WSOD^2:对于弱监督目标检测的自底向上和自顶向下的似物性蒸馏学习法

Zhaoyang Zeng1,2*, Bei Liu3, Jianlong Fu3, Hongyang Chao1,2, Lei Zhang3
1School of Data and Computer Science, Sun Yat-sen University
2 The Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education
3 Microsoft Research

我们研究了弱监督目标检测（WSOD），它在对象级标注任务中减少人工干预起到重要作用。将卷积神经网络（CNN）与区域提议机制相结合是当前的主流。尽管 CNN 在提取判别局部特征方面优势很大，但在测量边界框包含完整物体的概率方面仍然存在巨大的挑战（即 “似物性”）。本文通过设计一种适合于弱监督目标检测的训练机制，提出一个新的基于似物性蒸馏的 WSOD 框架（也就是 WSOD^2）。多元回归目标是采用自适应线性组合的方式，通过联合考虑低级特征的自下而上（BU）和自上而下（TD）似物性以及 CNN 置信度来确定的。因为在训练过程中边界框回归有助于区域提议学习，以接近它的具有高似物性的回归目标，因此通过优化，由自底向上论证学习得来的深层似物性特征可以逐渐精炼到 CNN 中。对于 BU/TD 似物性，我们探索了不同的自适应训练曲线，展示了我们提出的 WSOD^2 能够得到优秀的结果。

1. 介绍

识别和定位图片中的目标的能力显示了视觉信息的深度理解，并且在最近纪念备受关注。随着卷积神经网络（CNN）[5, 14, 19, 27]的发展，取得了重大进展。然而，当前先进的目标检测器大多依赖大规模的训练数据，要求人工标记边界框（例如 PASCAL VOC2007/2012 [7]，MS COCO [22]，Open Images [20]）。为了减轻繁重的打标签的工作，减少成本，通过仅利用图像级标注提出弱监督目标检测方法[2, 30, 37, 38]。

为了处理弱监督目标检测（WSOD）任务，大部分先前的工作采用多实例学习方法，将 WSOD 转换到多标签分类问题[2, 18]。后来，提出在线实例分类器（OICR）[29]和提议的聚类学习（PCL）[28]，通过明确分配实例标签来学习更有判别力的实例分类器。OICR 和 PCL 两个都采用了将初始目标检测器的输出作为伪真实标签的方法，两个方法显示了对提升 WSOD 的分类能力有益。然而，通常情况下分类模型是检测一个类别的物体是否存在，而不能预测图像中物体的位置、尺寸和数量。这个弱点通常会导致检测的边界框的不完全或过大，如图 1 的第一和第三行所示。OICR 和 PCL 和性能很大程度依赖初始目标检测结果的精确度，这限制了其进一步的提高。而且，他们忽视了对于边界框回归的学习，它在现代目标检测器[3, 4, 13, 21, 24]设计中起到重要作用。C-WSL 将边界框回归集成到 OICR 框架中以减少定位误差，然而它依赖一个贪婪的真实标记选择策略，需要额外的计数标注[9]。

依赖于初始弱监督目标检测结果的现有工作，通过卷积神经网络（CNN）从特征图中学习物体边界框。尽管 CNN 擅长以自上而下的方式用图像级标签学习物体的判别性局部特征，但是它在没有真实标记作为监督的情况下检测边界框是否包含一个完整物体表现很差。

人们已经提出了一些基于物体证明的低级特征（例如颜色对比度[23]和超像素跨度[1]），以测量通用的似物性，该似物性是以一种自底向上的方法测量边界框中包含某类物体的可能性。受到这些自底向上的物体证明的启发，此工作中，我们探索着使用它们的优势提升 CNN 模型在捕获图像似物性的能力。我们提出将这些自底向上的善于发现边界框的论证和在单网络中有着强大表现能力的 CNN 结合起来。

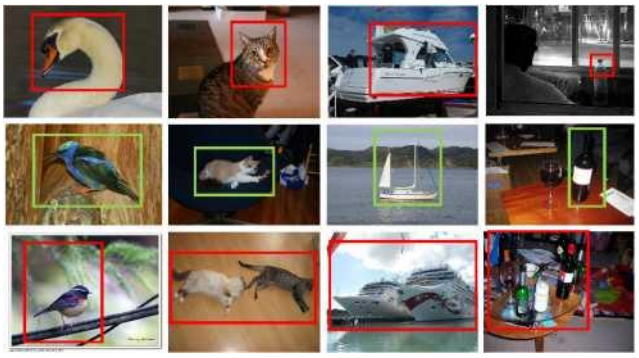


图 1 由 OICR [30]得到的典型弱监督目标检测结果。在第一、第二以及第三行中我们可以分别观察到物体的部分、正确的以及过大的检测结果。

我们提出一个带有似物性蒸馏机制的 WSOD 框架（WSOD^2），以利用自底向上的物体证明和自顶向下的带有新型训练机制的分类输出。首先，给定一个带有成千上万个区域提议（例如，由选择搜索方法[33]生成）的输入图片，我们学习一些实例分类器，以预测每一个区域提议的分类概率。这些分类器的每一个都可以帮助选择多个高置信边界框作为可能的物体实例（也就是说，伪分类和边界框回归真实值）。第二，我们使用一个边界框回归器来微调每一个提议的定位和尺寸。第三，因为每一个边界框不能仅仅通过 CNN 特征来捕获精确的目标边界，我们采用自适应线性结合法结合了自底向上的物体证明法和自顶向下的 CNN 置信分数，以测量每一个候选边界框的似物性，为每一个区域提议分配标签来训练分类器和回归器。

对于 CNN 喜欢的一些具有判别性的小边界框，自底向上的物体证明法（例如超像素跨度）往往非常低。WSOD^2 可以调节伪真实值以满足更高的 CNN 置信度和低级物体完整性。除此之外，集成进去一个边界框回归器来减小定位误差，同时在训练过程中增强自底向上的物体证明法的效果。我们设计了一个自适应训练策略，使监督过程逐渐精炼，这使得 CNN 模型能够得到充分的训练，以至于当模型收敛时可以描述物体的有判别性的局部和边界信息。

据我们所知，该工作是首个在弱监督目标检测任务中探究自底向上物体证明法的。贡献可以总结为以下几点：

1. 我们提出在弱监督目标检测任务中将自底向上物体证明法与自顶向下类别置信分数相结合。
2. 我们提出 WSOD^2(带有似物性蒸馏法的 WSOD)，通过一个边界框回归器和一个自适应训练机制，在 CNN 中提炼物体边界知识。

3. 我们在 PASCAL VOC2007/2012 和 MS COCO 数据集上的实验表明了我们提出的 WSOD^2 的效果。

2. 相关工作

2.1 弱监督目标检测

弱监督目标检测在最近几年备受关注。大多数现有工作采纳多实例学习[2, 6, 17, 28, 29, 31, 34]的想法，将弱监督目标检测转换成多标签分类任务。Bilen 等人[2]提出 WSDDN 方法，该方法在分类和检测分支的分数上执行乘法，因此可以选择高置信的正样本。Tang 等人[28]和 Tang 等人[29]发现，在线的将图像级标签转换到实例级监督是土生精确度的有效方法，因此提出基于前几个分支的输出在线改进实例分类器的几个分支。因为一个分类器生成的类别激活图可以大致定位物体[39, 40]，Wei 等人[36]试图应用它来生成导向性的检测结果，然后使用这些结果作为后续提升的参考。大多数前面的工作很大程度依赖伪真实标记挖掘，要么在线（在训练循环内部）要么离线（训练完成之后）。这种伪真实标记是由分类置信度[28, 29]或者手工规则[9, 38]决定的，这些规则不够精确不能测量区域的似物性。

2.2 边界框回归

边界框回归在[12]中被提出，被最近绝大多数以 CNN 为基础的全监督目标检测器[3, 4, 13, 21, 24]所采用，因为它可以减少预测框的定位误差。然而，由于缺乏监督，将边界框引入弱监督目标检测的工作还很少。一些工作将边界框回归看作后处理模块。在这些工作当中，OICR [29]直接使用训练集的检测结果去训练 Fast R-CNN。W2F [38]设计了一些策略，基于 OICR 的输出离线地选择高精度的伪真实标记。与此不同的是，Gao 等人[9]将边界框回归器集成到训练循环内的 OICR 中，使用额外计数信息来帮助选择伪真实标记。

在本文中，我们将边界框回归器集成到弱监督检测器中，通过新型使用自底向上的物体证明法分配回归目标。

3. 方法

我们所提出的带有似物性蒸馏的弱监督目标检测器（WSOD^2）的概述如图 2 所示。我们首先采用了基于多实例检测器（即 Cls 0）来获得初始的检测目标边界框。基于每一个所提出的边界框的定位，我们计算自底向上物体证明。这种证明充当向导，将图像级标签转换为实例级监督。我们以一种端到端并且自适应的方式对整个网络进行优化。这部分中，我们将详细介绍 WSOD^2。

3.1 基于多实例的检测器

弱监督目标检测中，只有图像级的标注是可用的。为了更好地理解图像内的语义信息，我们需要深入到区域级别，分析每一个框的特性。我们首先搭建一个基检测器来获得初始检测结果。我们按照 WSDDN [2]采用多实例学习的思想，通过将 WSOD 转换到多标签分类问题来优化基检测器。特别的是，给定一张输入图片，我们首先通过选择性搜索法[33]生成区域提议$\boldsymbol{R}$，通过 CNN 骨干网络、一个感兴趣区域池化层以及两个全连接层提取区域特征$\boldsymbol{x}$。

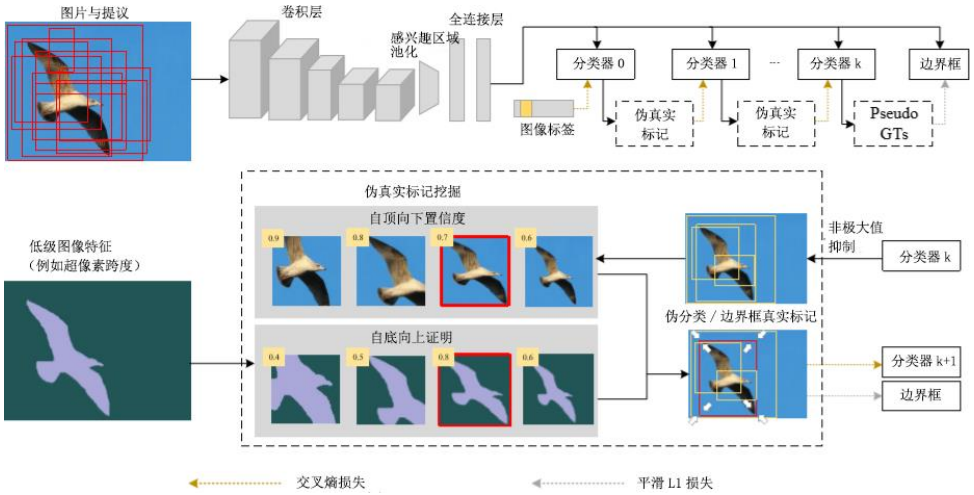


图 2 WSOD^2 的框架。将带有标签和预计算提议的图片输入到 CNN 中得到区域特征。然后区域特征会经过几个分类器和一个边界框回归器。应用非极大值抑制（NMS）从预测中挖掘正样本。分别通过分类分支和低级图像特征计算自顶向下（TD）置信度和自底向上（BU）证明。将它们结合起来对于每一个提议分配类别标签和回归目标。“Cls”表示分类器，“Bbox”表示边界框回归器。白色箭头表示两个范例区域提议的优化方向。[通过颜色观看效果最佳]

区域特征$\boldsymbol{x}$随后通过两个独立的全连接层输入到两个流中，生成的两个特征矩阵表示为$\boldsymbol{x}^c,\boldsymbol{x}^d \in \mathbb{R}^{C\times|\boldsymbol{R}|}$，其中C表示类别数量，$|\boldsymbol{R}|$表示提议数量。在$\boldsymbol{x}^c$和$\boldsymbol{x}^d$上向两个不同方向分别应用两个 softmax 函数，如下所示：

$$[\boldsymbol{\sigma}^c]_{ij}=\frac{e^{[\boldsymbol{x}^c]_{ij}}}{\sum_{k=1}^{C} e^{[\boldsymbol{x}^c]_{kj}}},\ [\boldsymbol{\sigma}^d]_{ij}=\frac{e^{[\boldsymbol{x}^d]_{ij}}}{\sum_{k=1}^{|\boldsymbol{R}|} e^{[\boldsymbol{x}^d]_{ik}}} \tag{1}$$

其中$[\boldsymbol{\sigma}^c]_{ij}$表示第$j$个区域提议的第$i$个类别标签的预测值，$[\boldsymbol{\sigma}^d]_{ij}$表示对于第$i$个类别的第$j$个区域提议所学习到的权重。我们通过对应元素乘积$\boldsymbol{s}=\boldsymbol{\sigma}^c\odot\boldsymbol{\sigma}^d$计算提议分数，在区域维度上聚合

来获得图像级分数向量$\boldsymbol{\phi}=[\phi_1,\phi_2,\cdots,\phi_{\mathrm{C}}]$，其中项$\boldsymbol{\phi}_c=\sum_{r=1}^{|\boldsymbol{R}|}[\boldsymbol{s}]_{cr}$。以这种方式，我们可以将图像级类别标签作为监督，应用二进制交叉熵损失函数去优化基检测器。基损失函数表示为：

$$L_{base}=-\sum_{c=1}^{\mathrm{C}}\left(\hat{\boldsymbol{\phi}}_c log(\boldsymbol{\phi}_c)+\left(1-\hat{\boldsymbol{\phi}}_c\right)log(1-\boldsymbol{\phi}_c)\right) \tag{2}$$

其中，$\hat{\boldsymbol{\phi}}_c=1$表示输入图像包含第$c$个类别，否则$\hat{\boldsymbol{\phi}}_c=0$。预测分数$\boldsymbol{s}$被当作初始检测结果。然而它不够精确，可以进一步精炼，在[29]中被讨论过。

3.2 自底向上和自顶向下的似物性

目标检测器的本质是边界框排序函数，函数中的似物性测量是一个重要因素。最近的以 CNN 为基础的检测器[13, 24, 25]通常将分类器置信度看作似物性分数。然而，这样的策略在弱监督方案中有缺陷，对于训练好的检测器来说很难从有判别性的物体部分和无关背景中判别完整物体。为了缓解这个问题，我们研究了自底向上的物体证明法（如超像素跨度），在传统目标检测中起到重要作用。

正如[1]中表述的，物体是具有良好的边界和中心的独立对象。因此，我们期望带有完成物体的框要比带有部分的、过大的或者背景框具有更高的似物性分数。自底向上的物体证明法总结了一般物体的边界特性，可以弥补 CNN 对于边界检测的弱点。

我们提出将自底向上的物体证明法应用到训练弱监督目标检测器中。特别的是，收到 OICR [29]的启发，我们在$\boldsymbol{x}$的顶部搭建K个实例分类器，将第k个分类器的输出看作是第$(k+1)$个分类器的监督，运用自底向上物体证明法来指导网络训练。每一个分类器都是在$\mathrm{C}+1$个类别（我们将背景看作第0个类别）上由一个全连接层和一个 softmax 层构成。正式的说，对于第k个分类器，我们定义第k个分类器的损失函数如下：

$$L_{ref}^{k}=-\frac{1}{|\boldsymbol{R}|}\sum_{r\in\boldsymbol{R}}\left(\boldsymbol{w}_r^k\cdot CE\left(\boldsymbol{p}_r^k,\hat{\boldsymbol{p}}_r^k\right)\right) \tag{3}$$

其中，$\boldsymbol{p}_r^k$表示提议r的$\{\mathrm{C}+1\}$个维度输出类别概率，$\hat{\boldsymbol{p}}_r^k$表示其真实标记的独热编码。

$CE\left(\boldsymbol{p}_r^k,\hat{\boldsymbol{p}}_r^k\right)=-\sum_{c=0}^{\mathrm{C}}\hat{\boldsymbol{p}}_{rc}^k log(\boldsymbol{p}_{rc}^k)$是标准的交叉熵函数。由于实际的实例级真实标签无法使用，因此我们使用在线策略动态地在训练循环中选择每一个提议的伪真实标签，在 3.4 节会进一步解释。我们基于提议r的似物性在线地分配损失权重$\boldsymbol{w}_r^k$。特别的是，我们首先提取了提议r的自底向上的似

物性特征，并将其表示为$O_{bu}(r)$，将$O_{bu}(r)$与$O_{td}^{k}(r)$相结合，$O_{td}^{k}(r)$是由第k个分类器得到的类别置信度。$\boldsymbol{w}_r^k$是自底向上证明结果与自顶向下置信度的线性结合，表达式如下：

$$\boldsymbol{w}_r^k = \alpha O_{bu}(r) + (1-\alpha)O_{td}^{k}(r) \tag{4}$$

其中，α为自底向上物体证明法的影响因子。公式 4 重点三个术语定义如下：

自底向上的物体证明法O_{bu}。这项工作中我们主要采用了超像素跨度（SS）作为自底向上的证明法，我们也研究了其他三种方法：多尺度显著性（MS）、颜色对比度（CC）以及边缘密度（ED）。这些方法的实验详见 4.2 节。

自顶向下的类别置信度O_{td}。我们基于前一个分支的输出计算当前分支的自顶向下的置信度。特别的是，一旦我们得到了第$(k-1)$个分支的类别概率$\boldsymbol{p}_r^{k-1}$，第k个分支的自顶向下类别置信度可以表示为：

$$O_{td}{}^{k}(r) = \sum_{c=0}^{\mathrm{C}} \left(\boldsymbol{p}_{rc}{}^{k-1} \cdot \hat{\boldsymbol{p}}_{rc}{}^{k} \right) \tag{5}$$

由于$\hat{\boldsymbol{p}}^k$是独热编码向量，仅有一个值的$\boldsymbol{p}^{k-1}$被选来计算$O_{td}^{k}(r)$。

影响因子α。α是平衡自底向上物体证明结果和自顶向下类别置信度效果的重要因素，有一些权重衰减函数计算得来。这样的设计使得边界知识得以蒸馏进入 CNN 中，详见 3.4 节。

因为自底向上的物体证明法和自顶向下的类别置信度可以从边界和语义信息的角度测量边界框中包含物体的可能性，我们将这两种表述分别看作自底向上似物性和自顶向下似物性。

3.3 边界框回归

自底向上物体证明法能够发现物体边界，所以我们研究如何让它指导预先计算出来的边界框在训练阶段更新。直观的想法是集成边界框回归来精调提议的位置和尺寸。

边界框回归在典型的全监督目标检测器中是必要的组件，因为它能减少定位误差。尽管边界框标注在弱监督目标检测中不可用，一些现有的工作[9, 28, 30, 38]显示在线或离线挖掘伪真实标签并且拟合它们可以提升很大性能。受此想法的启发，我们在$\boldsymbol{x}$的顶部集成了边界框回归器，使它能在线更新。边界框回归器和在 Fast R-CNN [11]中的表示形式一样。对于区域提议r来说，回归器预测定位和尺寸的偏置$\boldsymbol{t}_r = (t_r^x, t_r^y, t_r^w, t_r^h)$，可以按照下列形式进行进一步优化：

$$L_{box}=\frac{1}{|\boldsymbol{R}_{pos}|}\sum_{r=1}^{|\boldsymbol{R}_{pos}|}\left(\boldsymbol{w}_r^{\mathrm{K}}\cdot smooth_{L1}\left(t_r,\hat{t}_r\right)\right) \tag{6}$$

其中，$\hat{t}_r$是通过[12]中描述的r和$\hat{r}$之间的坐标和尺寸差值计算得来，$\hat{r}$表示回归参考。$\boldsymbol{R}_{pos}$表示正区域（非背景），详见 3.4 节。$smooth_{L1}$函数和[25]中定义的一样，w_r^{K}表示通过最后一个分类分支计算得来的回归损失权重。我们基于w_r^{K}的影响计算伪回归参考$\hat{r}$，这评价了我们在 3.2 节中表述的提议的似物性：

$$\hat{r}=\underset{\left\{m\in M(\mathrm{K},\boldsymbol{R})|IoU(m,r)>\mathrm{T}_{iou}\right\}}{arg\max}\left(\boldsymbol{w}_m^{\mathrm{K}}\right) \tag{7}$$

其中，M是正样本挖掘函数，详见 3.4 节，是T_{iou}特定的 IoU 阈值。公式 7 使每一个正区域样本得以靠近一个似物性很高的框。

我们采用边界框回归器在训练阶段增强框的预测。用下述方法更新公式 4：

$$\boldsymbol{w}_r^k=\alpha O_{bu}(r')+(1-\alpha){O_{td}}^k(r) \tag{8}$$

其中r'是r与t_r的偏置。我们保证$O_{td}^k(r)$不变，因为O_{td}^k包含一个 RoI 特征扭曲的操作，会受到边界框预测的影响。在这个新的形式中，提议的定位是在线更新的。更新后的框可能完成更高的似物性，这意味着会有更高的可能性选取到更加精确完整的回归目标。

3.4 似物性蒸馏

公式 3 和知识蒸馏[15. 16]有相似的形式，其中外部知识来自自底向上和自顶向下的似物性。其中的α是平衡每一个知识的权重。在训练开始时，自顶向下的分类器不够被信赖，所以我们期望自底向上的证明法可以在结合中（即公式 4）可以起到主导作用。有了自底向上的证明的指导，网络会试图调节自顶向下分类器的置信分布以服从自底向上的证明。我们称这个过程为似物性蒸馏。

随着训练的进行，O_{td}可信度增加，O_{td}从O_{bu}学到了边界的决策能力，而因为是有监督分类它仍然保持语义理解能力。因此，α可以逐渐将注意力从自底向上物体证明转移到自顶向下的 CNN 置信度上。特别的是，α可以由一些权重衰减函数算得。我们调查了几个权重要贱函数包括多项式、余弦以及常数函数，我们将在 4.2 节对比不同函数的效果。

除了α，为了实现似物性蒸馏，我们还需要测定$\hat{\boldsymbol{p}}_r^k$。我们想在保持语义识别能力的同时，应用自底向上证明来加强边界表述因此我们应用来自上一个分支的分类器的输出以挖掘正提议。

给定来自第$(k-1)$个分类器的输出，我们通过下面的步骤挖掘伪真实标记：

1. 使用预先定义的阈值T_{nms}，基于每一个提议r的类别概率$\boldsymbol{p}_r^{k-1}$，对$\boldsymbol{R}$应用非极大值抑制（NMS）。将保留下来的框定义为$\boldsymbol{R}_{keep}$。
2. 对于每一个类别$c(c>0)$，如果$\hat{\boldsymbol{\phi}}_c=1$，在$\boldsymbol{R}_{keep}$中搜索所有的框，框内对于类别$c$的类别置信度比另一个预先定义的阈值$\mathrm{T}_{conf}$大，分配类别标签$c$给这些框。特殊的是，如果没有框被选中，则选取分数最高的框。将所有被搜索的框的集合定义为$\boldsymbol{R}_{seek}$。
3. 对于$\boldsymbol{R}_{seek}$中的每一个搜索框，搜索在$\boldsymbol{R}$中所有它的邻近框。这里如果两个框的交并比（IoU）大于一个阈值T_{iou}，就认为两个框为邻近关系。将所有邻近框的集合定义为$\boldsymbol{R}_{neighbor}$。将相同的类别标签分配给所有的邻近框作为它们的种子框。其他非种子框和非邻近框当作背景。将已分配的标签转换为独热向量，以获得所有的$\hat{\boldsymbol{p}}_r^k$。
4. 最后，将$\boldsymbol{R}_{seek}$和$\boldsymbol{R}_{neighbor}$的并集作为正提议：$\boldsymbol{R}_{pos}=\boldsymbol{R}_{seek}\cup\boldsymbol{R}_{neighbor}$。

正如在 3.2 节和 3.3 节所提到的，将上述操作归为函数$M(k,\boldsymbol{R})$，函数返回正提议的集合。通过这样的方式，邻近的正样本会分配相同类别标签，而似物性高的样本会得到高权重。这样的信息会通过优化被蒸馏进 CNN 中，因此 CNN 会逐渐提升发现物体边界的能力。

3.5 训练和推断细节

训练。总体的训练目标的表达式为：

$$L=L_{base}+\lambda_1\sum_{k=1}^{\mathrm{K}}L_{ref}^k+\lambda_2 L_{box} \tag{9}$$

其中，λ_1和λ_2是用来平衡损失权重的超参数。这里我们取$\lambda_1=1$，$\lambda_2=0.3$，按照[29]设置$\mathrm{K}=3$。由于所有的K个分类器的监督都来自于前一分支，在第一个 2000 次迭代设置$\alpha=0$来预热。当挖掘伪真实值标记是，通常遵循[38]设置。

推断。模型有K个改良分类器和一个边界框回归器。对于每一个预测框来说，遵循[29]对来自所有的K个分类器的输出取平均，产生类别置信度，使用边界框回归器调节它的位置和尺寸。最后，应用阈值 0.3 的 NMS 去除冗余的检测框。

4. 实验

4.1 实验设置

数据集和评价标准。在三个目标检测基准上评价我们的方法：PASCAL VOC2007&2012 [7]和 MS COCO [22]。删掉这些数据集提供的边界框标注后，只使用图片和他们的标签信息来训练。PASCAL VOC 2007 和 2012 分别由 20 个类别的 9962 张和 22531 张图片组成。对于 PASCAL VOC 来说，在 trainval 上进行训练（2007 有 5011 张图，2012 有 11540 张图），在 test 上做均值平均精度（mAP）的汇报，而且还在 trainval 上采用修正定位（CorLoc）来测量定位精度。在符合标准设定$IoU > 0.5$的条件下，执行两个评价标准。MS COCO 包含 80 个类别。在 train2014 上训练，在 val2014 上评价，两个集合分别包含 82783 张和 40504 张图片。我们在 val2014 数据集上，在设定交并比为 0.5 的基础上报告了平均精度结果，在以交并比为 0.5 的基础，0.05 为递增，直到 0.95 的情况下报告平均精度结果。

实现细节。采用 VGG16 [26]作为 CNN 的骨干网络，使用 ImageNet [19]的与训练参数作初始化。对于所有新的层权重，使用均值为 0，标准差 0.01（除了边界框回归器是 0.001）的高斯分布随机地初始化方式，所有新的偏置值初始化为 0。我们遵从一个被广泛使用的设定[2, 29, 30, 37]使用选择性搜索法为每一张图片生成大约 2000 个提议。整个网络采用端到端优化，SGD 优化器参数为10^{-3}的初始学习率、0.0005的权重衰减以及0.9的动量。总体的迭代步数在 VOC 2007 上被设定为 80000，学习率在第 40000 步时除以 10.对于 VOC 2012 将迭代步数翻倍。学习率衰减步数也翻倍到 80000 步。对于 MS COCO 设定迭代步数为 360000，使学习率在第 180000 步衰减。遵从[28, 29]在训练时采用多尺度设定。特殊的是，输入图片的短边会随机地改变到$\{480, 576, 588, 864, 1280\}$尺寸，限定长边的长度不大于 2000。而且，所有训练图片水平翻转也会在训练时被使用。对于消融实验报告了单尺寸测试结果，当与前人工作对比时报告了多尺寸测试的结果。所有的实验都是基于在 4 个英伟达 P100 显卡上的 PyTorch 实现的。

4.2 消融实验

我们在 PASCAL VOC 2007 上进行模型消融测试，证明 WSOD^2 的有效性。

自底向上的证明。对于自底向上的物体证明，我们用独立和结合的方法测试四个证明法的效果。四种证明法列举如下：

1) 多尺度显著性（MS），在一些尺度上总结显著性；
2) 颜色对比度（CC），计算与邻近区域的颜色分布差异；
3) 边缘密度（ED），计算内环中边缘密度；
4) 超像素跨度（SS），分析所有超像素的跨度。

因为不同证明法的值域不同，我们将算得的值归一化到$[0-1]$。对于 CC，ED 和 MS 来说，固定参数，由于缺少监督因此按照经验设定$\theta^{MS}=0.2,\theta^{CC}=2,\theta^{ED}=2$。对于 SS 来说，遵循[8]，设置$\theta_{\sigma}^{SS}=0.8,\theta_{k}^{SS}=300$。请读者参阅[1]获取更多关于四种证明法的细节和$\theta^{MS},\theta^{CC},\theta^{ED},\theta^{SS}$的意义。

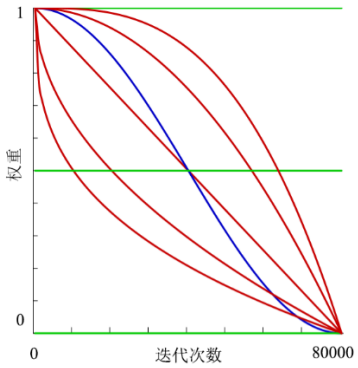
为了消融分析这些自底向上证明法的效果，在此模型消融分析实验中对于所有包含在证明法中的参数设置只保留$\alpha=1$，对于那些没有在自底向上证明法只作为对比基线的方法取$\alpha=0$。同样也测试了四个证明法均值结合的效果。正如[1]中讨论的那样，线性结合并不好，我们只是为了评价自底向上证明法的效果以及对以后工作有所启发进行的这个实验。

结果在表 1 中进行展示。从和基线对比可以看到，在自底向上证明法的指导下，性能可以显著提高。表 1 也包含了在所有类别上的 AP 值，从中可以发现不同的证明法对不同的类别效果不同。例如，对于单个的证明，ED 对于“船只”效果好，而在“电视”效果不好。此外，我们可以发现，这一结果也与[1]中所报告的测量结果一致，表明这些自底向上的证据与目标检测性能正相关。从它们的相结合的结果可以发现，可以完成比除了 SS 以外的所有单个证明法更好的性能。我们相信线性均值在结合这些证明法方面并不是正确的方法，在未来会有更好的方法被研究出来。在后续的实验中我们采用 SS 作为自底向上的物体证明法。

表 1 自底向上的物体证明法的模型对比实验。将每一种证明法集成到 WSOD^2 中，做了 PASCAL VOC 2007 测试集的均值平均精度（mAP）。也通过简单的平均结合了所有的证明法，结果列在最后一行。

方法	飞机	单车	鸟	船	瓶子	公交	汽车	猫	椅子	牛	桌子	狗	马	摩托	人	植物	羊	沙发	火车	电视	mAP
N/A	58.5	63.5	46.3	25.0	18.7	66.4	63.6	55.7	26.4	45.7	42.2	43.8	48.5	63.5	15.0	24.5	44.3	49.8	62.3	54.3	45.9
CC	62.0	64.5	44.9	24.5	19.6	70.3	62.9	52.6	20.6	54.5	44.2	49.0	55.7	64.9	15.1	22.0	49.2	56.2	52.7	58.6	47.2
ED	52.7	60.2	44.2	32.2	20.6	65.8	60.8	57.0	21.8	57.7	38.1	51.0	57.5	66.2	15.0	25.0	52.2	54.1	61.0	37.8	47.0
MS	62.0	66.2	41.2	25.1	19.2	68.1	61.5	60.7	12.2	52.9	47.9	61.6	58.8	65.6	18.1	17.6	47.2	59.0	54.3	51.4	47.5
SS	61.3	63.6	44.6	26.6	21.0	65.5	61.2	49.0	25.1	52.6	44.2	58.3	64.1	65.8	16.7	21.9	49.6	53.7	59.4	57.8	48.1
CC+ED+MS+SS	59.5	57.6	43.1	29.7	19.7	65.4	59.7	68.1	21.5	57.6	45.7	50.5	58.4	64.0	14.6	17.2	50.4	61.2	64.9	50.0	47.9

影响因子α。我们测试了几个权重衰减函数，包括常数（$\alpha=0,0.5,1$）、多项式（$\alpha=-(n/N)^{\gamma}+1$,其中$\gamma=2,3,1,1/2,1/3$）以及余弦函数（$\alpha=1+\cos(n\pi/N)/2$）。结果在图 3 中进行展示。从前三行对比中可以发现自底向上的物体证明帮助模型学习边界特征，取得更好的目标检测效果。在不同的设计中，线性衰减（即$\alpha=-(n/N)+1$）性能最好，后续的实验都是基于此设定进行。我们会为了未来的研究一直探索最好的参数。



(a)

α 衰减函数	γ	mAP
$\alpha=\gamma$（绿色曲线，从上向下为$\gamma=0,\frac{1}{2},1$）	0	45.9
	1/2	47.2
	1	48.1
$\alpha=-\left(\frac{n}{N}\right)^{\gamma}+1$（红色曲线，从上到下$\gamma=3,2,1,\frac{1}{2},\frac{1}{3}$）	3	49.2
	2	49.0
	1	50.3
	1/2	46.5
	1/3	46.3
$\alpha=\frac{\left(1+\cos\left(\frac{n\pi}{N}\right)\right)}{2}$（蓝色曲线）	-	49.7

(b)

图 3 对于α的权重衰减函数的模型消融实验分析。(a)是不同函数的权重衰减曲线。(b)是在 PASCAL VOC 2007 测试集上不同衰减设置的 mAP，n 和分别表示当前步数和总步数。[通过颜色观看效果最佳]

表 2 WSOD2 的不同组件的对比试验。√表示该组件被使用。当对于每一类别带有最高置信度的提议被用作种子边框时，“非极大值抑制”不勾选。

边界框	非极大值抑制	自底向上	α 衰减	mAP
				43.3
√				45.1
√	√			45.9
√	√	√		48.1
√	√	√	√	50.3

每一个组件的效果。表 2 展示了每一个组件的有效性。可以发现边界框回归器带来了至少 2.6 的 mAP 的提升。像 OICR 一样[29]不直接使用 NMS 方法的设置时将每一个类别的最高置信框看作种子边框。NMS 也可以提升 0.8 的 mAP。有关自底向上证明（BU）和α衰减函数的细节在上面已经讨论过，两个组件都可以带来 2.2mAP 的提升。

4.3 与当前先进技术对比

我们在 PASCAL VOC 2007 & 2012 [7]和 MS COCO [22]数据集测试 WSOD^2，报告了性能以及和现有的先进弱监督检测器进行对比。因为我们要对比的大多数方法采用了多尺度测试，所以我们我们报告了我们多尺度测试结果。

表 3 在 PASCAL VOC 2007 测试集上不同方法的均值平均精度。*表示在 07+12trainval 上进行的训练。

方法	飞机	单车	鸟	船	瓶子	公交	汽车	猫	椅子	牛	桌子	狗	马	摩托	人	植物	羊	沙发	火车	电视	mAP
WSDDN [2]	39.4	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
ContextLocNet [18]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
OICR [29]	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
PCL [28]	54.4	69.0	39.3	19.2	15.7	62.9	64.4	30.0	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63.0	43.5
Tang 等人的方法 [30]	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
C-WSL [9]	62.9	64.8	39.8	28.1	16.4	69.5	68.2	47.0	27.9	55.8	43.7	31.2	43.8	65.0	10.9	26.1	52.7	55.3	60.2	**66.6**	46.8
MELM [34]	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 [37]	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
WSCDN [35]	61.2	66.6	48.3	26.0	15.8	66.5	65.4	53.9	24.7	61.2	46.2	53.5	48.5	66.1	12.1	22.0	49.2	53.2	**66.2**	59.4	48.3
WSOD	**65.1**	64.8	**57.2**	**39.2**	**24.3**	**69.8**	66.2	**61.0**	**29.8**	64.6	42.5	**60.1**	**71.2**	**70.7**	21.9	28.1	**58.6**	**59.7**	52.2	64.8	**53.6**
WSOD^{2*}	68.2	70.7	61.5	42.3	28.0	73.4	69.3	52.3	32.7	71.9	42.8	57.9	73.8	71.4	25.5	29.2	61.6	60.9	56.5	70.7	56.0

表 4 在 PASCAL VOC 2007trainval 不同方法的正确定位结果。*表示在 07+12trainval 上进行的训练。

方法	飞机	单车	鸟	船	瓶子	公交	汽车	猫	椅子	牛	桌子	狗	马	摩托	人	植物	羊	沙发	火车	电视	CorLoc
WSDDN [2]	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
ContextLocNet [18]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
OICR [29]	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
ZLDN [37]	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
PCL [28]	79.6	**85.5**	62.2	47.9	37.0	83.8	83.4	43.0	38.3	80.1	50.6	30.9	57.8	90.8	27.0	58.2	75.3	**68.5**	**75.7**	78.9	62.7
C-WSL [9]	85.8	81.2	64.9	50.5	32.1	**84.3**	85.9	54.7	43.4	80.1	42.2	42.6	60.5	90.4	13.7	57.5	**82.5**	61.8	74.1	**82.4**	63.5
Tang 等人的方法[30]	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
WSCDN [35]	85.8	80.4	73.0	42.6	36.6	79.7	82.8	66.0	34.1	78.1	36.9	68.6	72.4	**91.6**	22.2	51.3	79.4	63.7	74.5	74.6	64.7
WSOD^2 (ours)	**87.1**	80.0	**74.8**	**60.1**	36.6	79.2	83.8	**70.6**	**43.5**	**88.4**	46.0	**74.7**	**87.4**	90.8	44.2	52.4	81.4	61.8	67.7	79.9	**69.5**
WSOD^{2*} (ours)	89.6	82.4	79.9	63.3	40.1	82.7	85.0	62.8	45.8	89.7	52.1	70.9	88.8	91.6	37.0	56.4	85.6	64.3	74.1	85.3	71.4

在 PASCAL VOC 上的 AP 验证。从表 3 中可以发现 PASCAL VOC 2007 上 WSOD^2 达到 53.6mAP，比其他端到端训练模型[28, 29, 35]表现显著，超出至少 5.3mAP。WSOD^2 在 PASCAL VOC 2012 也很鲁棒，实现 47.2 的 mAP，在表 5 中有所展示。

而且，我们遵循全监督目标检测的常用设定，在 PASCAL VOC 07+12 trainval 上训练 WSOD^2，将其表示为 WSOD^{2*}。这样的设定实现一个非常惊人的 mAP 分数 56.1，正如表 3 最后一行所示。

在 PASCAL VOC 上的 CorLoc 验证。CorLoc 是在训练集上验证检测器的定位精确度。在表 4 和表 5 中分别汇报了 PASCAL VOC 2007 和 2012 的结果。我们发现在 PASCAL VOC2007 和 2012 上 WSOD^2 都明显优于其他端到端训练模型[28, 29, 35]。

表 5 在 PASCAL VOC 2012 数据集上不同方法的对比。*指的是在 07+12trainval 上的训练。

方法	mAP	CorLoc
OICR [29]	37.9	62.1
PCL [28]	40.6	63.2
Tang 等人的方法[30]	40.8	64.9
ZLDN [37]	42.9	61.5
WSCDN [35]	43.3	65.2
WSOD^2	**47.2**[1]	71.9
WSOD^{2*}	52.7[2]	72.2

在 MS COCO 上的 AP 验证。表 6 中汇报了 MS COCO 数据集的结果。由于在 MS COCO 数据集上的工作结果汇报很少，我们只能与[10]和[28]进行性能对比。可以发现 WSOD^2 超出其他所对比的工作至少 2AP。

表 6 在 MS COCO 数据集上不同方法的实验结果

方法	AP@.50	AP@[.50:.05:.95]
Ge 等人的方法[10]	19.3	8.9
PCL [28]	19.4	8.5
PCL+Fast R-CNN [28]	19.6	9.2
WSOD^2	22.7	10.8

4.4 可视化与案例分析

和 OICR 对比，对 WSOD^2 的有效性进行了定量分析。提取了训练好的模型$conv5$特征，在图 4 中可视化了一些案例。高亮部分表示 CNN 中输入图片的高响应区域，相比与 OICR，WSOD^2 可以逐渐将相应区域从判别部分迁移到完整物体上。

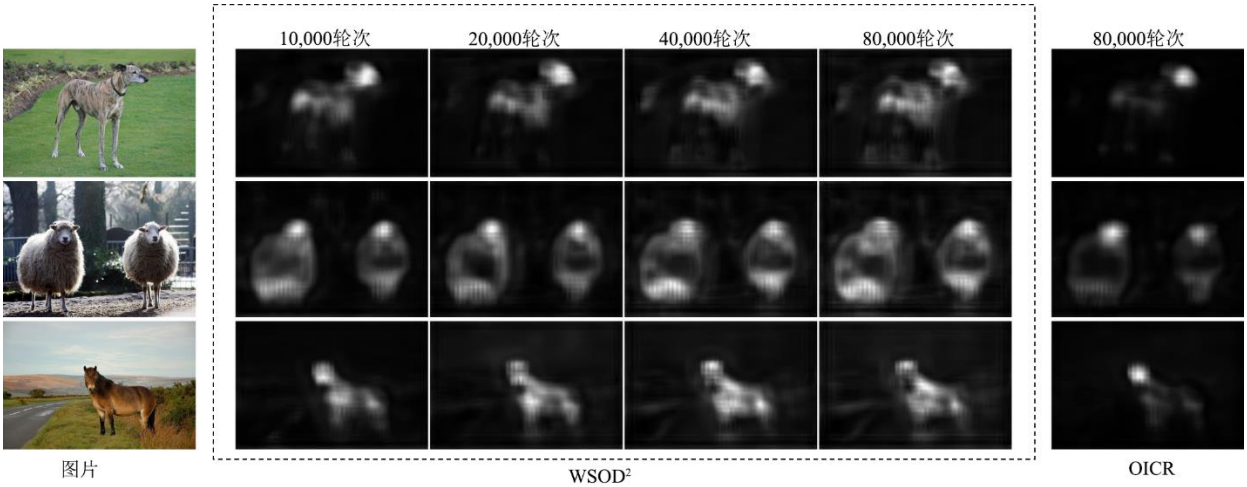


图 4 $conv5$ 特征图的可视化。响应图是由所有特征图通道取平均再标准化到$(0, 255)$后生成的。中间四列的特征图是由 WSOD² 在不同迭代伦茨提取来的。最后一列是通过 OICR [29]提取得到的特征图。

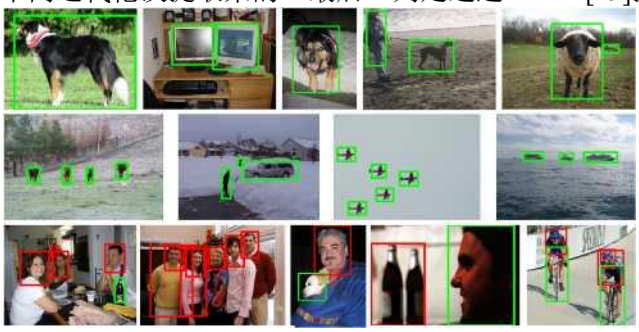


图 5 由 WSOD² 生成的示例结果。绿色边框表明正确的预测，红色边框表明失败案例。[通过颜色观看效果最佳]

图 5 展示了 WSOD² 的一些成功和失败的案例。可以观察到 WSOD² 可以很好的处理多离散实例，但在密集场景下解决检测问题任然存在挑战。我们也发现对于“人”类而言，大多数弱监督目标检测器常常是发现人的脸。原因是，在当前数据集下，人脸是“人”这一类别最普遍的模式，而其他部分在图片中经常是丢失的。这仍是一个有挑战性的问题，我们考虑未来应用人类结构的先验知识解决这一问题。

5. 结论

本文采用自底向上和自顶向下的似物性蒸馏法（即 WSOD²），提出新型弱监督目标检测，来提高 CNN 的深度似物性表示能力。自底向上的物体证明法可以测量一个边界框包含完整物体的概率，以自适应训练方式使用它来蒸馏 CNN 中的边界特征。我们也提出一种训练策略，以一种端到端的方式将边界框回归器和渐进式实例分类器结合起来。使用我们的方法，针对 WSOD 任务在一些标准数据集和设定进行实验。以定量和定性的方式，表明我们提出的 WSOD² 的有效性。我们也对 WSOD 问题的挑战和可能的提升（例如“人”类）进行了彻底的分析。

6. 致谢

本工作部分得到了中国国家自然科学基金（NSF）资助的 1672548、U1611461、61173081 和中国广州科技计划资助的 201510010165。