基于显著图的弱监督实时目标检测

李 阳 [1] 王 璞 [1] 刘 扬 [1] 刘国军 [1] 王春宇 [1] 刘晓燕 [1] 郭茂祖 [1, 2, 3]

摘 要 深度卷积神经网络 (Deep convolutional neural network, DCNN) 在目标检测任务上使用目标的全标注来训练网络参数, 其检测准确率也得到了大幅度的提升. 然而, 获取目标的边界框 (Bounding-box) 标注是一项耗时且代价高的工作. 此外, 目标检测的实时性是制约其实用性的另一个重要问题. 为了克服这两个问题, 本文提出一种基于图像级标注的弱监督实时目标检测方法. 该方法分为三个子模块: 1) 首先应用分类网络和反向传递过程生成类别显著图, 该显著图提供了目标在图像中的位置信息; 2) 根据类别显著图生成目标的伪标注 (Pseudo-bounding-box); 3) 最后将伪标注看作真实标注并优化实时目标检测网络的参数. 不同于其他弱监督目标检测方法, 本文方法无需目标候选集合获取过程, 并且对于测试图像仅通过网络的前向传递过程就可以获取检测结果, 因此极大地加快了检测的速率 (实时性). 此外, 该方法简单易用; 针对未知类别的目标检测, 只需要训练目标类别的分类网络和检测网络. 因此本框架具有较强的泛化能力, 为解决弱监督实时检测问题提供了新的研究思路. 在 PASCAL VOC 2007 数据集上的实验表明: 1) 本文方法在检测的准确率上取得了较好的提升; 2) 实现了弱监督条件下的实时检测.

关键词 弱监督, 实时目标检测, 显著图, 伪标注, 深度卷积神经网络

引用格式 李阳, 王璞, 刘扬, 刘国军, 王春宇, 刘晓燕, 郭茂祖. 基于显著图的弱监督实时目标检测. 自动化学报, 2020, **46**(2): 242−255

DOI 10.16383/j.aas.c180789

Weakly Supervised Real-time Object Detection Based on Saliency Map

LI Yang[1] WANG Pu[1] LIU Yang[1] LIU Guo-Jun[1] WANG Chun-Yu[1] LIU Xiao-Yan[1] GUO Mao-Zu[1, 2, 3]

Abstract Deep convolutional neural network (DCNN) trains model parameters by using object bounding-box annotations in object detection task, and its detection accuracy has been greatly improved. However, bounding-box annotations are very expensive and time-consuming. In addition, the real-time performance of object detection is another important problem that restricts its practicality. To solve these two problems, this paper proposes a new weakly supervised real-time object detector with image-level labels. The proposed method includes three sub-modules: 1) firstly, producing class-specific saliency maps based on a classification network and the back-propagation process, which provides object localization clues; 2) then, generating pseudo-annotations (pseudo-bounding-box) based on class-specific saliency maps; 3) finally, treating the pseudo annotations as ground-truth and optimizing the parameters of our real-time object detection network. Different from other weakly supervised object detection methods, our method avoids the computing process for obtaining object candidates. And we obtain object detection results of test images by feed-forward process, thus our method greatly speeds up the detection process (real-time). In addition, our method is simple and easy to use; for unknown class objects, we only need to train the classification and detection networks. So our method has a strong generalization ability and provides a new idea for weakly supervision real-time detection problem. Extensive experiments on PASCAL VOC 2007 benchmark show that: 1) the proposed method achieves a good improvement on detection accuracy; 2) it realizes real-time detection under weakly supervised condition.

Key words Weakly supervised, real-time object detection, saliency map, pseudo-annotations, deep convolutional neural network (DCNN)

Citation Li Yang, Wang Pu, Liu Yang, Liu Guo-Jun, Wang Chun-Yu, Liu Xiao-Yan, Guo Mao-Zu. Weakly supervised real-time object detection based on saliency map. *Acta Automatica Sinica*, 2020, **46**(2): 242−255

收稿日期 2018-11-27 录用日期 2019-06-24
Manuscript received November 27, 2018; accepted June 24, 2019
国家重点基础研究发展计划 (2016YFC0901902), 国家自然科学基金 (61671188, 61571164, 61976071, 61871020) 资助
Supported by National Key Research and Development Program of China (2016YFC0901902) and National Natural Science Foundation of China (61671188, 61571164, 61976071, 61871020)
本文责任编委 王立威
Recommended by Associate Editor WANG Li-Wei

1. 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001 2. 北京建筑大学电气与信息工程学院 北京 100044 3. 建筑大数据智能处理方法研究北京重点实验室 北京 100044
1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001 2. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044 3. Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing 100044

近年来, 深度卷积神经网络 (Deep convolutional neural network, DCNN) 在图像目标检测任务中取得了突破性的进展并成为主流的模型架构. 图像目标检测的性能在 DCNN 的帮助下取得了很大的提升[1−6].

目标检测方法需要目标的边界框 (Bounding-box) 标注及相对应的类别注释来优化神经网络的参数. 这种全监督方法的主要问题是获取大量的精确的标注极为耗时并且成本昂贵. 因此, 尽管这些方法能够获取非常准确的检测结果, 但是需要耗费大量的人力及时间资源, 因而其适用性受到一定限制. 为了解决这个问题, 研究者们尝试放宽图像标注的程度, 仅使用图像级标注 (Image-level annotations) 来训练模型, 并提出了一系列弱监督目标检测方法[7−14].

虽然弱监督目标检测方法具有图像级标注, 但是在缺少目标位置信息的情况下, 该方法使得目标检测成为极具挑战性的问题. 近年来, 很多研究工作致力于从图像级标注中学习到复杂的语义信息[10, 12−14]. 一些早期的方法采用多示例学习 (Multiple instance learning, MIL)[5, 8, 11], 在模型优化和正例选择之间迭代计算. 另一些方法通过 DCNN[15−16] 学习有效的特征表示[10, 17], 并提出构建两个子模块网络 (目标定位和目标挖掘模块) 来提升检测的准确率[12, 14]. 此外, 有的方法充分发挥 MIL 与 DCNN 模型的特性, 通过融合 MIL 网络与示例分类优化网络来构造目标检测网络[13]. 还有一些方法在 DCNN 基础上引入了有利于检测的先验信息, 如目标尺寸[8]、上下文信息[18]、位置线索[9, 19]、目标区域候选集[12−13] 等. 这些方法的实验结果表明:融合目标位置信息的方法比其他方法得到更精确的检测结果[10]. 因此, 受上述方法的启发, 本文提出一种通过构建目标位置的伪标注 (Pseudo-bounding-box) 来训练检测网络的方法. 所谓伪标注, 是针对目标的真实标注 (Ground-truth) 而言, 而目标的真实标注在弱监督设定下是无法获取的. 因此如何获取高质量的伪标注, 使得该伪标注能够较为准确地给出目标的位置信息并优化目标检测网络是本文的研究重点之一. 受分类网络可以给出一些目标位置线索的启发[19−20], 一些弱监督语义分割方法[21−22] 借助这些位置线索提升了分割的准确率. Shimoda 等[23] 提出基于分类网络和改进后的反向传递过程生成类别显著图的方法, 这些类别显著图提供了目标位置的可靠信息, 并在语义分割中得到较好的分割结果. 但是文献 [23] 的方法并不能实现实时的目标检测任务. 受该方法启发, 本文将类别显著图应用到目标检测任务中, 通过构建高质量的伪标注来训练实时的目标检测器.

目标检测任务中, 检测准确率和检测速度是两个重要的评价指标. 为了提升检测速度, 一些研究者提出快速目标检测方法[1−2, 4, 23−25], 但是这些方法需要精确的标注来训练模型的参数, 在弱监督的设定下还不能够实现实时检测. 本文借鉴文献 [3] 的方法, 提出弱监督条件下的实时目标检测模型, 对于测试图像, 仅通过网络的前向传递过程即可以获取检测结果, 而节省了其他预处理/后处理的计算时间. 总体来讲, 本文为了提升检测准确率采用分类网络获取目标的位置线索, 并构建伪标注信息来训练一个实时检测器, 从而提升了检测速度, 提出的方法具有以下创新点:

1) 为获取目标在图像中的位置线索, 本文利用分类网络生成类别显著图, 并根据该显著图生成目标的伪标注;

2) 利用伪标注优化实时目标检测网络, 该实时检测网络在提升检测速度的同时, 无需目标候选集合的生成过程;

3) 本文所提方法实现了弱监督条件下的实时目标检测任务;

4) 该模型简单易用, 具有一定的泛化能力, 并为弱监督实时目标检测问题提供了新的研究思路.

实验结果表明, 本文所提方法在 PASCAL VOC 2007 数据集上不仅具有优秀的检测准确率, 而且实现了目标检测的实时性.

1 相关工作

1.1 弱监督目标检测

目前大部分弱监督目标检测方法将该问题归纳为多示例学习 (Multiple instance learning, MIL) 问题[5, 8, 11, 13]. MIL 方法将每幅图像看作由目标候选集[26−27] 所构成的多个对象示例的包. 如果一幅图像在某个类别上被标注为正例, 那么该图像至少包含该类别的一个对象示例; 反之, 负例集图像只包含除此类别外的其他类别的对象示例. MIL 方法首先通过当前模型所选择的正示例学习模型参数, 再基于当前模型选择好的正示例, 这两个过程交替进行直到模型收敛[28].

然而, 这种交替学习的方法使得 MIL 方法过度依赖于模型初始化的好坏, 而且是非凸的求解问题, 所以模型的参数常常陷入局部最优解[29]. 为了解决这个问题, 一些研究学者致力于如何获取更好的模型初始化参数[5, 8]. 近年来, 随着深度卷积神经网络

(DCNN) 在很多视觉领域上带来的重大突破[12–14, 29–31], DCNN 常用来做特征提取[32]. 同时, 一些方法进一步融合 MIL 方法与深度学习方法, 通过 DCNN 进行示例选择[33–35] 并构建有效的端对端 (End-to-end) 网络模型[9–10, 18]. Diba 等[9] 提出级联网络结构, 有效地将 MIL 融合进深度网络模型中. 此外, Oquab 等[36] 利用网络中间层特征表示来确定对象的位置并完成分类任务, 开辟了弱监督视觉任务的新思路. Bilen 等[10]则设计了双数据流的网络结构用来做目标检测和类别判定. Kantorov 等[18] 在文献 [10] 方法的基础上加入了对象的上下文信息. Tang 等[13] 构造一种新型目标检测网络, 该网络结构在 MIL 网络的基础上加入了多个示例分类优化子网络, 每个子网络为其下一个子网络提供目标候选集的聚类结果, 基于这个聚类结果计算其损失函数, 迭代地优化整个网络参数. Teh 等[17] 专门设计了注意力网络 (Attention network)来提升目标检测的准确率, 该方法通过融合目标候选区域的注意力得分与网络特征来得到更有效的特征表示, 以便于选择出最优的目标候选区域作为最终的检测结果. Wan 等[12, 14] 为了解决 DCNN 模型得到的目标检测结果存在检测区域不完整或者包含背景等问题, 提出最小熵潜在模型 (Min-entropy latent model, MELM). MELM 模型将网络结构分为两个分支:目标挖掘和目标定位. 目标挖掘部件用于从众多目标候选集合中选择最具有判别性的目标区域; 而目标定位部件用于给出目标的准确位置信息. 但是这些方法[12–14, 17] 均采用目标候选集合作为网络的输入, 这样无法实现实时性. 受上述方法的启发, 本文同样利用 DCNN 的网络模型构建有效的实时检测网络, 既充分地利用了网络其强大的特征表示能力, 又通过构建伪标注来节省目标候选集合获取的时间, 从而实现了实时性.

1.2 实时目标检测

针对目标检测的实时 (Real-time) 问题, 一些研究工作试图加速变形部件模型 (Deformable part model, DPM)[25, 37], 包括加速 HOG 特征提取[25] 及构建级联 DPM 模型[37]. DPM 在 DCNN 出现之前曾广泛用于目标检测[38]. Sadeghi 等[24] 基于级联结构及矢量量化技术提出了实时 DPM 模型, 该模型 1 s 可以处理 30 幅图像但是检测准确率较低 (平均准确率为 26.1 %). Girshick 等[1, 4] 提出 Fast/Faster R-CNN 模型来加速原始 R-CNN[2] 模型的运行速度. 这两个方法将多次前向运算过程通过共享的方法压缩计算时间, 并且利用网络特征生成目标的候选集合, 从而节省了使用 selective search[26] 或 edge boxes[27] 生成图像目标候选集合的时间. Redmon 等[39] 通过对图像划分区域的方式提出了简单有效的 YOLO 方法. 而 Liu 等[3] 通过将预测检测框离散化为多个尺度和多个收缩比例的特征层检测框实现了实时检测的目标. 然而, 上述所列方法均需要图像中目标的真实标注来优化模型参数, 即为全监督方法, 但在弱监督目标检测中, 目前还没有方法可以实现实时性. 因此本文在弱监督的设定下, 提出了一个实时目标检测方法.

1.3 类别显著图

当训练数据集中只有图像级标注时, 弱监督的视觉问题[21–23] 希望通过分类网络来获取目标在图像中的一些位置线索. Zhou 等[19] 利用全卷积神经网络及全局平均池化 (Global average pooling) 运算生成类别激活图 (Class activation mapping, CAM), 该 CAM 能够定位出图像中目标的某些判别性区域. Simonyan 等[20] 研究表明可以通过网络的反向传递过程找到目标在图像中的大概位置. Springenberg 等[40] 还提出通过对前向计算所得到的最大损失值进行网络反向传递求导来定位目标的方法. Shimoda 等[23] 利用反向传递的思想, 提出改进后的类别显著图生成方法, 并且将该方法用于弱监督语义分割中. 文献 [23] 的方法所生成的类别显著图可以较为清晰地描述出目标的轮廓, 从而可以提供目标的位置线索. 由此可见, 上述基于分类网络挖掘目标位置线索的方法为弱监督视觉任务提供了有效的先验信息.

本文工作致力于解决弱监督实时目标检测问题. 在只有图像级标注时, 所提方法借助分类网络的反向传递过程生成类别显著图, 并根据该显著图构建伪标注, 最后使用伪标注来训练实时目标检测网络模型. 我们的方法能够实现目标检测的实时性, 并且简单易用; 针对未知类别的目标具有较强的泛化能力. 据我们所知, 所提方法是第一个实现了弱监督条件下的实时目标检测的模型, 该模型为弱监督的实时检测任务提供了新的研究思路.

2 基于显著图的弱监督实时目标检测方法

2.1 模块构成

本文所提出的弱监督实时目标检测方法由三个子模块构成: 1) 首先借助深度卷积神经网络的图像分类任务获取类别显著图 (Class-specific saliency

maps); 2) 其次基于类别显著图生成目标的伪标注 (Pseudo-bounding-box); 3) 最后将伪标注看作真实标注 (Ground-truth), 以全监督的方式训练实时 (Real-time) 目标检测网络. 该方法的整体架构及三个子模块的具体结构如图 1 所示.

2.2　类别显著图获取

目前一些弱监督目标检测方法[7, 9] 在训练检测网络之前需要获取目标的候选集, 这些目标候选集合明确地提供了图像中目标的位置信息. 常用的目标候选集合获取方法有 selective search[26] 和 edge boxes[27]. 但是这些候选集获取算法计算时间较长, 例如文献 [26] 的方法平均需 11.2 s[41] 完成一幅图像的计算并获取 2 000 个左右的目标候选. 因此这样较高的计算代价使得设计实时检测模型变得极为困难.

另一方面, 近年来一些自顶向下 (Top-down) 方法提出利用分类网络获取类别显著图, 从而为图像中的目标提供位置线索. 本文受该思想启发, 借鉴文献 [23] 的方法, 首先计算图像的类别得分对图像中间层卷积特征的导数, 再生成类别显著图和伪标注, 从而通过图像级 (Image-level) 标注获取目标的位置信息.

本文首先基于 VGG-16[42] 网络训练图像分类网络, 其损失函数定义为

$$L_c(\boldsymbol{\theta}) = \frac{1}{N}\sum_{j=1}^{N} -\bar{\boldsymbol{z}}_j \ln\left(f(I_j)\right) - (1-\bar{\boldsymbol{z}}_j)\ln(1-f(I_j)) \tag{1}$$

其中, $\bar{\boldsymbol{z}}_j$ 为图像的类别标签向量 (向量元素中 "1" 表示图像存在该类别的对象, 否则为 "0"), $f(I_j)$ 为网络预测的类别得分, I_j 为第 j 幅图像, N 为图像总数, $\boldsymbol{\theta}$ 表示为网络参数. 从式 (1) 可以看出, 本文将多标签分类问题看作 $|C|$ 个独立的二分类问题 ($|C|$ 为数据集的总类别数).

对于一幅图像 I_j 和该图像的一个真实类别 c, 分类网络输出该类别的得分 S_c. 那么类别得分 S_c 对于第 i 层特征 F_i 在激活值 F_i^0 的导数为

$$D_i^c = \left.\frac{\partial S_c}{\partial F_i}\right|_{F_i^0} \tag{2}$$

获取 D_i^c 之后, 通过线性插值运算上采样 D_i^c 到原始图像尺度, 记为 M_i^c. 从式 (1) 可以看出, 对于多类别图像 (该图像的类别集合表示为 $\bar{\boldsymbol{c}}_j$), 本文都会获取每个类别 c 的显著图 M_i^c. 但是多个类别的显著图会存在相互覆盖的情况, 为了解决这个问题并且突出当前类别 c 的显著图与其他类别的区别, 本文对 M_i^c 进行提纯处理

$$\bar{M}_{i,x,y}^c = \sum_{c'\in\bar{\boldsymbol{c}}_j'} \max(M_{i,x,y}^c - M_{i,x,y}^{c'}, 0)[c \neq c'] \tag{3}$$

其中, $\bar{\boldsymbol{c}}_j' = \bar{\boldsymbol{c}}_j \setminus c$, 下标 $\{x, y\}$ 为图像的横纵坐标. 通过式 (3) 的提纯运算 (当前类别显著图分别与其

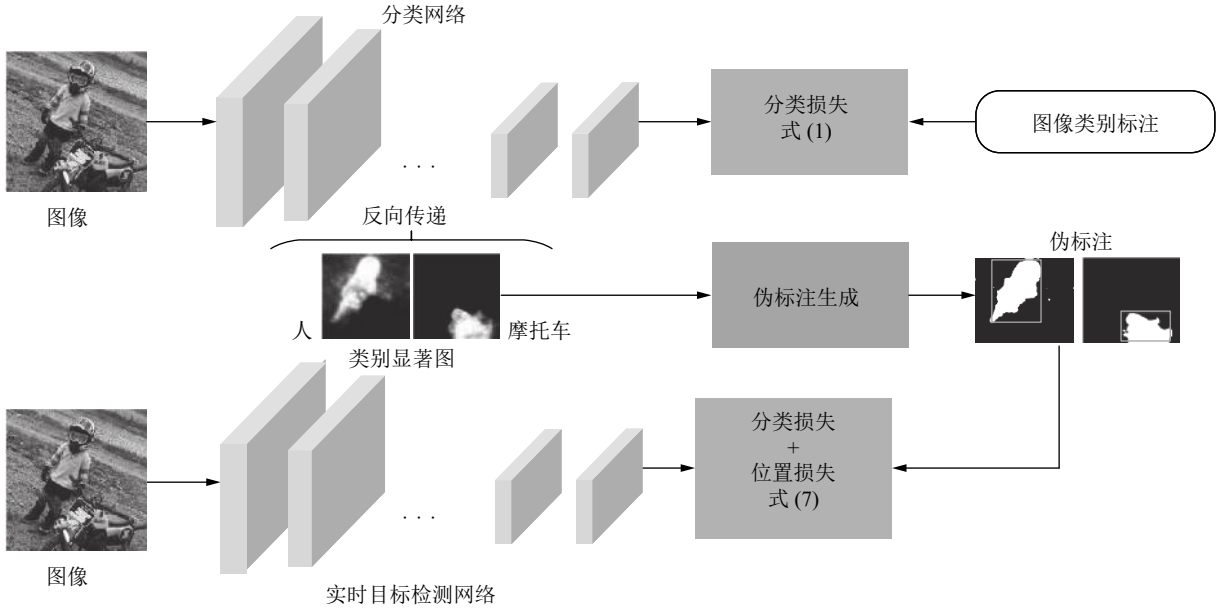


图 1　弱监督实时目标检测方法结构图

Fig. 1　Pipeline for weakly supervised real-time object detection

他类别显著图的相减运算), $\bar{M}^c_{i,x,y}$ 可以更显著地描述该类别目标的位置信息.

一些研究表明[20], 深度卷积神经网络在低层特征中 (conv2, conv3) 更能体现目标的轮廓等细节特征, 而高层特征 (conv4, conv5, conv6) 更倾向于描述目标的高层语义信息. 为了融合轮廓特征和高层语义信息, 我们对第 2 卷积层 (conv2) 到第 6 卷积层 (conv6) 特征分别通过上述运算得到多层显著图 $\{\bar{M}^c_i|i \in [2,3,4,5,6]\}$, 最后平均各个层的显著图

$$\tilde{M}^c_{x,y} = \frac{1}{|L|}\sum_{i\in L}\tanh(\alpha \bar{M}^c_{i,x,y}) \tag{4}$$

其中, L 为层数集合; α 为一个常数, 实验中 $\alpha = 1.5$; tanh 为正切激活函数. 图 2 展示了一些在 PASCAL VOC 2007 数据集上获取的显著图, 图中第 2~4 列对应类别得分排名前三的显著图. 从图 2 可以看出, 对于正确的类别 (前两行图像的正确类别显著图为第 2 列和第 3 列, 最后一行图像的正确类别显著图为第 2 列), 显著图可以较好地描绘出目标的轮廓和位置信息. 实验中, 我们对文献 [23] 方法中的原网络结构进行调整, 去掉原网络中的 "sort-id" 及 "signal" 层, 加入图像级标注作为网络的输入, 并且只对图像中所包含的类别利用式 (2) 求导. 因此我们只关注图像中确定存在的类别的目标显著图.

2.3 伪标注生成

本文的下一步工作是根据上述获取的类别显著图生成目标位置的伪标注 (Pseudo-bounding-box). 所谓伪标注是针对真实的人工标注 (Ground-truth) 而言, 本文的目标是尽可能获取较为准确的伪标注, 从而提升其检测的准确率.

一幅图像中常常存在同一类别的多个目标实例, 例如草原风景图中常常出现多只羊, 街道图像中存在多辆汽车, 如何将这些相同类别的目标分别用 Bounding-box 标注出来是伪标注生成方法需要解决的首要问题. 如第 2.2 节所描述, 图像 I 可以获取每个类别 c ($c \in \bar{\boldsymbol{c}}_j$) 的显著图, 但是这些显著图无法区别多个目标实例. 为了解决这个问题, 本文所提出的伪标注生成方法包含两个步骤: 1) 二值化类别显著图; 2) 融合多个联通区域生成目标的 Bounding-box 标注.

首先根据预先设定的阈值 th_c, 将类别显著图二值化

$$B^c_{x,y} = \begin{cases} 1, & \text{若 } c \in \bar{\boldsymbol{c}}_j \text{ 且 } \tilde{M}^c_{x,y} > th_c \\ 0, & \text{否则} \end{cases} \tag{5}$$

如果 $\tilde{M}^c_{x,y} > th_c$, 表示位置 $\{x,y\}$ 上的像素类别为 c. 同一图像中的不同类别的目标具有不同的大小、尺度、颜色, 为了获取更高质量的伪标注, 本文根据类别不同设定不同的二值化阈值 th_c. 通过实验验证可以发现对于目标尺寸较小的类别, th_c 的值应该较大, 以确保获取更精确的位置信息; 反之, 对于目标尺寸较大的类别, th_c 的值应该较小, 以确保发现更完整的目标位置. 表 1 列出了本文所使用的针对 PASCAL VOC 20 个类别的阈值 th_c.

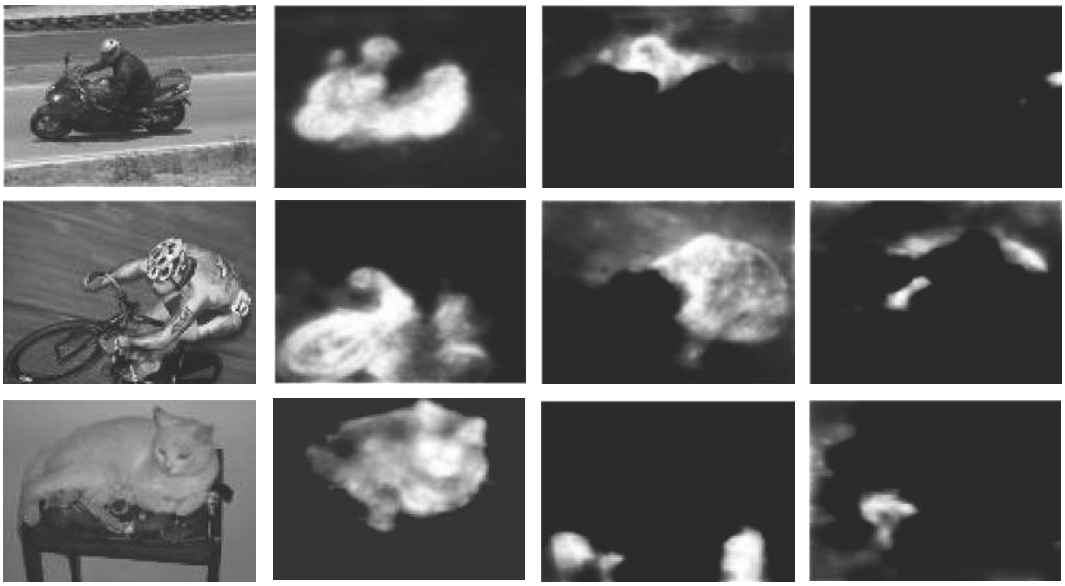


图 2 类别显著图

Fig. 2 Class-specific saliency maps

表 1　二值化类别显著图的阈值设置 (PASCAL VOC 数据集 20 个类别)

Table 1　Thresholds of the binarization class-specific saliency maps (20 categories for PASCAL VOC dataset)

类别	阈值	类别	阈值	类别	阈值
Plane	0.7	Cat	0.4	Person	0.5
Bike	0.7	Chair	0.6	Plant	0.7
Bird	0.5	Cow	0.5	Sheep	0.5
Boat	0.8	Diningtable	0.3	Sofa	0.5
Bottle	0.8	Dog	0.3	Train	0.7
Bus	0.6	Horse	0.3	TV	0.8
Car	0.6	Motorbike	0.5		

为了区别同类别的不同对象, 在生成 Bounding-box 时, 利用连通区域分析方法 (Connected component analysis-labeling, CCL) 对二值化类别显著图进行处理, 来标注相邻的连通前景区域. 使用 CCL 方法标注之后, 图像中常常存在多个分散的小区域. 根据设定好的阈值, 保留其像素个数大于该阈值的区域 (实验中阈值为 1 000 个像素) 并抛弃那些较小的区域. 图 3 列出 PASCAL VOC 2007 数据集上所获取的伪标注的可视化表示. 从图 3 可以看出, 真实标注 (原图中的边框) 与伪标注 (二值化显著图中的边框) 几乎相同, 从而可以证明本文的方法可以较好地生成用于训练目标检测网络的伪标注.

2.4　实时目标检测网络

最后一个模块利用伪标注训练实时目标检测网络. 本文采用 SSD (Single shot multibox detector)[3] 作为基础实时目标检测网络, 将伪标注作为真实标注来训练该模型. 这个目标检测网络可以直接预测目标的边界框/检测框 (Bounding-box) 及目标的类别, 它不仅表现出优秀的检测性能, 而且可以做到实时运算, 无需获取目标候选集.

这里 SSD 方法采用 VGG-16 网络结构. 为了捕捉不同层特征, SSD 方法选择多个卷积层并对这些特征进行卷积滤波操作, 生成一组目标检测框. 为了生成目标的 Bounding-box, 该方法将卷积特征划分成多个特征单元 (例如将二维卷积特征划分为 4×4、8×8 的特征单元), 每个特征单元根据所设定的尺度比例及长宽比对应一组检测框, 本文称为 "基检测框" (图 4 中虚线框, 一组基检查框对应一个特征单元). 假设本文所选择用于生成基检测框的特征层有 L 层, 那么每一层特征图上的基检测框的尺度比例为

$$s_l = s_{\min} + \frac{s_{\max} - s_{\min}}{L-1}(l-1),\ l \in [1, L] \qquad (6)$$

其中 $s_{\min}=0.2$, $s_{\max}=0.9$, 表示较低层的基检测框有较小的尺度比例, 较高层的基检测框有较大的尺度比例, 这也同时说明了较低层的特征可以捕捉到目标的更多细节, 而较高层的特征捕捉到更深层的语义信息. 其基检测框的长宽比集合为 $A=1, 2, 3, 1/2, 1/3$, 那么每个基检测框的宽为 $W_l = s_l\sqrt{A}$;高为 $H_l = s_l\sqrt{A}$. 由此可以看出, 一个特征图上的一个特征单元可以获取 6 个基检测框. 每个基检测框都对应 $|C|$ 个类别打分和 4 个偏移量 (w, h, x, y) , 那么一个特征图上的 $m\times n$ 个特征单元经过 k 个卷

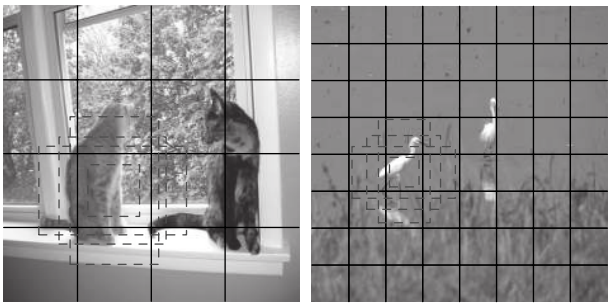


图 4　4×4 及 8×8 的特征单元和相对应的目标基检测框

Fig. 4　Feature map cells for 4×4 and 8×8 and its corresponding object default bounding-boxes

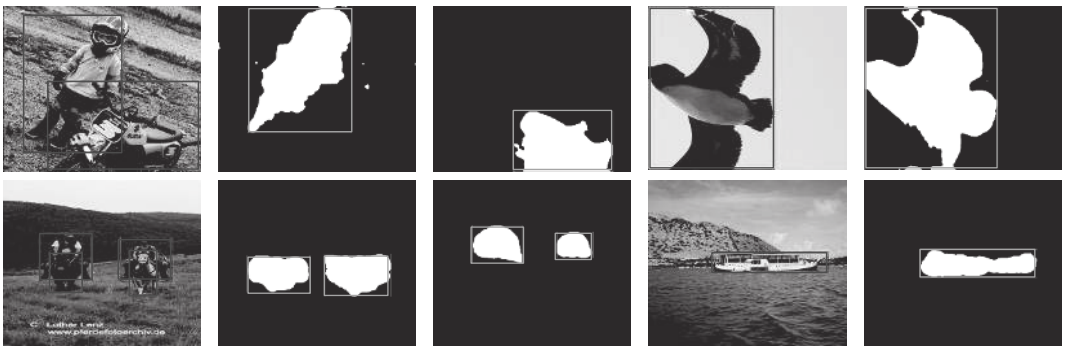


图 3　二值化类别显著图以及相应的伪标注

Fig. 3　Binarization class-specific saliency maps and the corresponding pseudo-bounding-boxes

积核的卷积操作之后其输出大小为 $m \times n \times k \times (|C|+4)$.

为了训练 SSD 网络, 本文使用第 2.3 节所生成的伪标注作为真实标注来优化网络参数. 这里用 $x_{i,j}^c$ 表示每个基检测框与伪标注的匹配程度, 如果 $x_{i,j}^c = 1$ 表示类别 c 的第 j 个伪标注与第 i 个基检测框的覆盖度 $IoU > 0.5$, 否则 $x_{i,j}^c = 0$. 对于图像的类别 c, $\sum_i x_{i,j}^c \geq 1$. 那么实时目标检测网络的损失函数包含两部分: 位置损失和分类损失, 具体定义为

$$L(x,c,l,g) = \frac{1}{N}(L_{\text{loc}}(x,l,g) + L_{\text{cls}}(x,c)) \tag{7}$$

其中, N 为匹配到的基检测框的个数. 当 $N=0$ 时, $L(x,c,l,g)=0$. $L_{\text{loc}}(x,l,g)$ 是 L_1 平滑损失, 用于回归预测检测框 (网络输出的检测结果) 及正例基检测框

$$L_{\text{loc}} = \sum_{i\in pos}^{N}\sum_{j=1}^{M}\sum_{m\in\{cx,cy,w,h\}} x_{i,j}^c smooth_{L_1}(\hat{l}_i^m - \hat{g}_j^m) \tag{8}$$

对于匹配到的基检测框 d_i 和预测检测框 $l_i, \hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w$; $\hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^y$; $\hat{g}_j^w = \ln(g_j^w/d_i^w)$; $\hat{g}_j^h = \ln(g_j^h/d_i^h)$; $\hat{l}_i^{cx} = (l_i^{cx} - d_i^{cx})/d_i^w$; $\hat{l}_i^{cy} = (l_i^{cy} - d_i^{cy})/d_i^y$; $\hat{l}_i^w = \ln(l_i^w/d_i^w)$; $\hat{l}_i^h = \ln(l_i^h/d_i^h)$ ($\{cx, cy, w, h\}$ 分别为检测框的几何中心值, 宽度及高度, g_j 为第 j 个伪标注). 此外 $L_{\text{cls}}(x,c)$ 定义为

$$L_{\text{cls}}(x,c) = -\sum_{i\in pos}^{N} x_{i,j}^c \ln(\hat{s}_i^c) - \sum_{i\in neg}\ln(\hat{s}_i^0) \tag{9}$$

其中, $\hat{s}_i^c$ 为第 i 个基检测框对于类别 c 的网络预测得分. 式 (8) 和 (9) 将基检测框按照与伪标注的 IoU 匹配值分为正例集 ($i \in pos$) 和反例集 ($i \in neg$). 显然通过这种方式会产生大量的反例样本. 为了平衡正例集与反例集的数量, 对所有反例基检测框进行类别得分排序, 选得分最高的一部分反例基检测框作为反例集, 使得反例集与正例集的数量比保持为 3 : 1.

3 实验及分析

3.1 数据集及评价指标

3.1.1 数据集

本文所有实验均采用 PASCAL VOC 数据集, 该数据集在计算机视觉的分类, 检测及分割任务中都是基准数据集, 提供了用于检测的标准 Bounding-box 标注和算法性能评价的官方系统. PASCAL VOC 数据集包含了自然场景中的 20 个类别的图像 (PASCAL VOC 2007 训练集: 2 501, 验证集: 2 510, 测试集: 5 011; PASCAL VOC 2012 训练集: 5 717). 本文使用训练集和验证集来训练目标检测模型, 并且只使用该数据集的图像级标注.

3.1.2 平均指标

标准的 PASCAL VOC 评价指标为平均准确率 (Mean average precision, mAP), 用于实验对比和分析的准则. 首先需要计算预测检测框 p 与真实边界框 g 的覆盖率 $IoU(p,g)$

$$IoU(p,g) = \frac{p \bigcap g}{p \bigcup g} \tag{10}$$

如果覆盖率 $IoU(p,g) \geq 0.5$ 则认为该目标已经被检测到. 那么根据每个类别的召回率和准确率计算每个类别的平均精确度 (Average precision, AP), 即可计算 20 个类别的平均准确率.

此外, 除了检测的平均准确率外, 图像目标检测的另一个重要评价指标为检测速度. 本文采用每秒帧率 (Frame per second, FPS) 来评价各方法在测试集上的检测速度, FPS 为每秒处理图像的个数.

3.2 实验设置

首先, 为获取类别显著图, 我们对 VGG-16 网络结构进行改进, 将所有的全连接层替换为连续的卷积层, 并且图像输入尺度剪裁为 512×512. 使用深度学习框架 Caffe 来搭建网络结构, 并采用 ImageNet[31]数据集上预训练的参数初始化网络, 利用随机梯度下降算法 (Stochastic gradient descent, SGD) 优化网络参数. 其网络训练的参数设置为: 初始学习率 (Learning rate) 为 0.001, 每 2 000 次迭代, 学习率降低 10 倍; 动量系数 (Momentum) 为 0.9; 权重衰减系数 (Weight decay) 为 0.0005; dropout 层的 drop_rate为 0.5; 最多迭代次数 20 000; 每次迭代的图像数量 (Mini-batch size) 为 20.

在训练实时目标检测网络时, 输入图像尺寸为 300×300, 优化方法同样采用 SGD 算法, 其参数设置为:初始学习率为 0.001, 分别在 8 000 次与 10 000 次迭代时学习率下降 10 倍, 动量系数为 0.9, 权重衰减系数为 0.0005, 最大迭代次数为 12 000. 该实时目标检测网络同样使用 Caffe 来构建网络模型. 实验配置为显存 11 GB 的 NVIDIA GeForce GTX 1 080 Ti.

3.3 实验结果及分析

3.3.1 与其他目标检测方法的平均准确率比较

本文与当前最好的 8 种弱监督目标检测方法进

行了比较, 如表 2 所示. 这 8 种方法中, Bilen 方法[11] 基于 SVM 和聚类方法; Cinbis 方法[7] 采用原始 MIL 优化; Wang 方法[32] 采用概率潜在语义分析模型 (Probabilistic latent semantic analysis, pLSA) 来做目标检测; Teh 方法[17] 提出注意力网络; WSDDN 方法[10] 采用双数据流卷积网络; WCCN 方法[9] 使用级联的检测网络结构; MELM 方法[12, 14] 采用具有目标挖掘与目标定位这两个分支的网络结构, 提出在训练目标检测网络与重新分配候选集标签之间的递归学习策略; PCL 方法[13] 融合了 MIL 与 DCNN 的模型优点, 提出了基于候选集聚类来学习网络参数的方法. 其中, Teh 方法[17], WCCN 方法[9], MELM 方法[12, 14], PCL 方法[13] 均使用目标候选集合作为网络的输入. PASCAL VOC 2007 测试集上的检测结果表明, 本文所提方法在仅使用 VOC 2007 训练集训练模型时, 所得到的平均准确率为 33.9 %; 而当添加 5 717 幅 VOC 2012 训练集图像时, 其平均准确率达到 39.3 %, 提高了近 6 个百分点. 从表 2 可以看出, 本文比 Bilen 方法[11]高出近 12 个百分点, 比 Cinbis 方法[7] 提升了 9 个百分点, 比 Wang 方法[32]高出近 8 个百分点, 此外比 Teh 方法[17] 及 WSDDN 方法[10] 分别提升了近 5 个百分点. 与 WCCN 方法[9] 相比, 本文的平均准确率稍微降低了近 3 个百分点, 这主要是因为 WCCN 使用了 Edge boxes[27] 生成的目标候选集来训练级联网络结构. 与 MELM 方法[12, 14] 及 PCL 方法[13] 相比, 本文的平均准确率分别降低了 8 个百分点和 9.5 个百分点, 这是因为 MELM 方法与 PCL 方法都借助于 Selective search[26] 获取图像中目标的候选集并在候选集合上学习到最优的区域作为检测结果. 但是由于获取目标候选集需要计算时间, 所以这些方法无法实现检测的实时性. 另外,表 2 中, WSDDN_Ens[10] 集合了三个类型的卷积网络并平均它们的检测结果. 由此可见, 本文只采用一种卷积网络就可以获取与 WSDDN_Ens 相同的平均准确率. 通过表 2 可以发现, 本文方法在 “Cat” 、“Dog” 这两个类别上取得了很大的提升, 而在 “Cow” 、“Sheep” 这两个类别上其检测准确率要比其他 8 个方法低, 这种情况的出现可能是因为这两个类别常常在图像中存在多个目标, 而且彼此之间的差异性很低, 从而在生成伪标注时就存在误差, 导致检测不准确.

3.3.2　与其他实时目标检测方法的检测速度比较

为了提高目标检测的运行速度, 很多研究者提出了一些快速算法[24−25, 43], 但是在弱监督的设定下,

表 2　PASCAL VOC 2007 测试数据集的目标检测准确率 (%)

Table 2　Object detection precision (%) on PASCAL VOC 2007 test set

方法	Bilen[11]	Cinbis[7]	Wang[32]	Teh[17]	WSDDN[10]	WSDDN_Ens[10]	WCCN[9]	MELM[12, 14]	PCL[13]	本文 (07数据集)	本文 (07+12数据集)
mAP	27.7	30.2	31.6	34.5	34.8	39.3	42.8	47.3	48.8	33.9	**39.3**
Plane	46.2	39.3	48.9	48.8	39.4	46.4	49.5	55.6	**63.2**	54.5	54.2
Bike	46.9	43.0	42.3	45.9	50.1	58.3	60.6	66.9	**69.9**	52.9	60.0
Bird	24.1	28.8	26.1	37.4	31.5	35.5	38.6	34.2	**47.9**	30.0	41.9
Boat	16.4	20.4	11.3	26.9	16.3	25.9	**29.2**	29.1	22.6	15.2	20.7
Bottle	12.2	8.0	11.9	9.2	12.6	14.0	16.2	16.4	**27.3**	7.8	11.4
Bus	42.2	45.5	41.3	50.7	64.5	66.7	70.8	68.8	**71.0**	47.3	55.9
Car	47.1	47.9	40.9	43.4	42.8	53.0	56.9	68.1	**69.1**	44.5	49.2
Cat	35.2	22.1	34.7	43.6	42.6	39.2	42.5	43.0	49.6	62.5	**71.3**
Chair	7.8	8.4	10.8	10.6	10.1	8.9	10.9	**25.0**	12.0	9.4	10.5
Cow	28.3	33.5	34.7	35.9	35.7	41.8	44.1	**65.6**	60.1	17.6	23.2
Table	12.7	23.6	18.8	27.0	24.9	26.6	29.9	45.3	**51.5**	39.3	48.5
Dog	21.5	29.2	34.4	38.6	38.2	38.6	42.2	53.2	37.3	48.9	**54.5**
Horse	30.1	38.5	35.4	48.5	34.4	44.7	47.9	49.6	**63.3**	45.7	52.1
Motorbike	42.4	47.9	52.7	43.8	55.6	59.0	64.1	**68.6**	63.9	49.9	56.4
Person	7.8	20.3	19.1	**24.7**	9.4	10.8	13.8	2.0	15.8	17.2	15.0
Plant	20.0	20.0	17.4	12.1	14.7	17.3	23.5	**25.4**	23.6	15.7	17.6
Sheep	26.6	35.8	35.9	29.0	30.2	40.7	45.9	**52.5**	48.8	18.4	23.3
Sofa	20.6	30.8	33.3	23.2	40.7	49.6	54.1	**56.8**	55.3	29.3	42.5
Train	35.9	41.0	34.8	48.8	54.7	56.9	60.8	**62.1**	61.2	43.7	47.6
TV	29.6	20.1	46.5	41.9	46.9	50.8	54.5	57.1	**62.1**	28.2	30.4

目前的方法还不能达到实时检测的要求. 本节实验综合比较了检测的平均准确率 (mAP) 和检测速度 (FPS), 以便直观地发现在准确率与检测速度上的最好权衡. 表 3 列出了本文方法及其他 10 种方法的检测速度及检测平均准确率, 其中包括 6 种全监督的方法 (第 2~7 行) 和 4 种弱监督方法 (第 8~11 行). 从表 3 可以看出, 本文的检测速度 (45 FPS) 与 SSD[3] 的检测速度 (46 FPS) 近似; 在准确率上, 由于 SSD 采用了有监督的训练模式, 因此本文的检测准确率要低于 SSD 方法. 本文方法与 30 HzDPM[24] 相比, 无论在检测平均准确率还是在检测速度上都表现出优秀的性能, 而该方法则认为当 FPS 达到 30 时可以满足实际中实时检测的需求[24], 因此可见本文在弱监督的情况下, 既提升了检测的准确率又达到了实时检测的目标.

表 3 PASCAL VOC 2007 测试数据集上的目标检测速度 (FPS) 及检测平均准确率

Table 3 Object detection speed (FPS) and detection mean average precision on PASCAL VOC 2007 test set

评价指标及数据	FPS	mAP	数据集
30 HzDPM[24]	30	26.1	07
Fast R-CNN[1]	0.5	70.0	07+12
Faster R-CNN[4]	7	73.2	07+12
YOLO_VGG[39]	21	66.4	07+12
Fast_YOLO[39]	155	52.7	07+12
SSD[3]	46	68.0	07
WSDDN_Ens[10]	0.5	39.3	07
WCNN[9]	–	42.8	07
MELM[12, 14]	–	47.3	07
PCL[13]	1.4	48.8	07
本文方法	**45**	**39.3**	07+12

表 3 中列出了另外 4 种弱监督目标检测方法 (第 8~11 行). WSDDN_Ens[10] 基于 Fast R-CNN[1] 方法提出了双数据流网络结构, 其检测速度与 Fast R-CNN检测速度近似 (0.5 FPS). WCNN[9] 与 MELM[12, 14] 在检测之前都需要为每幅图像生成近 2 000 个目标候选集, 这就大幅度地增加了检测时间, 导致无法实现实时检测, 因此这两种方法没有列出 FPS 值. 此外, 文献 [13] 中列出 PCL 方法在不考虑目标候选集获取时间的情况下, 其每幅图像的检测时间为 0.71 s (采用 NVIDIA GTX TitanX 显卡), 那么所对应的 FPS 值为 1.4. 本文使用 PCL[13] 所提供的源码, 使用 NVIDIA GeForce GTX 1 080 Ti 显卡时, 其每幅图像的检测时间为 0.81 s, 所对应的 FPS 为 1.2. 因此可以说明本文实验采用的 NVIDIA GeForce GTX 1 080 Ti显卡的性能稍稍低于 GTX TitanX 显卡. 表 3 中列出 PCL[13]使用 GTX TitanX 显卡时的 FPS 值. 即便是只考虑时间的情况下, 本文方法也要比 WSDDN_Ens[10]和 PCL[13] 花费更少的检测时间, 其 FPS 值要远远高于 WSDDN_Ens 与 PCL. 由此可以判定, 在弱监督设定下, 只有本文方法实现了实时检测的目标, 并且取得了较好的检测准确率.

3.3.3 阈值分析

本节通过变换二值化类别显著图时所设定的阈值来观测该阈值是否影响检测的准确率. 对于所设定的不同阈值, 该实验均采用相同的实验设置 (均使用 PASCAL VOC 2007 训练集来优化模型参数). 表 4 中列出了阈值取集合 $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ 中不同值和表 1 阈值组合在目标检测平均准确率 (mAP)上的对比结果. 图 5 给出了各个类别在不同阈值下的检测准确率直方图. 从表 4 可以看出, 当 20 个类别采用相同的阈值 0.5 时, 其 mAP 为 30.3 %, 而当 20 个类别采用表 1 所列阈值时, 其 mAP 获得提升 (33.9 %). 此外, 通过图 5 可以看出, 对于类别 “Bike” 、“Bottle” 、“Car” 、“Plant” 、“Train” 、“TV” , 阈值越大其检测准确率越高, 而对于类别 “Table” 、“Dog” 则阈值越小检测性能越好.

3.3.4 模型有效性验证

为了进一步验证本文所提方法的有效性, 本实验首先利用文献 [23] 的方法产生测试集图像中目标的显著图, 再采用第 2.3 节所描述的伪标注生成过程获取目标的检测结果. 该过程并没有训练任何目标检测网络. 本实验的检测平均准确率 (mAP) 在表 5 中列出. 通过表 5 可以看出, 在任何阈值设定下, 本文方法都要比文献 [23] 方法的检测准确率要高, 其准确率最高提升了 15.4 个百分点 (表 5, 第 7 列). 由此可以说明, 单纯使用文献 [23]的方法是无法获取较好的检测性能的. 为了进一步提升检测的准确率, 我们受文献 [23] 方法的启发, 通过网络的反向传递过程生成目标的显著图和伪标注, 最后利用该伪标注来训练实时目标检测网络. 本文方法既充分地利用了文献 [23] 方法的思想来挖掘目标在图像中的具体位置, 又发挥了网络模型利用上下文信息及优秀的特征表示来进一步提升检测性能的优势.

3.4 实验总结分析

为了更直观地分析本文的检测结果, 图 6 和图 7 可视化地表示出一些 PASCAL VOC 2007 测试集

表 4　不同二值化阈值对于检测准确率 (%) 的影响

Table 4　The influence of different binarization thresholds on the detection precision (%)

阈值	0.3	0.4	0.5	0.6	0.7	表1
mAP	26.6	28.7	30.3	29.4	26.7	**33.9**
Plane	30	39.8	45.5	47.7	51.1	**54.5**
Bike	45.8	37.9	51.6	53	**54.2**	52.9
Bird	22.8	30.8	**32**	29.2	26.9	30.0
Boat	6.3	5.2	8.2	10.5	15.1	**15.2**
Bottle	2.5	2.8	4.1	5.3	5.9	**7.8**
Bus	42.1	44.3	48.6	**50**	49.7	47.3
Car	32.9	36.9	39.5	44.6	**44.7**	44.5
Cat	59	**63**	58.5	47.6.	23.8	62.5
Chair	1	1.8	1	**9.5**	1.2	9.4
Cow	18.6	16.1	**22.5**	16.5	15.6	17.6
Table	**40.4**	36.5	38.1	29.9	24	39.3
Dog	**51.8**	47.8	36.7	25.1	7.6	48.9
Horse	41.6	44.5	44.2	40	31.3	**45.7**
Motorbike	47.9	51	**55.1**	51.3	49	49.9
Person	9.6	12.4	15.3	12.5	9.2	**17.2**
Plant	11.1	11.6	11.3	10.7	**17.1**	15.7
Sheep	12.7	14.4	**20.7**	17.7	15.6	18.4
Sofa	25.2	29	27.9	**31.8**	26.6	29.3
Train	26	28.8	34.6	39.6	42.9	**43.7**
TV	5.2	10	10.7	15	22.2	**28.2**

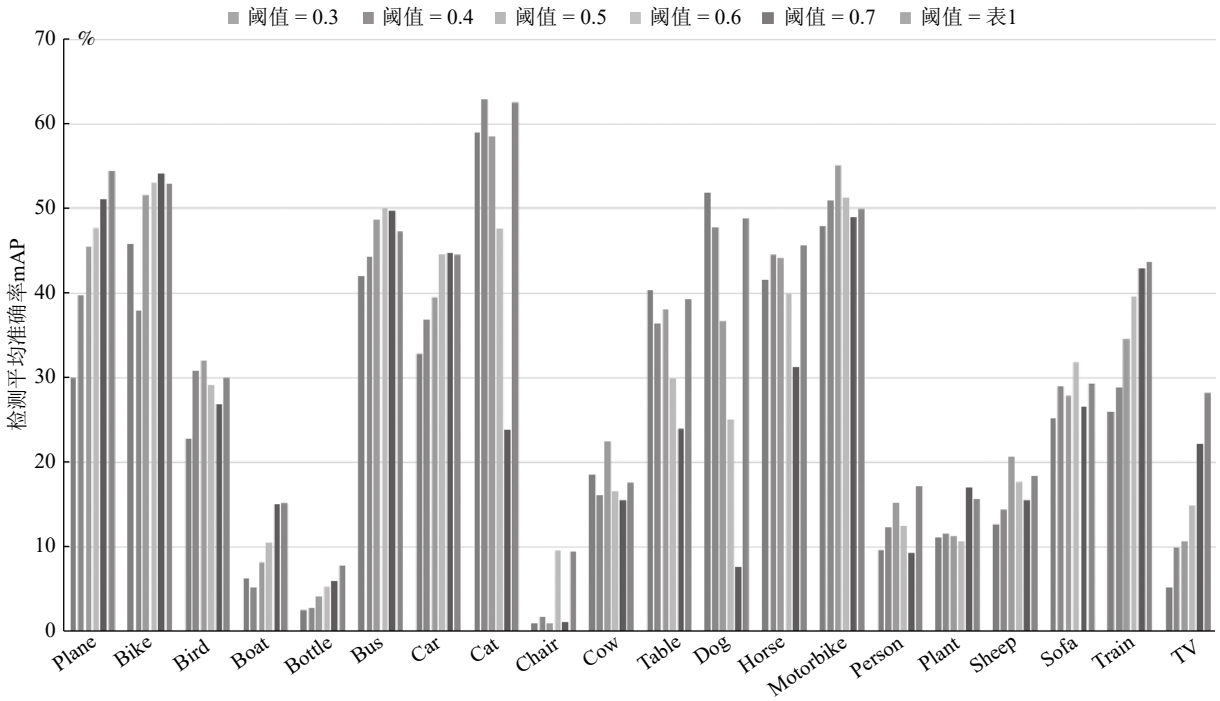


图 5　不同阈值设定下 20 个类别的目标检测准确率

Fig. 5　Object detection precision for 20 categories under different thresholds

表 5 在不同二值化阈值下, 通过文献 [23] 的方法生成的测试集检测结果 (mAP) 与本文方法的对比

Table 5 Under different binarization thresholds, the detection results (mAP) of test set generated by [23] and our method

阈值	文献 [23] (%)	本文方法 (%)	阈值	文献 [23] (%)	本文方法 (%)
0.3	16.7	26.6	0.6	16.1	29.4
0.4	17.5	28.7	0.7	13.7	26.7
0.5	17.5	30.3	表1	18.5	33.9

图 6 PASCAL VOC 2007 测试集上的成功检测样例

Fig. 6 Successful detection examples on PASCAL VOC 2007 test set

图 7 PASCAL VOC 2007 测试集上的失败检测样例

Fig. 7 Unsuccessful detection examples on PASCAL VOC 2007 test set

上的检测结果.图 6 和图 7 中, 检测的目标由不同颜色的矩形框表示出来并对应不同的类别, 图 6 列出了一些检测成功的案例. 通过对这些成功案例的分析, 我们发现如果目标具有以下特性: 1) 图像中包含不同类别的目标; 2) 对于同类别的目标, 各个目标之间彼此不相邻; 3) 要检测的目标对象占据图像的大部分区域. 则本文方法可以成功地检测到目标:

图 7 给出了检测失败的样例, 这些样例均具有以下特性: 1) 图像中存在多个同类别的目标, 并且彼此近邻; 2) 图像中的目标与背景对比度较低.

4 结束语

本文提出一种弱监督实时目标检测方法, 该方法首先利用分类网络的反向传递过程生成类别显著图, 再利用类别显著图生成目标的伪标注, 最后将伪标注作为真实标注训练实时检测网络. 通过实验分析表明, 该方法在 PASCAL VOC 数据集上比目前最先进的弱监督方法在检测准确率上获得了明显提升, 并且可以实现实时检测的目标.

我们的方法实现了弱监督条件下的实时目标检测任务, 模型简单易用, 具有一定的泛化能力. 从实验中我们发现, 伪标注的好坏将直接影响最后的检测性能, 因此如何提升伪标注的准确性是进一步提升实时检测性能的关键所在. 本方法提供了在弱监督条件下, 实现实时目标检测的一个可行的方案, 而且这个方案具有很大的提升空间. 因此本文为解决弱监督实时目标检测问题提供了新的研究思路.

我们下一步的研究计划是根据从失败的实验案例中获取的启发, 尝试构建更合理有效的伪标注, 并融合类别之间的关联性, 进一步优化弱监督目标检测网络结构, 提升检测的性能.

References

1 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Boston, USA: IEEE, 2015. 1440−1448

2 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 580−587

3 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. SSD: single shot multibox detector. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 21−37

4 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Procoeedings of the 2015 Advances in neural Information Processing Systems. Montréal, Canada: MIT Press, 2015. 91−99

5 Song H O, Girshick R, Jegelka S, Mairal J, Harchaoui Z, Darrell T. On learning to localize objects with minimal supervision. arXiv preprint.arXiv: 1403.1024, 2014.

6 Li Yong, Lin Xiao-Zhu, Jiang Meng-Ying. Facial expression recognition with cross-connect LeNet-5 network. *Acta Automatica Sinica*, 2018, **44**(1): 176−182
(李勇, 林小竹, 蒋梦莹. 基于跨连接 LeNet-5 网络的面部表情识别. 自动化学报, 2018, **44**(1): 176−182)

7 Cinbis R G, Verbeek J, Schmid C. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(1): 189−203

8 Shi M J, Ferrari V. Weakly supervised object localization using size estimates. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 105−121

9 Diba A, Sharma V, Pazandeh A, Pirsiavash H, Gool L V. Weakly supervised cascaded convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 914−922

10 Bilen H, Vedaldi A. Weakly supervised deep detection networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2846−2854

11 Bilen H, Pedersoli M, Tuytelaars T. Weakly supervised object detection with convex clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 1081−1089

12 Wan F, Wei P X, Jiao J B, Han Z J, Ye Q X. Min-entropy latent model for weakly supervised object detection. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1297−1306

13 Tang P, Wang X G, Bai S, Shen W, Bai X, Liu W Y, Yuille A L. PCL: proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. DOI: 10.1109/TPAMI.2018.2876304

14 Wan F, Wei P X, Jiao J B, Han Z J, Ye Q X. Min-entropy latent model for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. DOI: 10.1109/CVPR.2018.00141

15 Xi Xue-Feng, Zhou Guo-Dong. A survey on deep learning for natural language processing. *Acta Automatica Sinica*, 2016, **42**(10): 1445−1465
(奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, **42**(10): 1445−1465)

16 Chang Liang, Deng Xiao-Ming, Zhou Ming-Quan, Wu Zhong-Ke, Yuan Ye, Yang Shuo, Wang Hong-An. Convolutional neural networks in image understanding. *Acta Automatica Sinica*, 2016, **42**(9): 1300−1312
(常亮, 邓小明, 周明全, 武仲科, 袁野, 杨硕, 王宏安. 图像理解中的卷积神经网络. 自动化学报, 2016, **42**(9): 1300−1312)

17 Teh E W, Rochan M, Wang Y. Attention networks for weakly supervised object localization. In: Proceedings of the 2016 British Mahcine Vision Conference. York, UK: British Machine Vision Association, 2016.

18 Kantorov V, Oquab M, Cho M, Laptev I. Contextlocnet: con-

text-aware deep network models for weakly supervised localization. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 350−365

19 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2921−2929

20 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classifIcation models and saliency maps. arXiv Preprint. arXiv: 1312.6034, 2013.

21 Wei Y C, Feng J S, Liang X D, Cheng M M, Zhao Y, Yan S C. Object region mining with adversarial erasing: a simple classifIcation to semantic segmentation approach. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 1568−1576

22 Kolesnikov A, Lampert C H. Seed, expand and constrain: three principles for weakly-supervised image segmentation. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 695−711

23 Shimoda W, Yanai K. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 218−234

24 Sadeghi M A, Forsyth D. 30 Hz object detection with DPM V5. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 65−79

25 Dean T, Ruzon M A, Segal M, Shlens J, Vijayanarasimhan S, Yagnik J. Fast, accurate detection of 100, 000 object classes on a single machine. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013. 1814−1821

26 Van de Sande K E A, Uijlings J R R, Gevers T, Smeulders A W M. Segmentation as selective search for object recognition. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Colorado Springs, USA: IEEE, 2011. 1879−1886

27 Zitnick C L, Dollár P. Edge boxes: locating object proposals from edges. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 391−405

28 Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. *Artifficial Intelligence*, 1997, **89**(1−2): 31−71

29 Zhang D, Liu Y, Si L, Zhang J, Lawrence R D. Multiple instance learning on structured data. In: Proceedings of the 2011 Advances in Neural Information Processing Systems. Cranada, Spain: MIT Press, 2011. 145−153

30 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 3431−3440

31 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z H, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211−252

32 Wang C, Ren W Q, Huang K Q, Tan T N. Weakly supervised object localization with latent category learning. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 431−445

33 George P, Kokkinos I, Savalle P A. Untangling local and global deformations in deep convolutional networks for image classifIcation and sliding window detection. arXiv Preprint arXiv: 1412.0296, 2014.

34 Tang P, Wang X G, Bai X, Liu W Y. Multiple instance detection network with online instance classifier refinement. arXiv Preprint arXiv: 1701.00138, 2017.

35 Wu J J, Yu Y N, Huang C, Yu K. Deep multiple instance learning for image classifIcation and auto-annotation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 3460−3469

36 Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 1717−1724

37 Zhu W J, Liang S, Wei Y C, Sun J. Saliency optimization from robust background detection. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Rognition. Columbus, USA: IEEE, 2014. 2814−2821

38 Zhu L, Chen Y H, Yuille A, Freeman W. Latent hierarchical structural learning for object detection. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Rognition. San Francisco, USA, 2010. 1062−1069

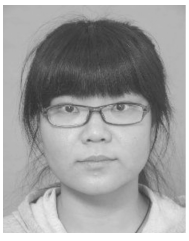
39 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unifled, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 779−788

40 Springenberg J T, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. arXiv Preprint arXiv: 1412.6806, 2014.

41 Cheng M M, Liu Y, Lin W Y, Zhang Z M, Posin P L, Torr P H S. BING: binarized normed gradients for objectness estimation at 300 fps. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 3286−3293

42 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv Preprint. arXiv: 1409.1556, 2014.

43 Yan J J, Lei Z, Wen L Y, Li S Z. The fastest deformable part model for object detection. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 2497−2504

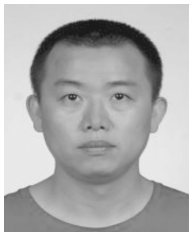


李 阳 哈尔滨工业大学计算机科学与技术学院博士研究生. 2013 年获得哈尔滨工业大学计算机科学与技术学院硕士学位. 主要研究方向为计算机视觉与机器学习.
E-mail: liyang13@hit.edu.cn
(**LI Yang** Ph.D. candidate at the School of Computer Science and Technology, Harbin Institute of Technology. She received her master degree from the School of Computer Science and Technology, Harbin Institute of Technology. Her research interest covers computer vision and machine learning.)

王 璞 2018 年获得哈尔滨工业大学计算机科学与技术学院硕士学位. 主要研究方向为计算机视觉与机器学习. E-mail: wangpu@hit.edu.cn
(**WANG Pu** received his master degree from the School of Computer Science and Technology, Harbin Institute of Technology in 2018. His research interest covers computer vision and machine learning.)

刘 扬 博士, 哈尔滨工业大学计算机科学与技术学院副教授. 主要研究方向为机器学习, 图像处理与计算机视觉. 本文通信作者.
E-mail: yliu76@hit.edu.cn
(**LIU Yang** Ph.D., associate professor at the School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers machine learning, image processing, and computer vision. Corresponding author of this paper.)

刘国军 博士, 哈尔滨工业大学计算机科学与技术学院副教授. 主要研究方向为计算机视觉, 图像处理与模式识别. E-mail: hitliu@hie.edu.cn
(**LIU Guo-Jun** Ph.D., associate professor at the School of Computer Science and Technology, Harbin Instititute of Technology. His research interest covers computer vision, image processing, and pattern recognition.)



王春宇 博士, 哈尔滨工业大学计算机科学与技术学院副教授. 主要研究方向为机器学习与计算生物学.
E-mail: chunyu@hit.edu.cn
(**WANG Chun-Yu** Ph.D., associate professor at the School of Computer Science and Technology, Harbin Instititute of Technology. His research interest covers machine learning and bioinformatics.)

刘晓燕 博士, 哈尔滨工业大学计算机科学与技术学院副教授. 主要研究方向为机器学习与计算生物学.
E-mail: liuxiaoyan@hit.edu.cn
(**LIU Xiao-Yan** Ph.D., associate professor at the School of Computer Science and Technology, Harbin Instititute of Technology. Her research interest covers machine learning and bioinformatics.)

郭茂祖 博士, 北京建筑大学电气与信息工程学院教授. 主要研究方向为机器学习, 数据挖掘, 生物信息学与计算机视觉.
E-mail: guomaozu@bucea.edu.cn
(**GUO Mao-Zu** Ph.D., professor at the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture. His research interest covers machine learning, data mining, bioinformatics and computing vision.)