软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3180−3194 [doi: 10.13328/j.cnki.jos.006619]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于数据合成和度量学习的台标检测与识别*

张广朋[1], 张冬明[2], 张 菁[1], 王川宁[1], 王立冬[3], 邹学强[2]

[1](北京工业大学 信息学部, 北京 100124)

[2](国家计算机网络应急技术处理协调中心, 北京 100029)

[3](北京广播电视台, 北京 100022)

通信作者: 张冬明, E-mail: zhdm@cert.org.cn

摘 要: 台标是视频的重要语义信息, 其检测与识别面临类别多、结构复杂、区域小、信息量低、背景干扰大等难题. 为提高模型的泛化能力, 提出将台标图像叠加到背景图像中合成台标数据, 来构建训练数据集. 进一步, 提出两阶段可伸缩台标检测与识别 (scalable logo detection and recognition, SLDR) 方法, 其采用 batch-hard 度量学习方法快速训练匹配模型, 确定台标类别. SLDR 的检测与识别分离机制使得其可将检测目标扩展到未知类别. 实验结果表明, 合成数据可以有效提升模型的泛化能力和检测精度. 实验亦显示 SLDR 方法在不更新检测模型的情况下, 即可获得与端到端模型相当的精度.

关键词: 数据合成; 度量学习; 可伸缩; 台标检测和识别

中图法分类号: TP391

中文引用格式: 张广朋, 张冬明, 张菁, 王川宁, 王立冬, 邹学强. 基于数据合成和度量学习的台标检测与识别. 软件学报, 2022, 33(9): 3180–3194. http://www.jos.org.cn/1000-9825/6619.htm

英文引用格式: Zhang GP, Zhang DM, Zhang J, Wang CN, Wang LD, ZOu XQ. TV Logo Detection and Recognition Based on Data Synthesis and Metric Learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3180–3194 (in Chinese). http://www.jos.org.cn/1000-9825/6619.htm

TV Logo Detection and Recognition Based on Data Synthesis and Metric Learning

ZHANG Guang-Peng[1], ZHANG Dong-Ming[2], ZHANG Jing[1], WANG Chuan-Ning[1], WANG Li-Dong[3], ZOU Xue-Qiang[2]

[1](Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

[2](National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

[3](Beijing Radio & Television Station, Beijing 100022, China)

Abstract: A TV logo represents important semantic information of videos. However, its detection and recognition are faced with many problems, including varied categories, complex structures, limited areas, low information content, and severe background disturbance. To improve the generalization ability of the detection model, this study proposes synthesizing TV logo data to construct a training dataset by superimposing TV logo images on background images. Further, a two-stage scalable logo detection and recognition (SLDR) method is put forward, which uses the batch-hard metric learning method to rapidly train the matching model and determine the category of TV logos. In addition, the detection targets can be expanded to unknown categories due to the separation mechanism of detection and recognition in SLDR. The experimental results reveal that synthetic data can effectively improve the generalization ability and detection precision of models, and the SLDR method can achieve comparable precision with the end-to-end model without updating the detection model.

Key words: data synthesis; metric learning; scalable; TV logo detection and recognition

---

* 基金项目: 国家重点研发计划 (2018YFB080402); 国家自然科学基金 (61672495, 61971016); 北京市自然科学基金-市教委联合资助项目 (KZ201910005007)

本文由“融合媒体环境下的媒体内容分析与信息服务技术”专题特约编辑汪萌教授、张勇东教授、俞俊教授以及张伟高级工程师推荐.

收稿时间: 2021-06-23; 修改时间: 2021-08-15; 采用时间: 2022-01-14; jos 在线出版时间: 2022-02-22

融合媒体发展是国家战略. 融合媒体业务中的媒体数据来源和内容将更为复杂, 移动终端短视频、网络视频分享、视频直播、广播电视等都将逐渐融合, 形成统一的内容服务平台, 如何进行高效的分析、管理和推荐, 是一个极具挑战性的问题. 台标是特定广播电台或组织的重要标志, 广播电台通过台标声明其对视频内容的所有权, 台标能够体现视频来源, 并协助进行视频内容分析. 台标检测与识别技术可以为视频内容分析提供丰富的语义信息, 因此其得到了广泛的应用[1]. 如在广电行业, 台标检测与识别也可以为电视质量检测、内容管理提供深度技术支持.

视频台标类别丰富, 而且仍在不断增长, 成为台标检测识别面临的首要问题. 同时, 台标作为嵌入式标志, 有别于其他标志, 多采用镂空、半透明设计, 受背景影响较大, 受视频编辑等重编码手段影响, 标志所在区域还可能出现模糊或变形, 这进一步加大了台标检测的难度. 基于此, 本文提出了数据合成方法, 来提高训练数据的多样性, 在无需大量人工标注的情况下, 提高网络模型的性能.

此外, 基于端到端深度网络的标志检测属于闭集 (close-set) 方法, 即基于封闭的训练数据并采用端到端网络, 直接完成目标的定位和检测. 闭集方法虽然可获得较高的识别精度, 但在可迁移性方面主要存在以下缺点: (1) 其所有支持的检测类别都是已知的, 如果新增台标类别, 需要重新训练网络更新模型. (2) 模型性能高度依赖训练样本, 既要大量标注数据, 还要注意类平衡的问题.

借鉴人脸识别[2,3]、物体识别[4,5]和标志识别任务[6,7]的解决思路, 我们提出一种可伸缩台标检测及识别方法, 构建检测与识别分离的两阶段网络模型.

本文主要贡献总结如下:

(1) 针对台标类别较多、现有标注数据较少的问题, 我们提出一种台标数据合成方法. 根据台标外形特点, 将标准台标区域直接叠加到背景中. 该方法有效减少了台标数据的人工标注. 结果表明台标数据合成方法可以有效提高模型检测能力, 提供可伸缩台标检测与识别模型训练所需的大量数据, 有效提升模型泛化能力和鲁棒性.

(2) 台标多采用镂空、半透明设计, 识别效果受背景影响较大. 本文在台标区域定位阶段引入实例分割分支获取台标掩码 (mask), 依赖台标掩码对图像进行预处理, 削减背景对台标类别匹配的影响, 提高台标识别精度.

(3) 针对闭集方法新增台标类型需对模型重新训练的缺点, 进一步提出一种可伸缩台标检测与识别方法, 适应多种台标类型, 提高台标检测与识别系统部署的灵活性. 其中针对台标的深度度量学习方法, 采用三元组损失训练匹配网络, 平衡批次样本, 使用 batch-hard 样本计算三元组损失, 快速收敛模型, 结合欧氏距离度量方法, 有效提高针对台标的识别能力.

本文第 1 节介绍台标检测与识别的相关工作. 第 2 节介绍台标数据合成工作. 第 3 节介绍可伸缩台标识别网络 SLDR. 第 4 节为实验设置与结果分析. 最后为本文结论.

1 相关工作

台标检测与识别是对视频画面中的台标进行定位和分类. 所采用的方法可分为基于传统手工特征的分类器方法和使用 YOLO[8]、SSD[9]、RetinaNet[10]等网络的深度学习算法. He 等人[11]使用基于 OTSU 的图像自动分割方法进行标志区域定位, 然后提取标志图像的 SURF 特征并用 k-means 聚类算法构建视觉词典, 最后使用支持向量机完成标志识别. 随着深度学习方法的发展, 徐佳宇等人[1]关注了台标镂空、半透明的特点, 提出一种逐图像的像素级半自动标注方法获得二值标签台标图像集, 并提出一个基于端到端全卷积网络的像素级台标识别网络 PNET, 在数据集上像素级分割的准确率高达 98.3%. 但这种方法采用端到端网络, 当新增台标类别时需要重新训练模型, 并不符合台标类别需经常变更的应用场景. 为满足台标类别变更的快速响应需求, 两阶段方法成为一种可行的技术方案.

顾名思义, 两阶段方法, 包括检测和识别两个阶段. 第 1 阶段实现目标的定位, 第 2 阶段进行匹配确定目标类别, 当新增识别类别时不需要重新训练模型, 是一种可以更灵活应对实际应用问题的开放数据集方法. 其实, 两阶段深度网络并不少见, 比如 R-CNN[12]、Fast R-CNN[13]、Faster R-CNN[14]就是典型的两阶段目标检测算法. 目前广泛用于考勤、海关等领域的人脸识别应用也是典型的两阶段方法. Taigman 等人[15]采用基于检测点的方法获取人

脸区域, 根据人脸基本点进行人脸对齐, 最后使用深度度量学习方法进行人脸的匹配. Schroff 等人[2]采用主流的深度神经网络来提取人脸区域特征, 使用基于 triplet 损失函数的深度度量学习方法进行人脸的识别, 有效提高了识别准确率. 也有很多学者在场景标志检测与识别任务中使用两阶段方法. Tüzkö 等人[6]提出两阶段的标志检测方法, 首先使用 Faster R-CNN 网络进行标志的定位, 然后用预训练后的模型提取标志特征, 最后通过检索确定标志的类别. Bastan 等人[7]提出了一个开放标志检测 (open-set logo detection, OSLD) 系统, 使用 RFBNet 定位标志区域, 提出 SDML 方法进行标志区域的匹配得到标志类别. Fehérvári 等人[16]在进行标志区域定位后, 使用了基于 Proxy-NCA 嵌入的度量学习方法来学习更好的嵌入, 以匹配候选标志区域和标准标志图像. Ayan 等人[17]使用 one-shot 学习技术实现标志的检测, 使用深度网络提取标准标志图像的特征表示, 将查询图像输入编解码网络, 并通过标准标志图像的特征表示加强编解码网络在分割图相应部分的响应, 获取目标标志的分割图.

与以上这些两阶段方法相比, 台标检测与识别有自身的难点和特点. 1) 与人脸具有相似“眼鼻嘴”分布模型相比, 台标类别较多, 组成元素也复杂多样, 往往含有矢量图、字符或图片等多种元素. 2) 台标可采用镂空或半透明设计, 其识别效果易受背景影响.

本文设计针对台标的可扩展检测与识别方案, 有效应对台标检测与识别存在的问题. 目前也有学者进行可泛化的目标检测与识别研究. Tian 等人[18]实现了无 anchor、无 proposal 的单阶段目标检测网络 FCOS, 避免了与 anchor 有关且对最终检测结果非常敏感的所有超参数. Yu 等人[19]则对视觉定位问题中候选框生成的问题进行深入探索, 提出一种兼顾多样性和鉴别性的候选框生成模型 DDPN, 以生成高质量的 proposal. 本文使用台标图像合成技术降低标注成本, 保证数据的多样性和鉴别性, 以提高深度模型泛化能力. 在开集方法中的台标区域定位阶段引入分割分支, 获取标准台标区域, 削弱台标镂空、半透明设计导致的背景干扰, 提高后续识别效果, 实现灵活高效的台标可扩展检测与识别.

2 台标数据合成方法

2.1 台标数据分析

台标作为电视台、组织的特定标志, 每类都具有独特的风格, 往往在较小区域内采用了大量规则图形、线段等元素, 同时例如 CCTV 系列台标、卫视台标及与其对应的二级电视台台标都较难区分, 涉及了细微的目标识别任务. 台标多采用镂空、半透明设计, 在图形、线段之间有大量的间隙, 台标图像如图 1 所示, 绿色框内为台标区域, 台标的镂空设计风格导致背景对台标检测的干扰较大. 已有学者针对台标的镂空性开展研究. Zhang 等人[20]依据先验知识对台标区域粗定位, 设计针对镂空台标的手工特征提取方法, 使用模板匹配方法完成对台标的检测, 但是此种方法难以适用目前多样、海量的台标图像.

图 1 台标图像

深度目标检测方法为解决台标检测中的镂空、半透明问题提供了新思路. 但目前, 该领域缺少公认的台标数据集, 现有文献中提及的数据集数据量少、缺少有效标注.

针对台标检测与识别任务, 本文构建了两个数据集, 其一为真实标注的数据集, 其二为合成数据集.

首先介绍真实标注的数据集. 为保证台标数据集的数量和种类, 对收集到的视频样本库采用 Sony Vegas Pro 进行解码, 并按照设定的抽帧策略从包含同种台标的不同视频中抽取代表帧. 所抽取的代表帧内容可能非常相似, 为保证模型的泛化能力, 通过相似度检测删除近似代表帧, 并侧重于选择背景丰富的代表帧. 为了减少画面清晰度、背景等因素给网络训练带来的干扰, 根据统计得到台标在画面中的位置规律, 只截取代表帧中相应台标区域. 图 1 展示了所获取的 4 个真实台标图像, 可以看出它们大多采用了镂空设计, 且 CCTV 综合、北京卫视存在半透明效

果, 台标区域受背景影响较大.

在获取真实的台标数据后, 为了给深度学习网络提供足够的监督信息, 需要为台标图片提供像素级的标注. 为了减少标注工作量, 我们依据同类台标出现的位置、大小, 将同类的台标分为不同组. 如图 2 所示, A 组、B 组、C 组分别展示了台标出现位置不同的同类台标图像. 每个分组仅需要标注其中一张图像即可完成本组的掩码标注, 这有效降低了标注的工作量. 最终, 构建了一个像素级标注的台标数据集, 共计 82372 幅图像, 包括 17 类.

图 2 真实台标数据标注组

2.2 台标图像合成

第 2 个数据集为合成台标数据集. 很多学者在标志检测与识别研究中涉及了数据合成研究[21−24]. 台标类别丰富, 而且往往会受到画面内容的影响, 所在区域变化多样, 逐一收集数据并进行手工标注需要大量人工成本. 合成数据可以自动生成大量数据, 并进行自动标注. 本文提出一种台标数据合成算法, 利用台标掩码信息和标准台标信息, 可与各种背景数据进行合成, 生成包含台标的模拟数据. 本文搜集了 1 198 类标准台标图像, 以及大量的背景图像, 通过台标数据合成算法共生成了约 112246 张图像. 接下来介绍合成技术的具体细节.

合成数据过程主要包括 4 个部分, 即标准台标的获取、台标掩码获取、背景图像选取以及台标与背景合成.

我们搜集了 1198 张标准台标图像, 图 3 列举了其中几个. 其中, 第 1 行是搜集的标准台标图像, 第 2 行是台标的掩码图像, 第 3 行是合成的台标图像.

图 3 台标图像合成示例

为了获得台标掩码, 首先采用 SLIC 方法处理台标图像, 自动获得超像素分割结果图, 作为粗糙的标签图像. 接着对较为粗糙的图像进行人工矫正, 背景区域像素值设置为 0, 以此获得最终的像素级标注. 我们共完成了 1 198 类台标的掩码生成.

电视台播放的视频背景丰富多样, 我们搜集了 20 900 张背景图像, 结合 FlickrLogos-32[25]的 no-logo 类图像, 共计 26 900 张图像作为背景图像源. 所选取的背景图像囊括了大范围的背景数据, 能够有效提升数据的多样性.

台标合成即利用标准台标的像素级标注将标准台标叠加到背景图像中, 叠加过程包括如下步骤:

1) 按设计的缩放比例, 对台标图像 A 和掩码图像 M 进行缩放, 获得缩放后的台标图像 As 和 Ms;

2) 按设计的叠加位置, 获取叠加图像 P, P 与 A_s 和 M_s 为同等大小矩阵;

3) 按照下述公式修改图像 P 的像素取值, 完成合成, 其中 $P(i,j)$、$A_s(i,j)$、$M_s(i,j)$ 代表相应图像在 (i,j) 位置处的像素值. 使用变量 α 是为了营造台标区域的半透明效果, 为拟合真实台标数据, 经过实验, α 取值范围为 (0.7, 1), 当 α 为 1 时没有半透明效果.

$$P(i,j)=\begin{cases} P(i,j) & M_s(i,j)=0 \\ \alpha\times A_s(i,j)+(1-\alpha)\times P(i,j) & M_s(i,j)\neq 0 \end{cases} \tag{1}$$

我们为每类台标合成 100 张图像. 传统图像数据增强方法已经被证明能够有效提升检测模型的鲁棒性, 我们对标准台标进行了预处理操作, 同时考虑同类台标颜色、形状基本无变化的特殊性, 在合成时只对台标进行了放缩处理, 为了保证台标数据的有效性, 对放缩处理进行了限定, 保证每张台标都能够完整出现在背景图像中, 同时对于台标缩小时分辨率小于 10×10 的台标不予叠加. 考虑到电视台视频中台标出现的位置随机性, 在叠加台标时随机选择位置, 同时, 除本类标准台标外, 对每次选择的背景图像随机叠加 1–2 类其他台标. 在台标图像合成后, 算法会自动生成合成图像的像素级标注.

合成数据实例如图 3 第 3 行所示, 合成图像能够较好地拟合真实台标图像, 不再需要手工标注. 第 4.2 节的实验能够证明合成数据的有效性, 其能够辅助实现台标检测算法性能的提升.

3 可伸缩台标识别网络 SLDR

通过数据合成技术, 可以显著提高台标检测与识别网络模型的泛化能力. 但对于端到端模型来说, 一旦目标台标集合发生变化, 即需要识别新的台标种类时, 需要更新模型, 而模型更新一般需要离线增量学习甚至完全重新训练, 无法满足目标台标集合经常改变的应用需求.

本文使用可伸缩开集方法进行台标的检测与识别, 具体方案流程如图 4 所示. 首先使用通用台标检测器定位台标位置, 如果只使用目标框进行标识的检测, 台标的镂空、半透明特点将加大识别难度, 因此在台标区域检测阶段引入分割分支获取台标掩码; 在获得台标掩码后, 为削弱背景对识别的影响, 需要对候选台标区域进行预处理操作, 即结合台标掩码对台标区域进行背景填充, 将背景像素值统一设置为 0; 最后将预处理后的台标区域和所有要识别的台标类的规范图像集相匹配得到最终类别, 匹配网络训练使用深度度量学习方法, 将检测到的台标区域使用距离最近的台标类进行标记. 应用 SLDR 方法, 当需要识别一个新的类别时, 将其标准台标图像添加到台标数据库即可, 不需要每一次增加类别都进行数据收集、标注和训练工作, 可以节省模型更新的成本, 提高对目标集变化时的响应速度, 同时结合所提出的数据合成方法, 可以满足两阶段模型训练所需的大量数据, 以提高模型的泛化能力和识别精度.

3.1 台标检测网络

台标检测即对台标所在区域进行定位, 其使用边界框和掩码定位候选区域, 而不识别台标的具体类别. 我们期望能够快速、准确地获取台标的像素级区域, 考虑到之后匹配需要得到台标的具体类别, 在此处的检测分割模型选择实例分割模型. 目前主要的实例分割模型和目标检测模型类似, 主要分为单阶段和两阶段的模型. 两阶段的实例分割模型主要为 Mask R-CNN[26], 虽然其检测的鲁棒性较高, 但是两阶段方法所消耗的时间较长, 难以满足实际应用需求. 单阶段实例分割模型不需要提前生成候选区域, 模型预测速度较快. Lee 等人[27]在 FCOS 的基础上提出

了单阶段实例分割模型 Centermask, 实现了无 anchor、无 proposal 的解决方案, 能够实现快速的目标分割. Yu 等人[28]提出了单像素重建网络的单阶段实例分割框架 SPRNet, 通过将单像素重建分支引入现成的单阶段检测器来执行有效的实例分割, 也能达到精度和速度的平衡. Bolya 等人[29]提出了极轻型的实例分割模型 YOLACT, 通过线性组合原型掩码生成分支和掩码系数预测分支两个并行支路的输出来获得最终的实例掩码, 能够在保证精度的同时提升预测的速度, 因此本文选择 YOLACT 进行台标的检测与分割.

图 4　可伸缩开集台标检测与识别方案

YOLACT 检测网络结构如图 5 所示, 其主干网络延用 RetinaNet 的基础架构, 主要包含基础特征提取网络和 FPN 层, 并在 RetinaNet 基础上添加了原型生成分支和掩码系数预测支路以完成实例分割任务. 台标检测网络使用 ResNet-50 作为基础特征提取网络, 提取 ResNet-50 第 3、4、5 卷积层特征{C3、C4、C5}, 并通过 FPN 网络生成对应的{P3、P4、P5}, 对 P5 层进行下采样得到{P6、P7}, 最后在{P3、P4、P5、P6、P7}这 5 个特征图上进行目标预测. 网络使用上述多个尺度的特征图, 可以检测到不同尺度的目标, 同时, 更深层的特征图能生成更鲁棒的目标掩码, 能保证最终的分割结果质量更高. YOLACT 在 FPN 的 P3 层之后添加了原型生成分支, 其由多个上采样卷积层组成, 最终输出尺度为 138×138 的 32 个原型掩码. YOLACT 预测层含有 3 个支路, 分别生成各候选框的类别、位置和掩码的置信度. 在支路中生成的 anchor 总数为 a, c 为需要预测的类别数, 全部的类别置信度的维度为 $a\times c$, 位置偏移的维度为 $a\times 4$, 掩码置信度为 $a\times k$, k 取值为 32, 对应原型生成分支输出的原型掩码个数. 掩码置信度用于加权原型生成分支产生的 32 个原型掩码, 具体计算方法见公式 (2), 其中 M 为最终生成的实例掩码数据, P 为原型生成分支生成的原型掩码, C 为掩码置信度, σ 为非线性函数 Sigmoid.

$$M=\sigma(PC^{\mathrm{T}}) \tag{2}$$

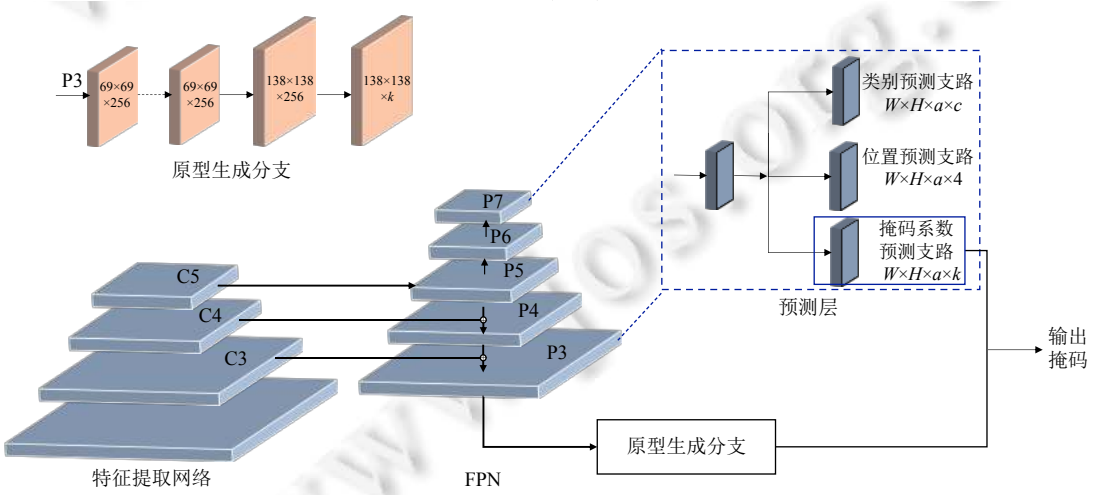


图 5　YOLACT 结构

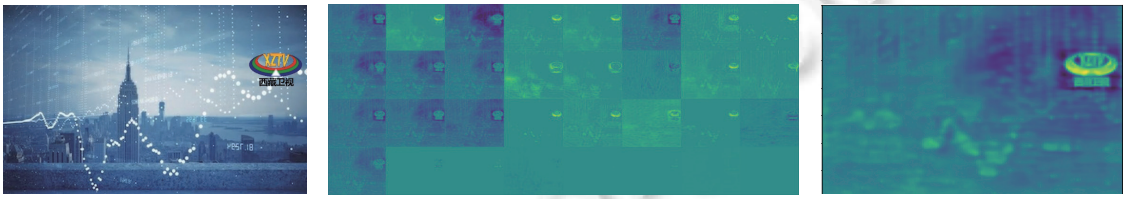
在模型训练过程中, 共使用 4 个 loss 函数, 分别是类别损失 L_{cls}、边界框回归损失 L_{box}、实例分割损失 L_{mask}、语义分割损失 L_s. L_{cls} 损失函数使用 Softmax 损失, L_{box} 损失使用 smooth L1 损失, L_{mask} 损失函数如下:

$$L_{\text{mask}} = BCE(M, M_{\text{gt}}) \tag{3}$$

其中, M 为预测的掩码, M_{gt} 为真实标注的掩码, BCE 代表像素级二进制交叉熵. 语义分割损失 L_s 为训练中额外的损失项, 可以在没有速度损失的情况下有效的提升特征的丰富性, 通过将含有 h 个输出通道的 1×1 卷积层直接附加到主干网络最大的特征图 (P3) 上, 使用交叉熵函数得到最终的 L_s 值, 总的损失函数如下:

$$L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}} + L_s \tag{4}$$

台标多镂空、半透明设计, 使用 YOLACT 实例分割模型能够对图像中的台标进行像素级分割, 削弱背景对台标识别的影响, 提升后续匹配的精度. 同时, 使用单阶段的实例分割模型, 可以更快速地完成台标的检测任务. YOLACT 实例分割模型在原型生成分支输出的 32 个原型掩码如图 6(b) 所示, 使用掩码置信度对获得的原型掩码进行加权之后生成的掩码如图 6(c) 所示, YOLACT 能够精确的进行台标的像素级分割.



(a) 待检测图像　(b) 原型掩码　(c) 掩码

图 6　台标分割示例

3.2　台标匹配度量学习

台标匹配即对提取的候选台标区域与目标台标进行匹配并判定其相似性, 该阶段使用深度度量学习方法. 图像深度度量学习的目的是学习一种低维的图像嵌入, 将相似的图像映射到嵌入空间的较近位置, 而将不同的图像映射到更远的位置, 以此区分不同类别的图像.

深度度量学习依赖于图像的正负对或三元组来优化损失函数, 本文选择使用三元组损失函数. 设 (a, n, p) 为输入的三元组数据, 其中 a 为训练数据集中随机选取的一个样本, a 和 p 属于同一类的样本, n 和 a 为不同类的样本, (x_a, x_n, x_p) 为特征提取网络生成的对应三元组的特征向量, 三元组损失函数 (triplet loss) 的公式如下所示:

$$L_{\text{triplet}} = \max(d(x_a, x_p) - d(x_a, x_n) + margin, 0) \tag{5}$$

通过 triplet loss 的学习后使得 x_a 和 x_p 之间的距离最小, 而 x_a 和 x_n 之间距离最大. 根据实验分析, 公式中的 $margin$ 取值为 1 最佳.

匹配网络训练时使用三元组损失函数, 基础特征提取网络使用 ResNet-50. 在训练深度度量网络时, 每个批次的训练样本选择是非常重要的[30]. 训练时三元组样本生成主要有离线方法和在线方法. 离线方法为随机生成三元组, 经过特定的迭代次数计算 triplet loss, 此方法计算效率不高, 需要较长时间训练, 而且模型收敛不稳定. 在线三元组生成即在训练时计算出当前批次数据之间的距离, 然后选择使用 semi-hard、hard 或 batch-hard 等三元组样本更新模型, 其主要原理即选择较难判断的三元组样本, 可以缩短模型训练的时间, 加快模型的收敛. 我们选择在线三元组样本生成方式, 同时, 在抽取每个批次的训练样本时平衡样本选择, 在每个批次中选择 m 个类, 每个类抽取 s 个样本, 共提取出 $m \times s$ 个样本.

在线训练模式中, 每个批次的样本可以使用 semi-hard、hard 或 batch-hard 等样本. 对应公式 (5), semi-hard 样本选择方法如下所示:

$$d(x_a, x_p) < d(x_a, x_n) < d(x_a, x_p) + margin \tag{6}$$

其中, semi-hard 样本代表 x_a、x_n 的距离较近, 引入 $margin$ 做权衡, 即选择不太困难的样本进行迭代更新. Hard 样本选择方法如下所示:

$$d(x_a, x_n) < d(x_a, x_p) \tag{7}$$

Hard 样本即 x_a、x_p 的距离很远, 代表模型会分错的样本, 使用符合公式 (7) 的样本进行训练可以在一定程度上提高模型收敛的速度.

Semi-hard 样本方案和 hard 样本方案是选择批次样本中的所有 triplet, 对其中的 semi-hard 和 hard 三元组样本的损失取均值. Batch-hard 样本选择方法与前两者不同, 其是遍历批次样本, 选择批次中最难的样本进行模型参数更新, 倾向于学习学不会的问题, 使用这种困难样本挖掘策略能够有效促进模型收敛. Batch-hard 样本即对于每一个 x_a, 只取那些 x_a 与 x_p 的距离最大且 x_a 与 x_n 的距离最小的样本组成三元组, 并只使用选择出的三元组样本损失值进行反向传播. 相对而言, batch-hard 的样本选择方式可以使模型更快收敛, 而且能取得较好的识别效果, 因此在训练时采用 batch-hard 样本选择方法.

公式 (5) 中的 d 代表两向量之间的距离, 目前, 主要的距离度量方法有欧氏距离、欧氏距离的平方和余弦距离. 欧氏距离 d_E 描述的是两个向量终点之间的距离, 欧氏的距离的平方 d_{E}^2 在计算上较为方便, 余弦距离 $d_{\cos}$ 表述两个向量之间的夹角余弦值, 设有向量 $A=[A_1,A_2,\ldots,A_n]$、$B=[B_1,B_2,\ldots,B_n]$, 各项距离计算公式如下所示:

$$d_{\mathrm{E}}=\sqrt{\sum_{i=1}^{n}(A_i-B_i)^2} \tag{8}$$

$$d_{\mathrm{E}}^2=\sum_{i=1}^{n}(A_i-B_i)^2 \tag{9}$$

$$d_{\cos}=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{10}$$

在第 4.3 节, 对不同的距离度量方法进行了实验对比, 最终选择了欧氏距离作为相似性度量的方法.

4 实验结果与分析

4.1 实验设置

数据集: 为验证本文方法的有效性, 在真实数据集和合成数据集进行评估.

真实数据集含有 17 类 82 372 张图像, 随机选择其中 74 134 张作为训练集, 8 238 张作为测试集. 在训练匹配网络时, 根据真实标注框对台标图像进行裁剪, 真实数据集中每张图像包含一个台标.

合成数据集共含有 1 198 类 112 246 张图像, 使用其中 1 178 类 111 195 张图像作为训练集, 测试集为 37 类 1 051 张图像. 训练匹配网络时, 同样依据真实标注框对台标进行裁剪. 在测试集中包含有训练集中没有的 20 类图像, 此举是为了验证可伸缩开集方法检测的有效性, 完成可扩展的台标检测与识别.

实验环境: 实验中采用的机器配置为 NVIDIA GeForce RTX 2080 Ti GPU, 所有网络模型的训练和测试均在 PyTorch 框架下进行, 测试时使用 COCO 数据集的评价方法.

4.2 合成数据效果测试

为验证合成数据的效果, 本节做了两组实验, 即在闭集检测器和通用检测器进行分别测试. 闭集检测器只能检测对应训练数据的类别, 即训练集和测试集的类别需严格一致; 通用检测器为可伸缩台标识别中台标定位阶段所使用的检测器. 通用检测器使用 YOLACT 实例分割网络, 在选择闭集检测模型时, 选择使用 YOLACT 的基础架构网络 RetinaNet.

(1) 在闭集检测器的效果测试. 测试集选择真实数据测试集. 参考其他研究对合成数据测试的方法[21,24], 使用较少真实数据训练 RetinaNet 闭集检测器, 在训练时用到真实数据只随机抽取每类的 30 张图像, 并选择 3 种策略验证合成数据的效果: ① RealImg: 仅使用真实数据进行训练; ② SynImg: 仅使用合成数据 (每类 100 张) 进行训练; ③ SynImg+RealImg: 首先使用合成数据 (每类 100 张) 进行训练, 然后使用真实数据 (每类 30 张) 进行微调. 需要特别说明的是, 闭集检测器的训练集类别是和测试集一一对应的, 因为其只能检测已知类别. 模型训练时, 输入图像的分辨率选择 550×550, 输入图像的标注为台标区域的 Box, 使用 Adam 优化器进行训练, 初始学习率为 0.000 01.

如表 1 所示, 使用合成数据结合真实数据进行训练, 可以取得更高的检测精度. SynImg+RealImg 方法相对 RealImg 方法提高了 3.3% 的 AP (Box), 证明使用数据合成方法生成的台标数据可以有效优化网络参数, 提高对台标的检测效果, 同时降低人工标注成本, 只需要较少的真实标注数据就可以取得较好的效果.

表 1 使用合成数据的性能结果 (%)

方法	AP (Box)	AP_{50} (Box)	AP_{75} (Box)
RealImg	92.3	98.9	97.4
SynImg	31.2	56.4	31.0
SynImg+RealImg	95.6	99.4	97.5

(2) 在通用检测器的效果测试. 通用台标检测网络只检测台标候选区域, 我们将所用训练集和测试集的不同类别台标标注统一为台标类. 在检测网络中添加了分割分支, 因此需要模型拥有更好的泛化性能, 才能够有效针对丰富类别的台标数据. 为了展现联合合成数据训练的效果, 本节设计了两种训练方法, 其一仅使用真实数据训练集进行训练, 其二使用真实数据训练集和合成数据训练集进行联合训练, 并分别在真实数据测试集和合成数据测试集测试模型效果. 通用检测器使用 YOLACT, 输入图像的分辨率选择 550×550, 输入图像的标注为像素级标注, 训练的检测器只需要检测出台标候选区域, 使用 Adam 优化器进行训练, 初始学习率为 0.000 01. 训练的模型在测试集的结果如表 2 所示.

表 2 台标检测器在真实数据集和合成数据集的检测性能 (%)

测试集	训练集	AP (Box)	AP_{50} (Box)	AP_{75} (Box)	AP (Mask)	AP_{50} (Mask)	AP_{75} (Mask)
真实数据	真实	84.4	99.0	98.9	55.3	98.9	58.8
	真实+合成	87.1	99.0	98.0	56.0	97.8	63.1
合成数据	真实	44.8	77.1	48.7	20.5	63.5	4.4
	真实+合成	91.0	99.0	98.0	50.7	94.3	46.8

在真实数据的测试集上进行测试时, 联合合成数据训练的模型得到的 AP (Box) 提升了 2.7%, AP (Mask) 提升了 0.7%; 在合成数据的测试集进行测试时, 联合合成数据训练的模型得到的 AP (Box) 提升了 46.2%, 而 AP (Mask) 提升了 30.2%. 如表 2 所示, 只使用真实数据训练时效果较低, 其中合成数据测试集上的 AP_{75} (Mask) 仅为 4.4%, 而联合合成数据后达到了 46.8%, 这有力说明数据的多样性能有效提高台标检测模型的泛化能力. 在真实数据测试集上的 AP 提升不及在合成数据集上的 AP 提升, 这主要是我们希望在引入合成数据时能够有效提高模型的泛化能力, 而不是过度拟合目前的真实数据集, 因此我们在合成数据时并没有针对性的引入我们目前所拥有的真实数据集的特性. 在合成数据测试集中存在 20 类没有在训练集中出现的类别, 而所提联合合成数据进行训练的方法达到了 91.0% 的 AP (Box), 台标检测网络可以鲁棒的开展扩展性检测, 将检测目标泛化到未知类别.

台标检测网络中添加了实例分割分支, 图 7 展示了对应表 2 结果的部分台标分割样例.

(a) 输入 (b) 检测 A (c) 检测 B (a) 输入 (b) 检测 A (c) 检测 B

(I) 真实图像 (II) 合成图像

图 7 YOLACT 台标图像分割结果

图 7 中 (I) 部分为真实数据图像的检测结果, 其中检测 A 为仅使用真实数据训练模型进行检测的结果, 检测 B 为使用真实数据和合成数据共同训练的模型进行检测的结果. 类似的, (II) 部分为合成图像的检测结果. 从图中可以看出, 联合合成数据进行训练的模型对真实数据和合成数据的检测效果均有改善, 合成数据的优势得以体现, 使用较小的合成成本就可以让模型拥有更强的泛化能力. (II) 部分中云南卫视和四川卫视两类图像并未用于训练, 其检测结果表明, 本文方法可以鲁棒地实现可伸缩检测, 将检测目标扩展到未知类别. 所提方法可以有效地实现台标检测, 提取高质量的候选台标区域.

4.3 台标匹配结果

本节评估不同距离度量方法、不同样本选择方案对台标匹配网络效果的影响. 训练时, 联合使用了合成数据和真实数据, 设置图像输入的大小为 224×224, 在进行相似性度量时生成的特征向量为 512 维, 输入图像源为对真实数据和合成数据根据真实框标注裁剪的目标区域, 使用 Adam 优化器进行训练, 初始学习率为 0.001. 测试时随机抽取每类的 2 张图像作为目标图像, 其余图像为待匹配图像. 匹配效果使用精度 *Precision* 和召回率 *Recall* 评估, 将 d_E 和预设阈值 T 进行比较, 得到相应的匹配状态:

$$status=\begin{cases} S_{TP}, & d_E \leqslant T \text{且匹配正确} \\ S_{FP}, & d_E \leqslant T \text{但匹配错误} \\ S_{FN}, & d_E > T \text{但匹配正确} \\ S_{TN}, & d_E > T \text{且匹配错误} \end{cases} \tag{11}$$

以此得到各种状态的图像数量 *TP*、*FP*、*FN*、*TN*, 匹配精度计算方式如下:

$$Precision=\frac{TP}{TP+FP} \tag{12}$$

召回率计算方式如下:

$$Recall=\frac{TP}{TP+FN} \tag{13}$$

在本节的测试中, 每张待匹配图像都有对应的目标图像, 我们在每次测试的 *Recall* 为 1 时记录 *Precision*.

表 3 中首先评估了不同的距离度量方法对结果的影响, 我们分别评估了欧氏距离、欧氏距离的平方、余弦距离的效果, 在训练模型时选择 batch-hard 样本, 可以得出使用欧氏距离度量方法的效果最佳.

表 3 不同距离度量方法的匹配精度 (%)

距离度量方法	*Precision* (真实数据)	*Precision* (合成数据)
欧氏距离	98.16	99.03
欧氏距离的平方	43.61	58.23
余弦距离	80.91	87.93

表 4 中展示了在训练匹配网络时不同样本选择方案对结果的影响, 在测试中距离度量方法使用欧氏距离, 分别在真实数据和合成数据的测试集上进行了评估. 特别地, 针对镂空、半透明台标识别问题, 我们设计了专门的预处理操作, 即利用掩码标注信息将背景颜色统一置为黑色. 预处理步骤如图 8 所示, (I) 部分为真实图像的预处理结果, 其中 (a) 为在真实图像中根据标注框裁剪出的台标区域, (b) 为 (a) 的掩码标注, (c) 为预处理结果, 类似的, (II) 部分为合成图像的预处理结果, 预处理之后可以降低甚至消除背景对识别的干扰. 表 4 中“*”代表对输入数据进行了预处理, 从表中可以得出, batch-hard 样本能够更好的收敛模型, 使用预处理操作在真实数据测试集上精度可以提升 1.29%, 在合成数据集上精度提升了 0.48%, 预处理操作可以降低背景对台标识别的影响, 提高台标识别的精度.

4.4 SLDR 和闭集方法结果对比

本节分别实现了开集方法 SLDR 和闭集方法进行台标检测与识别任务. SLDR 使用实例分割网络定位台标区域, 使用掩码信息对台标区域预处理去除杂乱背景的干扰, 然后匹配得到台标的具体类别. 为了对比体现 YOLACT

在台标数据集进行实例分割的优势, 我们同时使用 Mask R-CNN 实现了 SLDR 来进行效果的对比. 经过实验测试, 在匹配时阈值 T 取值为 10, 可以最大范围的滤除检测错误的台标区域. 鉴于 YOLACT 是对 RetinaNet 网络的改进, 本节实现的闭集方法采用的网络主要为 RetinaNet, 我们还使用了 YOLOv4[31]和 Faster R-CNN 来进一步对比展示检测与识别效果. YOLACT、Mask R-CNN、闭集模型的输入图像分辨率、优化函数等配置均相同, 输入图像的分辨率选择 550×550, 使用 Adam 优化器进行训练, 初始学习率为 0.000 01. SLDR 为开集方法, 使用了第 4.1 节中所述的大批量训练数据集进行训练; 闭集方法只能检测训练的已知类别, 实验中对合成数据和真实数据进行了分别训练, 训练类别对应测试集类别, 真实数据为 17 类 74 134 张图像, 合成数据为 37 类 3 513 张图像. SLDR 和闭集方法使用了一致的真实数据测试集和合成数据测试集.

表 4　不同样本选择方法的匹配精度 (%)

样本选择方法	*Precision* (真实数据)	*Precision* (合成数据)
Semi-hard	84.51	85.59
Hard	98.16	98.73
Batch-hard	98.16	99.03
Batch-hard*	99.45	99.51

(a) 台标区域　(b) 掩码标注　(c) 预处理结果　(a) 台标区域　(b) 掩码标注　(c) 预处理结果

(I) 真实图像　(II) 合成图像

图 8　台标图像预处理示例

在表 5 和表 6 中展示了闭集方法和 SLDR 方法在真实数据集和合成数据集的检测与识别结果, 此处的实验结果均是对得到的 Box 进行评价. 在闭集方法中, RetinaNet 检测的效果最好, 在真实数据集和合成数据集分别达到了 99.4% 和 92.5% 的 AP, 而且相较于 YOLOv4 和 Faster R-CNN, RetinaNet 拥有更快的检测速度. 在 SLDR 方法中可以选择使用预处理操作, “*”代表使用了预处理操作, 从表中可以看出, SLDR 在使用预处理之后得到的结果优于不使用预处理的. 同时预处理所需时间很小, 在真实数据中预处理时间可以忽略不计, 而在合成数据中基于 Mask R-CNN 和基于 YOLACT 的平均预测时间仅分别增加了 0.5 ms 和 0.4 ms. 基于 YOLACT 实现的 SLDR 比基于 Mask R-CNN 实现拥有更快的检测速度和更高的检测精度. 基于 YOLACT 的 SLDR 方法在真实数据集和合成数据集分别达到了 93.5% 和 93.6% 的 AP. 在真实数据测试集上, RetinaNet 达到的 AP 高于 SLDR 方法, 主要原因是真实数据集仅有 17 类, 而闭集方法只能检测已知类别, 在训练时也仅仅使用了对应测试集的类别, 但为了增强 SLDR 方法的可伸缩性, 使用了大量类别去优化 SLDR 的模型, 提高其泛化能力. 如表 6 所示, 在类别较多的合成数据测试集中, 基于 YOLACT 的 SLDR 方法取得的 AP 值高于所有的闭集方法.

同时, 为了说明 SLDR 方法的可扩展性, 在其训练数据中并不包含合成数据测试集的 20 类图像, 而从表 6 中可以看出, SLDR 依然取得了较好的效果. 这表明 SLDR 具有良好的可扩展能力, 即可准确检测未知类别台标区域. 此外, 我们注意到实验中真实数据相对于合成数据检测时间更短, 这是由于真实数据尺寸低于合成数据.

表 5 SLDR 和闭集方法在真实数据集的指标衡量结果

检测方法		平均每幅图像预测时间 (ms)	AP (%)	AP_{50} (%)	AP_{75} (%)
闭集方法	YOLOv4	23.5	84.5	99.1	97.9
	Faster R-CNN	42.2	88.6	99.7	99.4
	RetinaNet	23.3	99.4	99.8	99.3
SLDR	(Mask R-CNN)SLDR	77.3	86.3	95.4	95.0
	(Mask R-CNN)SLDR*	77.4	87.4	96.9	96.2
	(YOLACT)SLDR	24.3	91.9	96.6	96.1
	(YOLACT)SLDR*	24.3	93.5	98.5	97.9

表 6 SLDR 和闭集方法在合成数据集的指标衡量结果

检测方法		平均每幅图像预测时间 (ms)	AP (%)	AP_{50} (%)	AP_{75} (%)
闭集方法	YOLOv4	26.4	70.2	97.5	85.9
	Faster R-CNN	51.8	76.5	97.0	93.4
	RetinaNet	26.1	92.5	99.3	96.8
SLDR	(Mask R-CNN)SLDR	84.5	78.5	90.7	89.9
	(Mask R-CNN)SLDR*	85.0	79.7	92.4	91.3
	(YOLACT)SLDR	30.6	93.4	97.8	97.5
	(YOLACT)SLDR*	31.0	93.6	98.0	97.7

基于 YOLACT 实现的 SLDR 方法可以获得更高的精度以及更快的检测速度, 我们在图 9 中展示了对基于 YOLACT 实现的 SLDR 方法检测出的台标区域进行预处理的样例图像. (I) 部分为对真实图像的检测结果进行预处理的样例, 其中 (a) 为在真实图像中检测出的台标区域, (b) 为 (a) 的预测掩码, (c) 为预处理结果, 类似的, (II) 部分为对合成图像的检测结果进行预处理的样例. 需要说明的是, 在 (II) 部分中出现的台标类别均未在训练集中出现, 其依然取得了较好的分割结果. 为了更直观展示检测效果, 图 8 中为使用真实标注信息进行预处理的样例, 图 9 中为对台标图像进行检测并预处理的样例, 通过对比可以得出 SLDR 能够较为准确地进行台标的检测和台标区域的分割. 结合表 5 和表 6 中的数据, 通过将台标区域和分割信息结合进行预处理的操作能够提高后续的匹配精度, 增强整体台标检测和识别效果, 实现鲁棒的台标可扩展检测与识别.

(a) 台标区域 (b) 预测掩码 (c) 预处理结果 (a) 台标区域 (b) 预测掩码 (c) 预处理结果

(I) 真实图像 (II) 合成图像

图 9 台标图像检测结果预处理示例

表 5 和表 6 中, SLDR 检测方法使用的是同一个模型, 而闭集方法在更改测试集类别时均需要重新训练模型. 为验证 SLDR 在目标台标集合变化后的检测效果, 进一步说明其可伸缩性检测的优势, 我们专门合成了 300 类台标测试数据共 8513 张图片, 将表 5 和表 6 中所得结果使用的 SLDR 模型直接在 300 类的合成数据测试集上测试, 得到的结果如表 7 所示. 由于 YOLACT 实现的 SLDR 方法效果最好, 表 7 和表 8 中对应的 SLDR 方法均基于

YOLACT 实现. 同时 YOLACT 网络模型是对 RetinaNet 的改进, 并且在表 5 和表 6 展示的实验结果中, RetinaNet 能取得闭集中最好的效果, 因此对应表 7 和表 8 实验的闭集模型选择 RetinaNet. 表 7 中也列出了训练闭集模型需要的时间和模型测试结果, 训练闭集模型时使用了 19 806 张图片, 检测方法栏中的数字代表迭代次数, 闭集模型在训练数据上迭代 7 次后达到了和 SLDR 方法相似的性能. 从表中可以看出, 在 300 类台标数据上训练时, 闭集方法一次迭代需要 48 min, 而 SLDR 方法只需要使用大量数据集训练一次提高泛化能力, 在检测新的台标类别时一般不需要更新模型, 因此 SLDR 在检测灵活性方面有很大的优势, 其可伸缩开集的特点, 可以节省大量的模型更新代价, 特别是当类别逐渐增多时, 闭集模型训练需要更大的内存支持和更多的时间代价, 而 SLDR 方法在训练后可以拟合大量的台标类别, 实现高效灵活的可扩展检测.

表 7　SLDR 和闭集方法在 300 类合成数据集的指标衡量结果

检测方法	训练时间 (min)	AP (%)	AP_{50} (%)	AP_{75} (%)
闭集1	48	3.0	4.0	3.5
闭集4	192	58.3	66.3	65.4
闭集7	336	78.4	85.9	85.2
闭集10	480	83.6	89.8	89.0
闭集11	528	83.7	90.2	89.2
SLDR*	—	79.5	83.8	83.7

表 8　SLDR 和闭集方法在 600 类合成数据集的指标衡量结果

检测方法	训练时间 (min)	AP (%)	AP_{50} (%)	AP_{75} (%)
闭集11+	146	66.8	72.5	72.1
SLDR*	—	72.3	76.0	75.9
SLDR*+	69	72.3	75.9	75.8

目标集合在出现变化时, 可以对模型进行微调, 本节评估了两种检测方法在更新模型时需要的代价和达到的效果. 本实验测试集在 300 类合成数据的基础上新增 300 类图片, 新增 300 类图片的训练集包括 20 990 张图片, 测试集包括 5 032 张图片. 模型测试结果如表 8 所示, 表中“+”代表对模型进行了更新, 更新训练时对所有训练数据迭代一次.

微调闭集模型时, 时间相较于在 300 类数据上迭代一次增加了 98 min, 最终达到了 66.8% 的 AP. 初始 SLDR 模型是基于丰富类别的训练集进行训练的, 在检测时具有较好的泛化能力, 其在 600 类数据集上取得了 72.3% 的 AP. 在将初始 SLDR 模型在 600 类模型上进行更新时, 精度基本无变化, 其单次迭代训练时间相较于闭集模型减少了 77 分钟. SLDR 更新模型只需要更新定位网络参数, 台标度量网络学习一种低维的图像嵌入, 将相似的图像映射到嵌入空间的较近位置, 因此在台标集变化时不需要重新训练 SLDR 中的匹配模型. 根据表 7 和表 8 的结果显示, SLDR 可伸缩检测与识别的特点可以有效应对海量的台标数据, 进行灵活、高效的台标检测与识别.

5 结　语

针对台标检测与识别任务, 我们提出了数据合成方法来创建合成数据集, 自动生成标注样本, 有效应对台标类别多、标注数据少的问题. 进一步提出一种面向开集的两阶段网络 SLDR, 应用数据合成和度量学习方法, 实现可伸缩的台标检测与识别, 一阶段使用台标检测网络定位台标区域, 二阶段使用度量学习获得台标类别. 在大量合成数据训练的基础上, SLDR 拥有了强大的泛化能力, 可以在台标定位阶段精确获取台标区域及台标掩码. 针对台标背景对识别影响较大的问题, 使用定位阶段获取的台标掩码预处理台标图像, 可以有效提高检测的精度. SLDR 可伸缩的特点有效提高了台标检测的灵活性, 该方法同样可适用于其他目标检测任务.

为了进一步提高台标检测与识别的鲁棒性和灵活性, 可在以下几方面开展工作: (1) 在数据合成工作中, 可以

对台标模板模糊处理来进一步增加数据多样性; (2) 本文在台标检测网络中添加了分割分支, 但是分割分支训练在只有台标一类标注的数据集中没有取得预想的效果, 而预测框的效果表现较好, 可以考虑在预测出目标框之后再进行台标的分割, 从而更有效地去除背景的干扰, 提高台标识别的精度; (3) 本文着重研究了静态台标检测与识别, 动态的台标检测与识别也是亟需解决的问题, 可以引入时序关系辅助进行动态台标的检测; (4) 文字信息识别同样有助于台标识别精度提高, 可通过增加文字识别网络来进一步提升包含字符的台标识别精度.

References:

[1] Xu JY, Zhang DM, Jin GQ, Bao XG, Yuan QS, Zhang YD. PNET: Pixel-wise TV logo recognition network. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(10): 1878–1889 (in Chinese with English abstract). [doi: 10.3724/SP.J.1089.2018.16944]

[2] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 815–823. [doi: 10.1109/CVPR.2015.7298682]

[3] Wang M, Deng WH. Deep face recognition: A survey. Neurocomputing, 2021, 429: 215–244. [doi: 10.1016/j.neucom.2020.10.081]

[4] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv: 1703.07737, 2017.

[5] Wu D, Zheng SJ, Zhang XP, Yuan CA, Cheng F, Zhao Y, Lin YJ, Zhao ZQ, Jiang YL, Huang DS. Deep learning-based methods for person re-identification: A comprehensive review. Neurocomputing, 2019, 337: 354–371. [doi: 10.1016/j.neucom.2019.01.079]

[6] Tüzkö A, Herrmann C, Manger D, Beyerer J. Open set logo detection and retrieval. arXiv: 1710.10891, 2017.

[7] Bastan M, Wu HY, Cao T, Kota B, Tek M. Large scale open-set deep logo detection. arXiv: 1911.07440, 2019.

[8] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

[9] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot multibox detector. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]

[10] Lin TY, Goyal P, Girshick R, He KM, Dollár P. Focal loss for dense object detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2999–3007. [doi: 10.1109/ICCV.2017.324]

[11] He JM, Xie YX, Luan XD, Niu X, Zhang X. A TV logo detection and recognition method based on SURF feature and bag-of-words model. In: Proc. of the 2nd IEEE Int'l Conf. on Computer and Communications. Chengdu: IEEE, 2016. 370–374. [doi: 10.1109/compcomm.2016.7924725]

[12] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

[13] Girshick R. Fast R-CNN. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1440–1448. [doi: 10.1109/ICCV.2015.169]

[14] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/tpami.2016.2577031]

[15] Taigman Y, Yang M, Ranzato MA, Wolf L. DeepFace: Closing the gap to human-level performance in face verification. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1701–1708. [doi: 10.1109/CVPR.2014.220]

[16] Fehérvári I, Appalaraju S. Scalable logo recognition using proxies. In: Proc. of the 2019 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2019. 715–725. [doi: 10.1109/WACV.2019.00081]

[17] Bhunia AK, Bhunia AK, Ghose S, Das A, Roy PP, Pal U. A deep one-shot network for query-based logo retrieval. Pattern Recognition, 2019, 96: 106965. [doi: 10.1016/j.patcog.2019.106965]

[18] Tian Z, Shen CH, Chen H, He T. Fcos: Fully convolutional one-stage object detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 9626–9635. [doi: 10.1109/ICCV.2019.00972]

[19] Yu Z, Yu J, Xiang CC, Zhao Z, Tian Q, Tao DC. Rethinking diversified and discriminative proposal generation for visual grounding. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI, 2018. 1114–1120. [doi: 10.24963/ijcai.2018/155]

[20] Zhang L, Xia T, Zhang YD, Li JT. Hollow TV logo detection. In: Proc. of the 18th IEEE Int'l Conf. on Image Processing. Brussels: IEEE, 2011. 3581–3584. [doi: 10.1109/ICIP.2011.6116491]

[21] Su H, Zhu XT, Gong SG. Deep learning logo detection with data expansion by synthesising context. In: Proc. of the 2017 IEEE Winter Conf. on Applications of Computer Vision. Santa Rosa: IEEE, 2017. 530–539. [doi: 10.1109/WACV.2017.65]

[22] Su H, Zhu XT, Gong SG. Open logo detection challenge. In: Proc. of the British Machine Vision Conf. Newcastle: BMVA, 2018. 16.

[23] Montserrat DM, Lin Q, Allebach J, Delp EJ. Logo detection and recognition with synthetic images. Electronic Imaging, 2018, 30(10): 3371–3377. [doi: 10.2352/issn.2470-1173.2018.10.imawm-337]]

[24] Jiang YC, Gao C, Ji LX, Wu YC. Context-based synthetic data for logo recognition. In: Proc. of the 2019 Int'l Conf. on Artificial Intelligence and Advanced Manufacturing. Dublin: IEEE, 2019. 60–65. [doi: 10.1109/AIAM48774.2019.00019]

[25] Romberg S, Pueyo LG, Lienhart R, Van Zwol R. Scalable logo recognition in real-world images. In: Proc. of the 1st ACM Int'l Conf. on Multimedia Retrieval. Trento: ACM Press, 2011. 25. [doi: 10.1145/1991996.1992021]

[26] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

[27] Lee Y, Park J. CenterMask: Real-time anchor-free instance segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13903–13912. [doi: 10.1109/CVPR42600.2020.01392]

[28] Yu J, Yao JH, Zhang J, Yu Z, Tao DC. SPRNet: Single-pixel reconstruction for one-stage instance segmentation. IEEE Trans. on Cybernetics, 2021, 51(4): 1731–1742. [doi: 10.1109/TCYB.2020.2969046]

[29] Bolya D, Zhou C, Xiao FY, Lee YJ. YOLACT: Real-time instance segmentation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 9156–9165. [doi: 10.1109/ICCV.2019.00925]

[30] Wu CY, Manmatha R, Smola AJ, Krähenbühl P. Sampling matters in deep embedding learning. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2859–2867. [doi: 10.1109/ICCV.2017.309]

[31] Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv: 2004.10934, 2020.

附中文参考文献:

[1] 徐佳宇, 张冬明, 靳国庆, 包秀国, 袁庆升, 张勇东. PNET: 像素级台标识别网络. 计算机辅助设计与图形学学报, 2018, 30(10): 1878–1889. [doi: 10.3724/SP.J.1089.2018.16944]

张广朋(1997－), 男, 硕士, 主要研究领域为人工智能系统设计与集成.

王川宁(1997－), 男, 硕士, 主要研究领域为人工智能系统设计与应用.

张冬明(1977－), 男, 博士, 研究员, 博士生导师, CCF 专业会员, 主要研究领域为视频编码, 多媒体内容检索.

王立冬(1967－), 女, 教授级高级工程师, 主要研究领域广播电视工程技术, 视音频信号处理, 媒体网络.

张菁(1975－), 女, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多媒体内容分析与处理.

邹学强(1978－), 男, 博士, 高级工程师, 主要研究领域为网络安全.