 框架的输出来构造一个知识图，该知识图的多路割可以为相应的 SOP 分配类别标签。如果没有知识图，我们还可以使用 MIL 所学的概率来标记 SOP。在表1 - 表5中，我们报告了多路割前后的结果。知识图可以在所有情况下提高性能。因此，我们可以得出结论，知识图对于我们的系统至关重要。

平衡因子 δ. 在公式 (10) 中，我们使用平衡因子 δ 来控制特征的 cosine 相似度对图的边权重的贡献。在表6中，我们研究了不同的 δ 值的影响。当 $\delta \geq 2$ 时，我们获得了类似的结果。根据结果，我们将 δ 设置为 5 作为默认值，因为 $\delta = 5$ 具有略好的性能。

表 8
在 VOC2012 segmentation val 数据集 [37] 上对 Mask R-CNN 训练之后的 LIID 每个组件的评测。符号✗表示删除 LIID 中的一个组件。第一行（编号 1）是 LIID 的默认版本。

编号	Strategy	AP_{50}	AP_{75}	ABO
1	-	48.4	24.9	50.8
2	CAM-Based Loss $L^{(i)}_{\text{Att}}$ ✗	38.3	17.1	45.4
3	MIL Loss $L^{(i)}_{\text{MIL}}$ ✗	46.9	24.1	48.1
4	Center Loss $L^{(i)}_{\text{Cent}}$ ✗	45.8	23.0	48.6
5	Knowledge Graph ✗	46.1	22.8	48.1
6	$(R^j_i)_{k'}$ (Box → Mask)	45.2	22.9	48.9

关于 CAM 的讨论. 如果我们为公式 (9) 设置 $\alpha = 1.0$ 和 $\beta = 0.0$ 而不使用第5节中的知识图，则模型将退化为仅依赖 CAM 进行训练。在表2中，我们可以看到，按 AP_{50}、AP_{75} 和 ABO 计，结果分别为 28.7%、13.9% 和 35.5%。使用我们的其他设计，就 AP_{50}、AP_{75} 和 ABO 而言，结果分别提高到了 34.8%、16.7% 和 39.4%。请注意，我们模型的这个基于 CAM 的简单变体还包括一些我们的有效设计，如表3 - 表5所示。因此，我们的系统并不是直截了当的。

LIID 的上界. 我们还使用标注的真值边界框过滤和标记 SOP 来评测 LIID 的上限。具体来说，如果一个 SOP 的边界框与任何一个真值边界框的 IoU 大于 0.5，则保留该 SOP，并且为其分配与具有最大 IoU 的真值边界框相同的标签。否则，该 SOP 被丢弃。我们在表7中展示了实验结果。LIID 和其上界之间存在很大的性能差距，为将来的改进留有余地。

Mask R-CNN 训练之后的每个组件. 我们继续在 VOC2012 segmentation val 数据集 [37] 上评测每个组件在 Mask R-CNN 训练之后的影响。具体来说，我们逐个忽略每个分量，如损失函数或多路割，然后采用生成的伪实例分割来训练 Mask R-CNN。结果汇总在表8中。我们可以观察到，LIID 的每个组件都会对最终性能产生重大影响，因为删除任何组件都会导致性能大幅下降。

6.3 VOC2012 上的实例分割

由于仅由图像级别监督的弱监督实例分割是 Zhou 等人 [14] 最近提出的问题，所以以前对此问题的研究非常有限 [14]–[18].。因此，我们遵循 [14] 来基于一些弱监督的物体定位模型 [11], [22], [85] 生成的边界框构建一些基准模型。为了获得实例分割，我们应用了三种简单的蒙版提取策略：i）矩形（Rect），即仅使用边界框作为分割结果；ii）椭圆（Ellipse），即仅填充每个边界框所包含的最大椭圆；iii）MCG，即为每

图 3. PASCAL VOC2012 segmentation val 数据集 [37] 上的实例分割的定性结果。

表 9
在 VOC2012 segmentation val 数据集 [37] 上比较我们的方法和其他弱监督实例分割模型。浅色方法 [9] 使用边界框作为监督，而其他方法仅使用图像级标签作为监督。

Method		AP_{50}	AP_{75}	ABO
CAM [22]	Rect.	2.5	0.1	18.9
	Ellipse	3.9	0.1	20.8
	MCG	7.8	2.5	23.0
SPN [85]	Rect.	5.2	0.3	23.0
	Ellipse	6.1	0.3	24.0
	MCG	12.7	4.4	27.1
MELM [11]	Rect.	14.6	1.9	26.4
	Ellipse	19.3	2.4	27.0
	MCG	22.9	8.4	32.9
PRM [14]		26.8	9.0	37.6
IAM-S5 [15]		28.8	11.9	41.9
Cholakkal et al. [16]		30.2	14.4	44.3
Ahn et al. [17]		46.7	17.4	-
Hsu et al. [9]		58.9	21.6	-
Label-PEnet [18]		30.2	12.9	41.4
LIID (Ours)		48.4	24.9	50.8

个边界框以最大 IoU 检索 MCG SOP [19]。我们使用训练集的伪实例分割来训练 Mask R-CNN 模型 [4]，并将测试结果与 [9], [14]–[18] 和这 9 个基准模型进行比较。

在表9中总结了 VOC2012 segmentation val 数据集 [37] 上的数值的实验结果。请注意，Hsu 等人 [9] 使用边界框作为监督，因此直接将其他方法与其进行比较是不公平的。尽管如此，相对于 [9] 而言，所提出的 LIID 在度量指标 AP_{75} 上实现了 3.3% 的提高，这证明了 LIID 对于准确分割物体实例的有效性。看到 [9] 在指标 AP_{50} 方面胜过 LIID 并不奇怪，因为 [9] 使用的边界框先验将极大地帮助其查找物体实例，从而粗略地分割它们，使其与真值有些重叠。对于图像级监督的方法，所提出的 LIID 在各种评测指标下均达到最佳性能。相比于第二好的方法，即 [17]，LIID 的 AP_{50} 和 AP_{75} 分别高 1.7% 和 7.5%。请注意，AP_{75} 是实例分割中最重要的度量指标，因为它反映了检测紧密覆盖物体的能力。AP_{75} 方面的显著提高表明 LIID 擅长准确地分割与真值高度重叠的物体。最近，使用由 MCG 生成的 SOP 的弱监督物体检测模型 MELM [11] 的性能还不错，但比 PRM [14] 和 LIID 差。这证明了弱监督物体检测与弱监督实例分割密切相关，但不能直接应用于弱监督实例分割。我们在图3中展示了一些我们的实例分割结果的示例。我们可以看到 LIID 可以产生很好的实例分割。即使对于包含多个相同类别实例的图像，也可以很好地分割每个实例。

运行时间和内存消耗. 对于运行时间和内存占用，多路割需要大约 5 分钟的时间和 26 GB 的 CPU 内存才能处理 VOC2012 训练数据集。MIL 框架需要大约 0.02 秒来处理一张图像。因此，每张训练图像的平均运行时间为 $5 \times 60/10\text{K} + 0.02 = 0.05$ 秒。测试图像的运行时间与 Mask R-CNN [4] 相同，因为我们采用伪真值来训练 Mask R-CNN 进行测试。

6.4 MS-COCO 上的实例分割

在本部分中，我们将与 [29], [86] 进行比较，这些方法在 MS-COCO 数据集 [38] 上报告了弱监督实例分割结果。我们使用与 VOC2012 数据集相同的实验设置为 SOP 分配类别标签，并训练 Mask R-CNN [4] 模型。除了 [29], [86]，我们还报告

表 10
COCO test-dev 数据集 [38] 上的实例分割的蒙版 AP。关于度量指标的详细信息可以在 [38] 中找到。浅色的方法是全监督的，而 [29], [86] 和我们的 LIID 是弱监督的。

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [40]	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [87]	29.2	49.5	-	7.1	31.3	50.0
Mask R-CNN [4]	35.7	58.0	37.8	15.5	38.1	52.4
Fan et al. [29]	13.7	25.5	13.5	0.7	15.7	26.1
WS-JDS [86]	6.1	11.7	5.5	1.5	7.1	12.2
LIID (Ours)	16.0	27.1	16.5	3.5	15.9	27.7

了三种全监督的方法的结果，包括 MNC [40]，FCIS [87] 和 Mask R-CNN [4]。评测结果汇总在表10中。所提出的 LIID 的性能明显优于 [29], [86]，这表明所提出的 LIID 对不同的数据集具有鲁棒性。与 [29] 相比，LIID 的 AP、AP_{50} 和 AP_{75} 分别实现了 2.3%、1.6% 和 3.0% 的性能提升。这证明 LIID 相对于 [29] 的改进是不平凡的。

6.5 弱监督语义分割

上述实验在实例分割上评测了我们的方法，而与我们高度相关的另一个挑战性任务是仅在图像级监督下的弱监督语义分割。语义分割可以看作是一个逐像素分类问题，其中每个像素都分配有类别标签。与实例分割不同，语义分割不需要识别具有相同类别的物体。对于训练图像，我们在每张图像中合并具有相同语义类别的实例分割蒙版。然后，我们将产生的语义分割视为伪真值，并采用与以前的方法 [27], [29], [55], [91], [93] 相同的设置来训练 DeepLab [83] 模型。

在表11中，我们在 PASCAL VOC2012 segmentation val 和 test 数据集 [37] 上与最新的方法 [17], [18], [23]–[33], [47], [48], [55], [56], [66], [68], [71], [88]–[94] 用平均重叠率（mean Intersection-over-Union, mIoU）进行比较。为了公平起见，如果原始论文提供了，我们将以 ResNet101 [21] 作为骨干网络报告这些方法的结果（最近的方法通常报告 ResNet101 的结果）。除了 10K VOC2012 训练图像以外，某些方法 [23], [26], [31], [32], [68], [92] 还使用了额外的训练数据，例如网络抓取的图像 [23], [68], [92]，网络抓取的视频 [26], [31] 和像素级标签 [32]，以提高性能，这已在表11中进行了标记。我们提供两种形式的 LIID：一种没有额外的训练数据，另一种在 ImageNet 的简单子集 [69] 上进行了预训练。ImageNet 的简单子集 [69] 从 ImageNet 数据集 [72] 中选择与 PASCAL VOC 具有相同类别的 24K 图像。无论有没有额外的数据，LIID 的表现都优于所有最近的方法。与同时为实例分割和语义分割而设计的 [29] 相比，当 [29] 和 LIID 都使用 24K ImageNet 的简单图像 [69] 作为额外的训练数据时，LIID 比 [29] 在 val 集合和 test 集合上的 mIoU 分别高 3.3% 和 2.7%。这再次证明了 LIID 相对于 [29] 的改进既不琐碎也不简单。最近的新方法 [31] 使用包含 960K 视频帧的 4.6K 视频 [26] 作为额外的训练数据，比

表 11
在 PASCAL VOC2012 segmentation val 和 test 数据集 [37] 上对弱监督语义分割的比较。除了 10K VOC2012 训练图像外，某些方法还使用额外的数据进行训练。24K ImageNet 表示 [69] 中的简单 ImageNet 图像。4.6K Videos 来自 Web-Crawl 数据集 [26]，包括 960K 视频帧。除了图像级别的监督之外，半监督方法 [32], [47], [48] 还分别使用像素级别的标签、点和涂鸦作为监督。为了进行公平的比较，我们使用 ResNet101 [21] 作为骨干网络（如果原始论文提供的话）报告了各种方法的结果。“†” 表示使用 Res2Net101 [34] 作为主干网络的结果。

Method	Year	Extra Data	mIoU (%)	
			val	test
CCNN [88]	ICCV'15	✗	35.3	-
EM-Adapt [89]	ICCV'15	✗	38.2	39.6
MIL [71]	CVPR'15	✗	42.0	-
SEC [56]	ECCV'16	✗	50.7	51.7
AugFeed [66]	ECCV'16	✗	54.3	55.5
Bearman et al. [47]	ECCV'16	Points	49.1	-
ScribbleSup [48]	CVPR'16	Scribbles	63.1	-
STC [23]	PAMI'17	40K Web	49.8	51.2
Roy et al. [25]	CVPR'17	✗	52.8	53.7
Oh et al. [90]	CVPR'17	✗	55.7	56.7
AE-PSL [24]	CVPR'17	✗	55.0	55.7
WebS-i2 [68]	CVPR'17	19K Web	53.4	55.3
Hong et al. [26]	CVPR'17	4.6K Videos	58.1	58.7
DCSP [27]	BMVC'17	✗	60.8	61.9
DSRG [55]	CVPR'18	✗	61.4	63.2
MCOF [91]	CVPR'18	✗	60.3	61.2
AffinityNet [33]	CVPR'18	✗	61.7	63.7
Wei et al. [28]	CVPR'18	✗	60.4	60.8
GAIN [32]	CVPR'18	1464 Pixel	60.5	62.1
Shen et al. [92]	CVPR'18	80K Web	63.0	63.9
Fan et al. [29]	ECCV'18	✗	63.6	64.5
Fan et al. [29]	ECCV'18	24K ImageNet	64.5	65.6
Ahn et al. [17]	CVPR'19	✗	63.5	64.8
FickleNet [93]	CVPR'19	✗	64.9	65.3
Label-PEnet [18]	ICCV'19	✗	-	57.2
Lee et al. [31]	ICCV'19	4.6K Videos	66.5	67.4
SSDD [94]	ICCV'19	✗	64.9	65.5
OAA [30]	ICCV'19	✗	65.2	66.4
LIID (Ours)	-	✗	66.5	67.5
LIID (Ours)	-	24K ImageNet	67.8	68.3
LIID† (Ours)	-	✗	69.4	70.4

LIID 的额外数据多 40 倍。但是，LIID 的性能仍然比它更好，这证明了 LIID 的优越性。我们在图4中展示了一些语义分割结果的示例。结合第6.3节和第6.4节中的实验，我们可以得出结论：对于弱监督实例分割和语义分割，LIID 均达到了最新的性能。

7 总结

在本文中，我们致力于基于图像级监督的弱监督实例分割问题。我们的工作始于一些通用的 SOP。有了这些 SOP，我们首先提出了一个 MIL 框架，该框架可以同时预测概率分布并提取语义特征向量。然后，我们使用获得的信息为所有训练图像构造一个大型知识图。最后，提出了一种改进的多路割

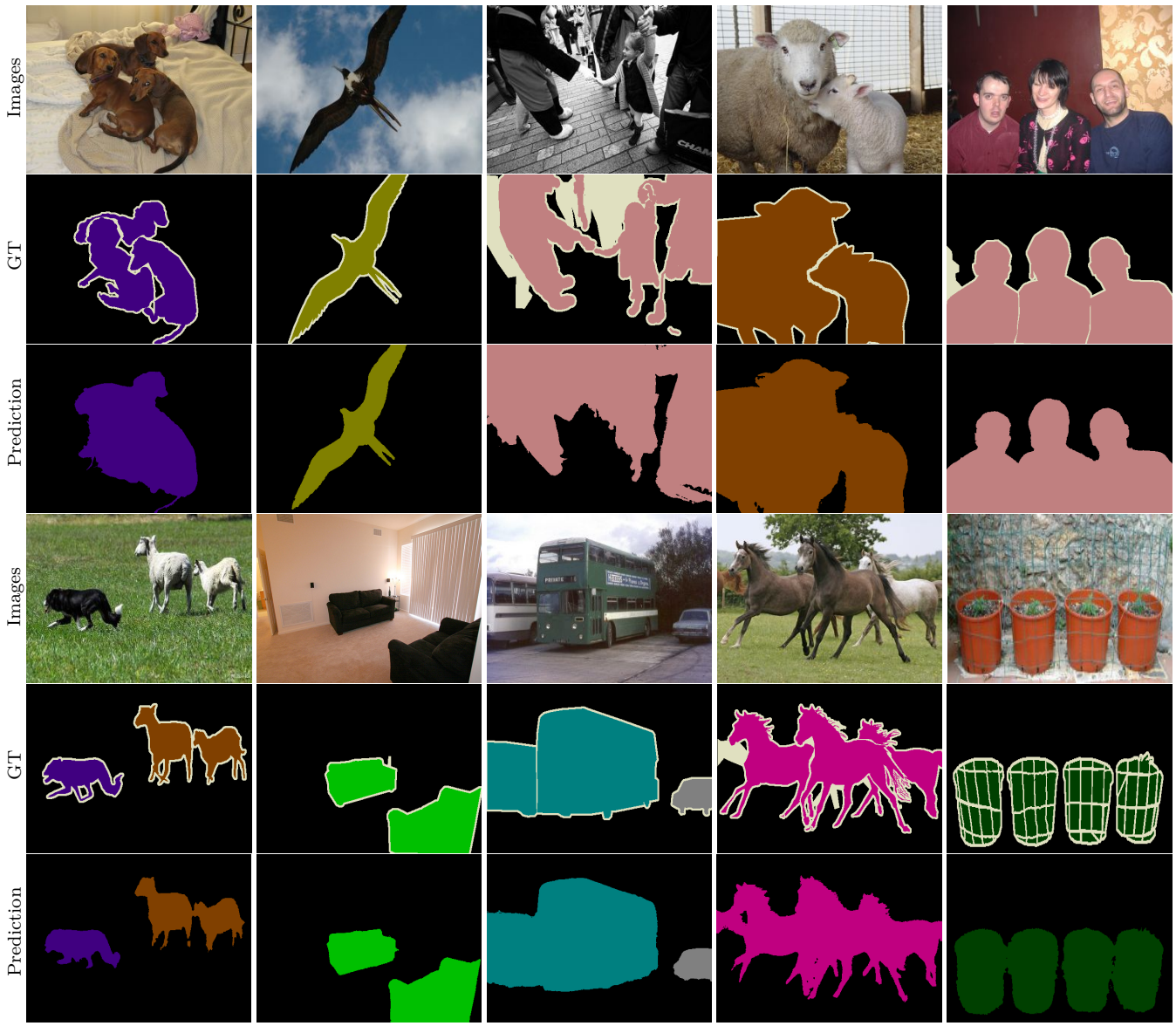

图 4. PASCAL VOC2012 segmentation val 数据集 [37] 上语义分割的定性结果。**从上到下**：原始图像，真值和 LIID 的预测结果，底部三行重复此顺序。

算法，将每个 SOP 分类为一个类别。属于背景类别的 SOP 将被视为嘈杂的数据并被删除。因此，所提出的方法利用了实例、图像和数据集级别的信息来检索 SOP 并为其分配正确的标签。与以前的方法相比，该方法在弱监督实例分割和语义分割方面都可以实现更好的性能。此外，我们对 PASCAL VOC2012 和 COCO 数据集使用相同的超参数，这表明我们的方法的超参数对于不同的数据集具有鲁棒性。将来，我们将尝试将所提出的基于 SOP 的 MIL 框架和多路割表示应用于其他弱监督的视觉任务。

致谢

本研究得到了项目编号为 2018AAA0100400 的“新一代人工智能重大项目”、“国家自然科学基金”（61922046）、“天津自然科学基金”（18ZXZNGX00110）和“教育部指导高校科技创新规划项目”的资助。

参考文献

[1] Y. Liu, Y.-H. Wu, P. Wen, Y. Shi, Y. Qiu, and M.-M. Cheng, “Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 1–1, 2021.

[2] R. Girshick, “Fast R-CNN,” in Int. Conf. Comput. Vis., 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Adv. Neural Inform. Process. Syst., 2015, pp. 91–99.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Int. Conf. Comput. Vis., 2017, pp. 2980–2988.

[5] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 3431–3440.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes

dataset for semantic urban scene understanding," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 3213–3223.

[7] A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 876–885.

[8] Q. Li, A. Arnab, and P. H. Torr, "Weakly- and semi-supervised panoptic segmentation," in Eur. Conf. Comput. Vis., 2018, pp. 106–124.

[9] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," in Adv. Neural Inform. Process. Syst., 2019, pp. 6582–6593.

[10] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 4233–4241.

[11] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 1297–1306.

[12] X. Zhang, J. Feng, H. Xiong, and Q. Tian, "Zigzag learning for weakly supervised object detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 4262–4270.

[13] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang, "Generative adversarial learning towards fast weakly supervised detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 5764–5773.

[14] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 3791–3800.

[15] Y. Zhu, Y. Zhou, H. Xu, Q. Ye, D. Doermann, and J. Jiao, "Learning instance activation maps for weakly supervised instance segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 3116–3125.

[16] H. Cholakkal, G. Sun, F. S. Khan, and L. Shao, "Object counting and instance segmentation with image-level supervision," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 12 397–12 405.

[17] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 2209–2218.

[18] W. Ge, S. Guo, W. Huang, and M. R. Scott, "Label-PEnet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation," in Int. Conf. Comput. Vis., 2019, pp. 3345–3354.

[19] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 1, pp. 128–140, 2017.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Int. Conf. Learn. Represent., 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 770–778.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 2921–2929.

[23] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2314–2320, 2017.

[24] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 1568–1576.

[25] A. Roy and S. Todorovic, "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 3529–3538.

[26] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using web-crawled videos," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 7322–7330.

[27] A. Chaudhry, P. K. Dokania, and P. H. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," in Brit. Mach. Vis. Conf., 2017.

[28] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 7268–7277.

[29] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu, "Associating inter-image salient instances for weakly supervised semantic segmentation," in Eur. Conf. Comput. Vis., 2018, pp. 371–388.

[30] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H.-K. Xiong, "Integral object mining via online attention accumulation," in Int. Conf. Comput. Vis., 2019, pp. 2070–2079.

[31] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation," in Int. Conf. Comput. Vis., 2019, pp. 6808–6818.

[32] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 9215–9223.

[33] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 4981–4990.

[34] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[35] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," Int. J. Comput. Vis., vol. 104, no. 2, pp. 154–171, 2013.

[36] P. Krahenbuhl and V. Koltun, "Learning to propose objects," in IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 1574–1582.

[37] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," Int. J. Comput. Vis., vol. 111, no. 1, pp. 98–136, 2015.

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[39] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in Eur. Conf. Comput. Vis., 2014, pp. 297–312.

[40] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 3150–3158.

[41] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 8759–8768.

[42] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang et al., "Hybrid task cascade for instance segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 4974–4983.

[43] A. Arnab and P. H. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 441–450.

[44] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 2858–2866.

[45] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother, "InstanceCut: from edges to instances with multicut," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 5008–5017.

[46] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," ACM Trans. Graph., vol. 23, no. 3, pp. 309–314, 2004.

[47] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in Eur. Conf. Comput. Vis., 2016, pp. 549–565.

[48] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 3159–3167.

[49] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 3136–3145.

[50] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in Int. Conf. Comput. Vis., 2017, pp. 3544–3553.

[51] D. Kim, D. Cho, D. Yoo, and I. So Kweon, "Two-phase learning for weakly supervised object localization," in Int. Conf. Comput. Vis., 2017, pp. 3534–3543.

[52] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 1325–1334.

[53] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 642–651.

[54] T. Durand, N. Thome, and M. Cord, "Exploiting negative evidence for deep latent structured models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 337–351, 2018.

[55] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 7014–7023.

[56] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in Eur. Conf. Comput. Vis., 2016, pp. 695–711.

[57] Y. Liu, M.-M. Cheng, X. Zhang, G.-Y. Nie, and M. Wang, "DNA: Deeply-supervised nonlinear aggregation for salient object detection," IEEE Transactions on Cybernetics, 2020.

[58] Y. Qiu, Y. Liu, H. Yang, and J. Xu, "A simple saliency detection approach via automatic top-down feature fusion," Neurocomputing, vol. 388, pp. 124–134, 2020.

[59] Y. Qiu, Y. Liu, X. Ma, L. Liu, H. Gao, and J. Xu, "Revisiting multi-level feature fusion: A simple yet effective network for salient object detection," in IEEE Int. Conf. Image Process., 2019, pp. 4010–4014.

[60] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," Computational Visual Media, vol. 5, no. 2, pp. 117–150, 2019.

[61] M.-M. Cheng, N. Mitra, X. Huang, and S.-M. Hu, "SalientShape: group saliency in image collections," The Vis. Comput., vol. 30, no. 4, pp. 443–453, 2014.

[62] S. Xie and Z. Tu, "Holistically-nested edge detection," Int. J. Comput. Vis., vol. 125, no. 1-3, p. 3, 2017.

[63] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 8, pp. 1939–1946, 2019.

[64] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J. Bian, and D. Tao, "Semantic edge detection with diverse deep supervision," arXiv preprint arXiv:1804.02864, 2018.

[65] Y. Liu, S.-J. Li, and M.-M. Cheng, "RefinedBox: Refining for fewer and high-quality object proposals," Neurocomputing, 2020.

[66] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, "Augmented feedback in semantic segmentation under image level supervision," in Eur. Conf. Comput. Vis., 2016, pp. 90–105.

[67] W. Shimoda and K. Yanai, "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in Eur. Conf. Comput. Vis., 2016, pp. 218–234.

[68] B. Jin, M. V. O. Segovia, and S. Süsstrunk, "Webly supervised semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 1705–1714.

[69] Q. Hou, D. Massiceti, P. K. Dokania, Y. Wei, M.-M. Cheng, and P. H. Torr, "Bottom-up top-down cues for weakly-supervised semantic segmentation," in International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), 2017, pp. 263–277.

[70] F. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in Eur. Conf. Comput. Vis., 2016, pp. 413–432.

[71] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 1713–1721.

[72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in IEEE Conf. Comput. Vis. Pattern Recog., 2009, pp. 248–255.

[73] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in Eur. Conf. Comput. Vis., 2014, pp. 725–739.

[74] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," Computational Visual Media, vol. 5, no. 1, pp. 3–20, 2019.

[75] Z. Zhang, Y. Liu, X. Chen, Y. Zhu, M.-M. Cheng, V. Saligrama, and P. H. Torr, "Sequential optimization for efficient high-quality object proposal generation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1209–1223, 2017.

[76] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-

M. Hu, "S4Net: Single stage salient-instance segmentation," Computational Visual Media, vol. 6, no. 2, pp. 191–204, 2020.

[77] K. P. Murphy, Machine learning: A probabilistic perspective. MIT press, 2012.

[78] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis, "The complexity of multiterminal cuts," SIAM Journal on Computing (SICOMP), vol. 23, no. 4, pp. 864–894, 1994.

[79] N. Garg, V. V. Vazirani, and M. Yannakakis, "Approximate max-flow min-(multi) cut theorems and their applications," SIAM Journal on Computing (SICOMP), vol. 25, no. 2, pp. 235–251, 1996.

[80] G. Călinescu, H. Karloff, and Y. Rabani, "An improved approximation algorithm for multiway cut," Journal of Computer and System Sciences (JCSS), vol. 60, no. 3, pp. 564–574, 2000.

[81] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in Eur. Conf. Comput. Vis., 2016, pp. 499–515.

[82] C. Bliek1ú, P. Bonami, and A. Lodi, "Solving mixed-integer quadratic programming problems with IBM-CPLEX: A progress report," in RAMP Symposium, 2014, pp. 16–17.

[83] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, 2018.

[84] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "HCP: A flexible CNN framework for multi-label image classification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 9, pp. 1901–1907, 2016.

[85] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal networks for weakly supervised object localization," in Int. Conf. Comput. Vis., 2017, pp. 1841–1850.

[86] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao, "Cyclic guidance for weakly supervised joint detection and segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 697–707.

[87] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 2359–2367.

[88] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in Int. Conf. Comput. Vis., 2015, pp. 1796–1804.

[89] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in Int. Conf. Comput. Vis., 2015, pp. 1742–1750.

[90] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, B. Schiele et al., "Exploiting saliency for object segmentation from image level labels," in IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[91] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 1354–1362.

[92] T. Shen, G. Lin, C. Shen, and I. Reid, "Bootstrapping the performance of webly supervised semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 1363–1371.

[93] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 5267–5276.

[94] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in Int. Conf. Comput. Vis., 2019, pp. 5208–5217.