软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007033]

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

注意力引导的标志检测与识别*

张冬明[1], 靳国庆[1], 鲁鼎煜[2], 张 菁[2], 张勇东[1,3]

[1](人民网 传播内容认知全国重点实验室, 北京 100733)

[2](北京工业大学 信息学部, 北京 100124)

[3](中国科学技术大学 信息科学技术学院, 安徽 合肥 230026)

通信作者: 靳国庆, E-mail: jinguoqing@people.cn

摘 要: 自然场景中的实体标志, 如商标、交通标志等, 易受拍摄角度、所依附物体形变、尺度变化等影响, 导致检测精度降低. 为此, 提出一种注意力引导的标志检测与识别网络 (attention guided logo detection and recognition network, AGLDN), 联合优化模型对多尺度变化和复杂形变的鲁棒性. 首先通过标志模板图像搜集及掩码生成、标志背景图像选取和标志图像生成创建标志合成数据集; 然后基于 RetinaNet 和 FPN 提取多尺度特征并形成高级语义特征映射; 最后利用注意力机制引导网络关注标志区域, 克服目标变形对特征鲁棒性的影响, 实现标志检测与识别. 实验结果表明, 所提方法可以有效降低尺度变化、非刚性形变的影响, 提高标志检测准确率.

关键词: 标志检测和识别; 数据合成; 多尺度特征融合; 注意力引导

中图法分类号: TP391

中文引用格式: 张冬明, 靳国庆, 鲁鼎煜, 张菁, 张勇东. 注意力引导的标志检测与识别. 软件学报. http://www.jos.org.cn/1000-9825/7033.htm

英文引用格式: Zhang DM, Jin GQ, Lu DY, Zhang J, Zhang YD. Attention Guided Logo Detection and Recognition. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7033.htm

Attention Guided Logo Detection and Recognition

ZHANG Dong-Ming[1], JIN Guo-Qing[1], LU Ding-Yu[2], ZHANG Jing[2], ZHANG Yong-Dong[1,3]

[1](State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China)

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

[3](School of Information Science Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: In natural scenes, logos such as trademarks and traffic signs are susceptible to shooting angle, carrier deformation, and scale changes, which reduces logo detection accuracy. Thus, this study proposes an attention guided logo detection and recognition network (AGLDN) to jointly optimize the model robustness for multi-scale and complex deformation. First, a logo synthesis dataset is established by image collection and mask generation of logo templates, image selection of logo background, and logo image generation. Then, based on RetinaNet and FPN, multi-scale features are extracted and high-level semantic feature mapping is formed. Finally, the attention mechanism guided network is employed to focus on the logo area, and the influence of logo deformation on feature robustness is suppressed to improve logo detection and recognition. Experimental results show that the proposed method can reduce the influence of scale changes and non-rigid deformation, and improve detection accuracy.

Key words: logo detection and recognition; data synthesis; multi-scale features fusion; attention guidance

标志 (logo) 作为一种信息传播的标识性视觉符号, 在产品或组织宣传上起到关键作用. 随着标志图片的快速增长, 标志检测与识别在广告效果评估、商标侵权分析、智能交通系统、仿冒商品检测等各个领域具有广泛的应

* 基金项目: 国家重点研发计划 (2021YFF0901600); 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)
收稿时间: 2022-11-21; 修改时间: 2023-03-16, 2023-04-25; 采用时间: 2023-08-18; jos 在线出版时间: 2023-12-06

用价值[1−4]. 标志类别丰富、表现形式多样, 如何利用智能化的分析技术实现标志检测与识别是亟待解决的实际工程应用课题.

标志更多为自然场景中包含的实体标志, 如商标、交通标志等, 而由于自然场景的视频或图片的摄制条件不同, 包括拍摄位置、角度、光线, 而导致标志呈现多变尺度、非刚性形变, 同时亮度、对比度也会存在复杂变化. 现有目标检测技术没有专门针对多尺度和非刚性形变等问题进行模型联合优化, 在标志面临复杂变化时难以获得满意的结果.

自然场景中标志存在尺度变化现象, 会造成提取的特征不对齐问题. 多尺度目标特征提取和融合是一种有效的解决方法, 它通过融合低层特征的描述性内容和高级特征的语义信息, 来提高尺度多变目标的特征鲁棒性[5]. RetinaNet[6]作为使用特征金字塔网络 (feature pyramid network, FPN)[7]的单阶段目标检测网络, 已经在目标检测领域取得了优异的效果. 另一方面, 标志出现在柔性载体或者非平面实体上容易存在非刚性形变. 深度卷积网络基于普通卷积产生的局部感受野进行特征提取时, 如果采样的特征包含了过多超出目标区域的内容, 将会导致特征受到图像中非目标内容的影响, 降低检测与识别的精度[8]. 针对提高网络对形变目标的建模能力, Zhu 等人[8]提出了可变形卷积, 来提高感受野中目标特征的权重, 也有很多网络通过引入可变形卷积来改善对形变目标的检测与识别能力. 例如 Liu 等人[9]提出一种门控特征融合模块, 通过注意力机制来有效融合可变形卷积特征和普通卷积特征, 综合提高在目标变形时网络的检测与识别能力. 考虑到标志虽然存在非刚性形变, 但是也有部分标志只含有尺度变化问题, 因此, 可以使用注意力机制来引导可变形卷积特征和普通卷积特征的融合, 从而有效提升感受野中目标特征的权重, 降低目标变形对检测性能的影响. 此外, 标志类别丰富且新类别不断涌现, 现有标志数据集无法覆盖所有标志类别, 因此, 标志检测与识别同样面临标注数据集不足的难题.

为此, 本文提出一种注意力引导的标志检测与识别网络, 通过联合模型优化, 提高其对尺度变化、非刚性形变的鲁棒性.

本文主要贡献总结如下.

(1) 针对标志类别丰富且新类别不断涌现, 导致数据集不能覆盖所有标志类别的问题, 本文提出一种标志数据合成方法, 合成符合真实标志图像内容分布的数据集, 提高模型泛化能力.

(2) 受拍摄位置、角度的影响, 标志容易出现尺度变化. 本文使用 RetinaNet 和 FPN 提取多尺度特征构建多个高级语义特征映射, 并通过融合低层特征的描述性内容和高级特征的语义信息提高尺度多变目标的特征鲁棒性.

(3) 受到拍摄角度、所依附载体变形的影响, 标志容易产生非刚性形变. 本文使用注意力引导机制来加权融合普通卷积特征和可变形卷积特征, 从而有效提升感受野中目标特征的权重, 降低目标变形对检测性能的影响. 并在统一网络内对多尺度和注意力机制联合优化, 实验表明所提方法显著提升了标志检测性能.

本文第 1 节介绍标志检测与识别的相关工作. 第 2 节介绍标志数据合成工作. 第 3 节介绍注意力引导的标志检测与识别网络 AGLDN. 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

标志检测与识别是对自然场景中的标志进行定位和分类, 其属于目标检测领域, 所采用的主要方法可分为基于手工特征的方法和基于深度学习的方法. 基于手工特征的视觉标志检测与识别主要包括建议区域提取、手工特征提取和建议区域分类[10], 其性能严重依赖手工设计的特征, 而在更换任务或者数据集变化时手工特征的鲁棒性难以保证. 此外, 这类方法通常使用传统分类器, 计算复杂度高且泛化能力难以保证, 限制了视觉标志检测与识别中速度和精度的提升.

基于深度学习的标志识别方法主要分为两阶段方法和单阶段方法[10]. 两阶段检测方法以 R-CNN (Region-CNN)[11]、Fast R-CNN[12]、Faster R-CNN[13]为代表, 在第 1 阶段进行目标建议区域的提取, 第 2 阶段进行目标的分类, 其大幅度提升了目标检测与识别的精度, 因此有研究者将两阶段方法应用在视觉标志的检测与识别中. Hoi 等人[14]分别使用了 R-CNN、Fast R-CNN、SPP-Net (spatial pyramid pooling in deep convolutional network)[15]进行视觉标志检测与识别, 并比较了相关精度. Eggert 等人[16]通过使用 Fast R-CNN 计算的特征表示直接对大目标进行识别,

同时以高概率预测包含小对象的显著区域并放大, 并计算这些放大区域的新特征表示, 在保持较低计算开销的同时提高了多尺度目标的检测质量. Li 等人[17]提出使用 Faster R-CNN 进行视觉标志的检测与识别, 使用迁移学习策略优化网络参数, 基于所使用的数据集使用 K 均值聚类获取合适的锚框尺度, 以此提高检测精度.

虽然两阶段方法提高了建议目标区域生成的质量, 同时获得较高的检测准确率, 但是整体检测速度较低, 无法满足实时性需求. 为了提高目标检测速度, 开始研究单阶段检测器, 代表性方法包括 YOLO[18]、SSD[19]和 RetinaNet[6]. 单阶段方法提高了检测与识别的灵活性, 但是整体检测精度相较于两阶段方法有一定差距, 因此有研究者提出利用特征图之间的上下文关系, 增强多尺度目标的特征鲁棒性, 以提高单阶段目标检测方法的性能. RetinaNet[6]通过结合基础特征提取网络和 FPN 提取多尺度特征, 并在不同尺度的特征图上设定不同尺度的锚框 (anchor), 使用大尺寸特征图预测较小的目标, 小尺寸特征图预测较大的目标, 有效提升了对不同尺度目标的检测与识别能力. Yang 等人[20]提出了一个 Inception-Text 模块, 通过不同大小的卷积来处理目标的多尺度, 并在每个分支的最后添加了一个可变形卷积 (deformable convolution)[21]处理多方向问题, 然后融合不同大小卷积分支提取的特征, 来提高多尺度多方向目标的检测性能. 虽然引入多尺度特征融合可以提高对多尺度目标的检测能力, 但进行特征融合时, 高级特征的表达能力将影响融合后的特征质量, 因此, 有研究者致力于提高多层次特征质量, 进而提升高级语义特征和底层特征的融合效果. Hou 等人[5]提出了一种多尺度特征解耦网络, 在网络中引入了一个平衡的特征金字塔模块 (balanced feature pyramid, BFP), 使用相同的深度特征映射来多层次特征, 在 FlickrLogos-32 数据集[22]上 mAP 达到了 86.2%. 但是使用 BFP 方法会显著增加特征融合的复杂度, Yang 等人[20]针对多方向多尺度的场景字符, 并借鉴了 Inception 网络[23]的构建思想提出了一个 Inception-Text 模块, 将 Inception-Text 添加在网络的高层, 在不过多提高计算复杂度的同时提高语义特征的质量, 在 ICDAR2015 数据集[24]上取得了 90.5% 的检测精度, 提高了多尺度多方向目标的检测性能.

在提高特征质量时, 有研究者在网络中引入注意力机制提高目标特征权重, 优化特征的表达能力, 进而提升目标检测器性能. 例如, Woo 等人[25]提出了卷积注意力机制模块 (convolutional block attention module, CBAM), 其首先使用通道注意力加权特征, 以增加有效通道的权重, 在通道注意力模块后引入空间注意力模块, 提高空间上有效特征的关注程度. 虽然 CBAM 被广泛应用于检测器的特征细化, 但是其通过级联通道注意力和空间注意力来优化特征, 而且使用了全连接层学习注意力权重, 导致参数量远远大于卷积结构, 严重影响了模型推理速度. 为了降低计算复杂度, Hu 等人[26]提出了 SE (squeeze-and-excitation) 模块, 采用了一种“特征重标定”策略自适应地重新校准通道的特征响应, 增强有效通道的权重, 提高特征质量. 而 SE 结构通过全连接层 (fully connected layer, FC) 降低特征维度以控制模型的复杂性, 虽然取得了一定的效果, 但是降维会影响特征的优化效果. Wang 等人[27]提出了 ECA (efficient channel attention) 模块, 在 SE 的基础上添加适当的通道交互, 使用卷积核大小为 k 的一维卷积来捕获跨通道的交互信息, 不仅大幅降低了参数量与复杂度, 而且能够有效提升感受野中目标特征权重.

除了上述基于通道和空间选择的方法之外, 也有研究者提高网络对变形目标的建模能力, 以此提升感受野中目标特征的权重. Dai 等人[21]提出了可变形卷积来提高对形变目标的建模能力, 其在普通卷积的操作上加入偏移量, 偏移量是通过另外一个平行的普通卷积单元计算得到, 并通过偏移量提高感受野中目标特征的权重. 可变形卷积在建模变形目标表现出的优异性能, 在很多网络中得到应用. Zhang 等人[28]引入可变形卷积提出了一种位置感知可变形卷积, 将普通卷积和可变形卷积输出的特征进行通道组合并使用 1×1 卷积输出最终的特征映射, 能够自适应提取非均匀分布的上下文特征, 提高复杂场景中目标特征的鲁棒性. 但是上述方法只是将普通卷积和可变形卷积提取的特征进行通道上的组合, 并没有关注普通卷积和可变形卷积所提取特征的有效性分布. Liu 等人[9]提出一种门控特征融合模块, 通过引入注意力门控机制来平衡普通卷积特征和可变形卷积特征, 当目标的变形被精确建模时, 注意力门控机制可以引导普通卷积特征和可变形卷积特征之间的融合, 以有效提升在目标变形、环境变化时的目标识别能力. 考虑到视觉标志虽然存在非刚性形变问题, 但是也可能有标志只含有尺度变化问题, 并无明显的形变, 因此, 引入这种注意力门控机制有可能提高模型自适应注意能力.

此外, 基于深度学习方法的目标检测与识别都需要大量的标注数据去训练模型参数, 以提高模型检测与识别的能力. 生成高质量标注数据集的主要方法是广泛收集后进行人工标注, 但是这不仅需要消耗大量的人力, 也难以

满足数据的多样性需求. 因此针对视觉标志检测与识别任务, 有研究者提出使用标志图像合成方法生成大量的数据集. 标志图像合成即将视觉标志模板图像经颜色抖动、仿射变换等预处理后叠加到背景图像中, 同时自动生成标注信息. 对视觉标志模板图像做大量预处理可以提高生成图像的多样性, 例如 Su 等人[29]对视觉标志模板进行缩放、旋转、颜色变换等预处理, 通过随机选取背景图像区域进行叠加来生成合成视觉标志图像; Jiang 等人[30]则考虑了视觉标志模板图像的选择, 其使用了透明背景的视觉标志模板和非透明背景的视觉标志模板, 同时也对视觉标志模板图像进行仿射变换、随机裁剪、颜色变换、高斯模糊等一系列预处理操作, 以此来提高样本的多样性. 而仅考虑对视觉标志模板图像进行预处理, 只能在一定范围内提高样本多样性, 因此有研究者提出对背景图像的考虑, Jiang 等人[30]考虑在现实场景中, 视觉标志会和周围场景有一定的上下文联系, 为了让样本尽可能拟合真实数据的分布, 其使用场景分类模型分类收集的图片, 并手工配置每类视觉标志可能出现的场景, 提高数据的真实程度; Song 等人[31]自动生成包含各种立体图形的背景图像, 提出了基于随机化的视觉标志数据合成, 将视觉标志图像粘贴到随机颜色、形状、纹理等特征的圆柱体、盒子和平面等形状上, 以此提高合成样本的多样性. 以上基于对视觉标志模板预处理和背景图像选取的方法没有关注视觉标志叠加的位置, 而其影响上下文关系的和谐性, 是判断合成图像是否拟合真实数据的主要依据. 为此, 有研究者针对合成方法开展研究, Su 等人[32]提出了情景对抗学习方法, 对初步合成的图像进行进一步的亮度、对比度等处理, 用来生成上下文关系更接近真实数据的合成图像. Gupta 等人[33]研究了场景字符图片的合成技术, 通过估计背景图像的深度和分割信息获取相对平滑区域, 并将字符做随机颜色、尺度等预处理添加到背景图像的平滑区域中完成合成操作, 有效拟合了真实图像的上下文关系. 基于该工作, Montserrat 等人[34]对视觉标志图像做旋转、颜色变换, 并使用泊松融合将模板添加到背景图像的平滑区域中, 但是泊松融合的羽化效果不符合真实图像的内容分布. 这些方法主要考虑了视觉标志或场景字符预处理提高多样性、合成位置和合成方法的选择来使合成图像符合真实图像内容分布. 但是存在如下几方面问题, 主要包括: (1) 在合成时首先对视觉标志进行随机仿射变换, 不符合实际情况; (2) 任意选择叠加位置, 或根据颜色相关性选择叠加位置, 导致视觉标志不能跟随所依附背景的变形而变形; (3) 在合成时使用泊松融合方法造成羽化或透明效果, 不符合真实场景中视觉标志图像. 因此, 生成拟合真实图像的多样化合成数据, 进而提高模型的泛化能力, 依然是目前亟需解决的问题.

2 标志数据合成方法

标志图片数量庞大、种类多样, 人工标注数据耗时长, 目前公开的数据集无法满足大量数据训练的需求. 本文提出一种标志数据集生成方法, 来拟合真实标志图像数据. 生成方法主要借鉴 Gupta 等人[33]对场景字符图片的合成, 并根据标志的特点进行了改进, 主要流程图如图 1 所示, 分为标志模板图像搜集及掩码生成、背景图像选择和标志模板图像与背景合成这 3 个步骤.

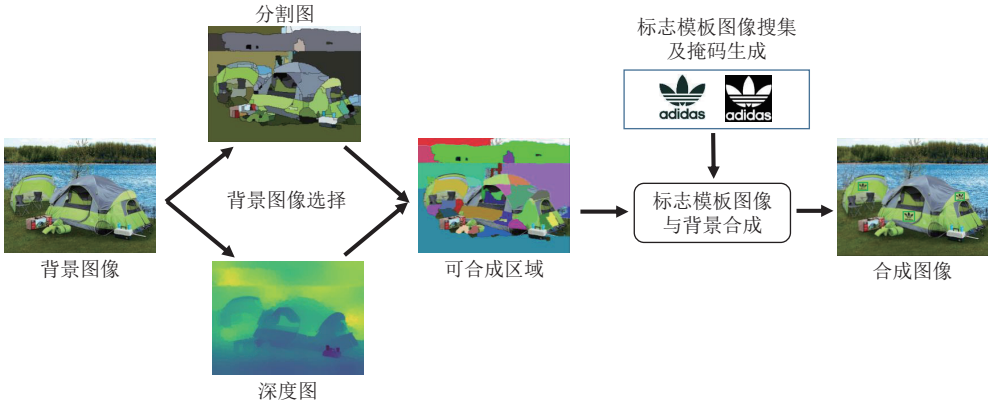


图 1 标志数据集生成流程图

(1) 标志模板图像搜集及掩码生成

目前标志识别主要使用的数据集为 FlickrLogos-32[22], 共包含 32 类标志, 本文同样基于 FlickrLogos-32 数据集, 因此在生成标志数据集时, 搜集的标志类别主要参考 FlickrLogos-32 中的 32 类标志图片. 首先在网络中搜集对应标志图片, 采用简单线性迭代聚类像素分割方法处理标志图像, 自动获得超像素分割结果图, 并经过人工矫正后获得标志区域, 即为标志模板图像.

(2) 背景图像选择

标志数据集生成需要保证最终生成的标志图像拟合真实场景中的标志图像, 因此背景图像的选择需要符合标志经常出现的场景, 本文选择了 FlickrLogos-32 的 6000 张 no-logo 类图像作为背景图像数据源. 同时考虑到标志合成到背景图像中的位置会影响到最终的生成效果, 需要将标志模板图像叠加到背景图像的相对平滑区域, 才能有效拟合真实场景的标志图像. 本文迁移使用 Gupta 等人[33]在场景字符图像合成时对背景图像可合成区域选择方法, 其主要通过图像分割和深度估计确定合适区域.

(3) 标志模板图像与背景合成

颜色常用于增加标志辨识度, 但标志图像由于光照、拍摄方法等影响, 会有相应的颜色变化. 在合成时, 本文不使用泊松融合的合成方法, 而是将标志模板直接变换到背景图像中, 具体步骤如下.

1) 获取每个可合成区域的轮廓坐标.

2) 将对应轮廓坐标转为 3-D 形式, 并将区域进行旋转使其在视线正向区域.

3) 将旋转后的区域平铺到平面上, 即只保留其 x 轴和 y 轴的坐标信息.

4) 获取平面的最小外包矩形, 矩形可以是有角度的, 并根据角度对平面区域进行旋转, 使最终外包矩形角度为 0.

5) 根据变换前后分割区域轮廓坐标的变化获得单应性变换矩阵 H_0:

$$H_0 = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{1}$$

6) 使用单应性变换矩阵 H_0 将变换后的区域外包矩形坐标 (l_1, r_1, r_2, l_2) 扭曲到原图中, 得到原图中区域的 4 点坐标 $(l_1'$, r_1', r_2', $l_2')$. 以 $l_1(x_1, x_2)$ 到 $l_1'(x_1'$, $x_2')$ 的变换为例, 计算公式如下:

$$\begin{bmatrix} x_1' \\ y_1' \\ 1 \end{bmatrix} \sim \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \tag{2}$$

7) 读取标准标志图片和其掩码图片, 将尺度缩放为和背景图片可合成区域的外包矩形一致大小, 并进行颜色抖动预处理. 首先将 RGB 图像转换为 HSV 图像, 然后只改变色调 *hue* 的值, 变换公式如下:

$$hue = (hue + huec)\%180 \tag{3}$$

其中, *huec* 为随机值, 取值范围为 5–30. 在变换之后转化为 RGB 图像.

8) 为使标志合成效果更具有真实性和随机性, 通过外包矩形 4 个点坐标和 6) 中获取的坐标 $(l_1'$, r_1', r_2', $l_2')$ 重新估计单应性变换矩阵 H_1.

9) 根据单应性变换矩阵 H_1 对标准标志图片和其掩码图像进行变换, 并将变换后的图像根据掩码信息直接叠加到背景图像中, 叠加方法如下所示:

$$P(i,j) = \begin{cases} P(i,j), & M_s(i,j) = 0 \\ A(i,j), & M_s(i,j) \neq 0 \end{cases} \tag{4}$$

其中, $P(i,j)$ 为背景图像的对应位置像素, $A(i,j)$ 为标准标志图片对应位置像素, $M_s(i,j)$ 为掩码图像对应位置像素. 将标准标志图片变换到背景图像后可以直接得到标注框.

图 2 中展示了标志图像生成示例, 可以看出本文生成方法可使标志随依附背景的变形而产生变形, 符合真实标志图像的内容分布. 同时, 标志合成区域为较平滑区域, 不会产生突兀感, 而且也会表现出不同尺度, 提高了样本的多样性, 进一步增强网络的泛化能力.

图 2　标志图像合成示例

3　注意力引导的标志检测与识别网络 AGLDN

针对标志尺度变化、非刚性形变的特点, 本文设计注意力引导的标志检测与识别网络 (attention guided logo detection and recognition network, AGLDN), 网络架构如图 3 所示. AGLDN 基于 RetinaNet 构建, 主要是因为 RetinaNet 中已经包含基础特征提取网络和基于 FPN 的多尺度特征融合机制, 并且其作为单阶段目标检测器, 可以在保证精测精度的同时提高检测的速度. 但是 RetinaNet 中仅简单集成了基于 FPN 的多尺度特征融合机制, 不足以应对标志尺度变化、非刚性形变的问题, 因此, 本文在 RetinaNet 的基础上进行了相关改进, 形成针对标志的检测与识别网络 AGLDN, 以提高网络对标志的检测与识别能力.

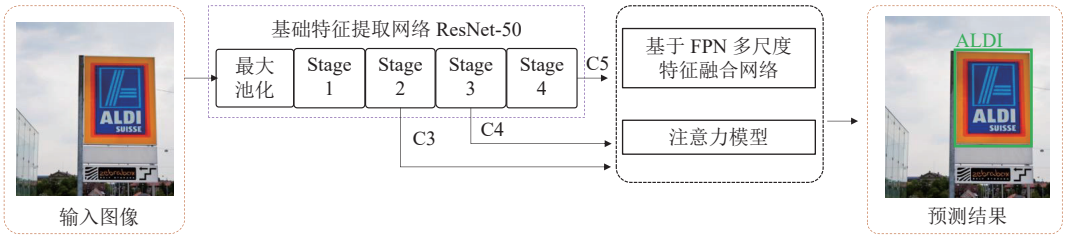


图 3　注意力引导的标志检测与识别网络

首先, 为提高网络应对标志尺度变化的问题, 在网络中引入多尺度特征融合, 主要包括 Inception-Logo 模块和基于 FPN 的多尺度特征融合. Inception-Logo 模块通过多尺度的卷积核提取特征并进行融合, 可以提高语义特征质量, 有效促进后续低层特征和高级特征融合的效果. 进一步, 改进基于 FPN 的多尺度特征融合: 将深特征 P6′和 P7 引入后续特征融合中, 这是 RetinaNet 中不具备的. 其次, 为了有效应对标志非刚性形变问题, 在网络中引入了注意力引导机制来提高感受野中目标特征权重. 注意力引导包括注意力门控模块 (attention gated module, AGM) 和基于双重高效通道注意力 (efficient channel attention, ECA), 其中, AGM 用于平衡普通卷积特征和可变形卷积特征, 当目标物体的变形被精确建模时, 注意力门控机制可以引导普通卷积特征和可变形卷积特征之间的融合, 以有效提升在标志变形时的特征鲁棒性. ECA 机制作为轻量级的通道注意力机制, 通过捕获跨通道的交互信息来生成

特征权重信息, 可以较小的计算代价提高特征鲁棒性.

具体地, AGLDN 的基础特征提取网络选择 ResNet-50, 包含 4 个特征提取阶段, 其中共含有 49 个卷积层; 在 ResNet-50 之后引入多尺度特征融合, 多尺度特征融合主要包括 Inception-Logo 模块以及基于 FPN 的多尺度特征融合, 其可以提高多尺度特征的鲁棒性; 在多尺度特征融合中引入由 AGM 和双重 ECA 组成的注意力引导机制来提高感受野中目标特征的权重, 提高网络对变形目标的特征提取能力, 实现高效的标志的检测与识别.

3.1 多尺度特征融合

真实场景的标志因拍摄角度、位置、焦距的变化引起尺度变化, 为应对尺度多变目标的检测与识别, 往往采用多尺度特征融合的方法, 比如 FPN 通过融合低层特征的描述性内容和高级特征的语义信息, 来提高尺度多变的特征鲁棒性[7]. 但在使用 FPN 进行特征融合时, 高级特征的表达能力将直接影响融合后的特征质量. 为此, 借鉴 Inception-Text 来优化标志图像特征.

(1) 标志特征优化模块 Inception-Logo

Inception-Text 模块, 通过不同大小的空间可分离卷积来处理场景字符的多尺度问题, 并在每个分支的最后添加了一个可变形卷积处理场景字符的多方向问题, 最后融合不同大小卷积支路提取的特征来有效提高特征鲁棒性. 考虑到字符和标志的表现形式不同, 本文将 Inception-Text 中的空间可分离卷积改为标准卷积, 以有效地对较为规则的标志提取特征. 下文将改进后适配于标志的多尺度特征提取的模块称为 Inception-Logo, 其具体结构如图 4 所示. Inception-Logo 通过多尺寸卷积核提取输入特征并进行特征融合, 同时考虑到传统卷积单元针对输入特征图在固定的位置进行采样, 而不同的位置可能对应不同尺度或者不同形变的物体, 这些卷积单元需自动调整尺度或者感受野, 因此, 在 Inception-Logo 中引入了可变形卷积, 提高目标的特征鲁棒性. Inception-Logo 被添加在 C5 卷积层之后, 输入特征的尺度为 19×19, 特征维度为 1 024 维. 首先对输入特征进行 1×1 卷积将特征维度降至 256 维, 降低卷积核提取特征的计算复杂度, 后续分为 a、b、c、d 这 4 个分支处理; 对 c、d 分支分别使用 3×3 和 5×5 的卷积核提取特征, 同时在 b、c、d 分支引入可变形卷积, 提高对尺度和形状多变标志的特征提取能力; 之后使用 Concat 操作将 b、c、d 分支的输入进行通道维度的叠加, 并添加 1×1 卷积做通道之间的特征融合; 最后将融合后的特征与 a 分支输出的特征直接进行相加, 并使用 ReLU 激活函数进行非线性映射, 完成特征的优化.

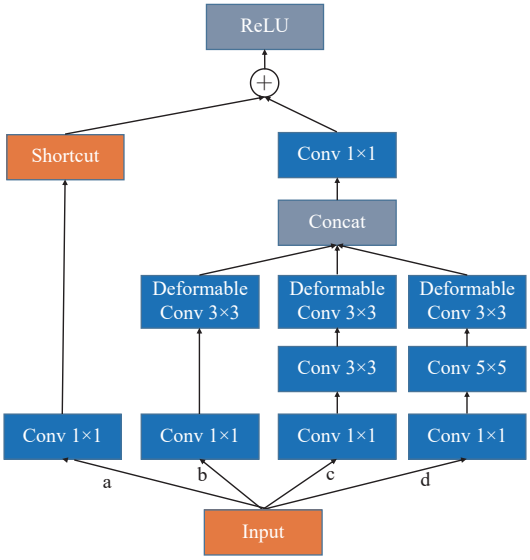
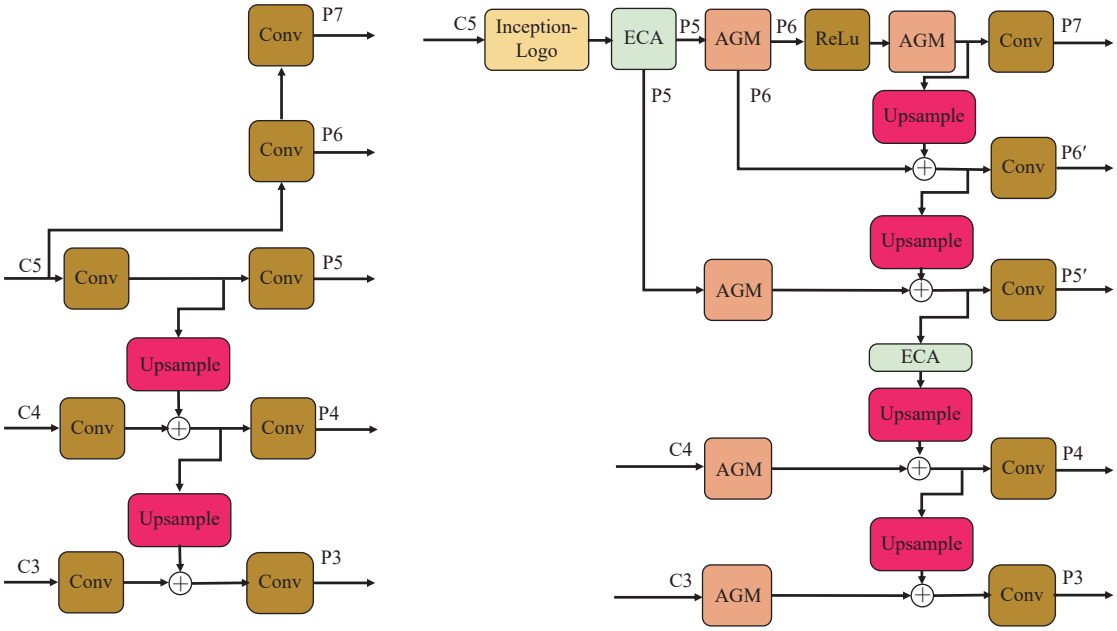


图 4 Inception-Logo 模块网络结构

Inception-Logo 模块通过不同尺度的卷积核提取不同尺度的特征, 完成特征优化. 但 Inception-Logo 的多分支提取特征可能会带来额外的时间代价, 因此将模块添加至网络的高层, 避免增加过多计算量.

(2) 基于 FPN 的多尺度特征融合

RetinaNet 中包含了特征融合结构 FPN, FPN 为 top-down 类型的金字塔特征融合模块, 可以利用高层语义信息, 增加了特征映射的分辨率, 能够有效应对小目标的检测. RetinaNet 中 FPN 的结构如图 5(a) 所示, 其对 C5 层下采样生成了 P6 和 P7 层, 并在{P3、P4、P5、P6、P7}层进行目标的预测, 但是 P6 层和 P7 层并没有参与后续特征融合. 为充分利用语义信息提高特征表达能力, 本文改进了基于 FPN 的特征融合方法, 改进后的多尺度特征融合结构如图 5(b) 所示, P5 层为经过 Inception-Logo 优化后的特征, 在 P5 层特征下采样得到 P6 和 P7 层特征后, 将 P6 层和 P7 层的特征分别与 P5、P6 层特征融合得到 P5′和 P6′.



(a) RetinaNet 中基于 FPN 的多尺度特征融合　　(b) AGLDN 中基于 FPN 的多尺度特征融合

图 5　基于 FPN 的多尺度特征融合网络结构

融合前后的特征图尺度变化如表 1 所示, 融合后共包含{P3、P4、P5′、P6′、P7}这 5 种不同尺度的特征图, 为降低后续计算的复杂度, 在融合初始就将特征通道降低为 256 维, 网络最后的检测层在这 5 种尺度的特征图上做类别和位置的预测. 通过整合多个特征图组成 top-down 类型的金字塔特征融合结构, 可以有效融合低层特征的描述性内容和高级特征的语义信息, 来提高特征对尺度变化的鲁棒性, 进一步提升不同尺度标志特征的检测精度.

表 1　多尺度特征融合后的特征图尺度变化

融合前		融合后	
卷积层	特征图尺度	卷积层	特征图尺度
C3	75×75×512	P3	75×75×256
C4	38×38×1024	P4	38×38×256
C5	19×19×2048	P5′	19×19×256
		P6′	10×10×256
		P7	5×5×256

3.2　注意力引导机制

考虑到标志图像背景复杂, 标志出现在柔性载体或者非平面实体上容易存在非刚性形变, 因此, 提高感受野中

目标特征的权重, 对提高标志检测与识别的性能具有重要的意义. 借鉴门控特征融合机制[9]的原理, 提出注意力门控模块, 用于特征融合来提升目标特征的鲁棒性. 此外, 考虑特征融合若能进行通道间的内容交互, 则可依赖通道选择来有效提高特征权重, 而 ECA[27]作为轻量化的通道注意力机制, 已经被证明可以有效提升视觉目标识别的精度. 因此, 提出使用双重 ECA 在特征融合时进行特征优化, 进一步提高标志检测与识别的性能.

(1) 注意力门控模块 AGM

AGLDN 在 5 种不同尺度的特征图上进行位置和类别的预测, 提高特征质量对检测与识别的性能影响较大. 因此, 本文在特征融合时使用注意力门控模块替换原本的普通卷积, 以获得更具分辨力的特征.

考虑到普通卷积特征通常无法对变形的目标的进行有效建模, 而可变形卷积可以根据变化的对象外观自适应地提取特征. 但是当目标对象处于外观变化较小的普通场景中时, 标准卷积特征可能是更有效的, 因此仅仅依靠可变形卷积可能存在不足. AGM 通过注意力门控机制来平衡正常卷积特征和可变形卷积特征, 当目标的变形被精确建模时, 注意力门控机制可以引导可变性卷积特征和标准卷积特征之间的融合, 从而提升特征鲁棒性.

AGM 的结构如图 6 所示, 首先对输入特征 F 进行 3×3 卷积, 将特征维度统一为 256, 降低后续计算复杂度, 同时有助于通道间的特征交互, 得到特征图 F_n; 然后分别对 F_n 进行 1×1 卷积和可变形卷积提取特征, 分别得到 F_s 和 F_d; 接着对 F_s 和 F_d 按照公式 (5) 进行加权求和, 获得特征图 F_o.

$$F_o = \alpha \otimes F_s + (1-\alpha) \otimes F_d \tag{5}$$

其中, $\otimes$ 代表逐元素相乘. α 是可学习的权重参数, 为使用 Sigmoid 非线性映射所得; 本文使用 α 和 (1–α) 分别作为标准卷积特征和可变形卷积特征的权重. 当标志出现变形时, α 较小, 从而增大可变形卷积特征的权重, 而当标志并没有出现变形时, α 增大, 从而提高标准卷积特征的权重. 依赖注意力引导机制, 可以有效结合标准卷积和可变形卷积特点, 获得鲁棒的更具表征力的标志特征.

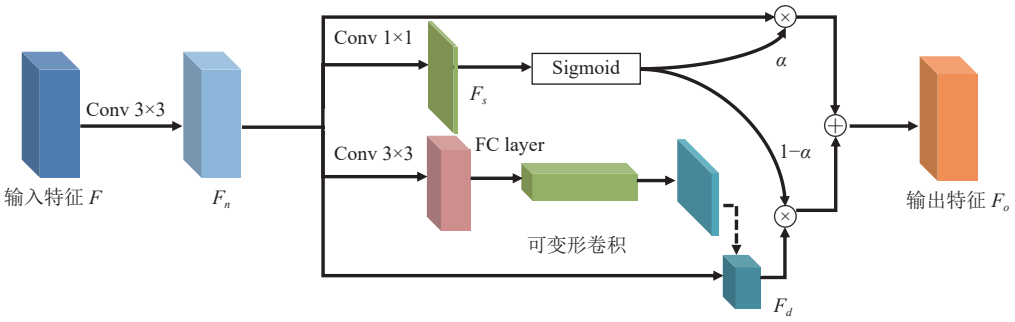


图 6 注意力门控模块结构

(2) 基于双重 ECA 的特征优化

标志容易存在非刚性形变, 导致大量的背景信息干扰标志特征的鲁棒性, 注意力机制可以提高网络对目标的关注度, 在目标变形时, 引导提高感受野中目标区域的特征权重, 优化特征的表达能力[35]. 在注意力机制中, ECA 通过内核大小自适应的一维卷积实现局部跨通道的信息交互, 可以保证较低计算复杂度的同时提高有效特征的权重, AGLDN 也采用 ECA 模块来调整感受野中标志特征的权重.

双重 ECA 结构如图 7 所示. 添加的第 1 处位于 Inception-Logo 后, Inception-Logo 模块中涉及使用不同大小的卷积核提取特征并进行融合, 在添加 ECA 模块后可以有效选择特征. 第 2 处在 P5′特征之后, 其可以更有效地促进高层特征和底层特征的融合. 首先对输入特征 F 进行通道级全局平均池化 (global average pooling, GAP) 得到特征图 F_G; 然后使用卷积核大小为 k 的一维卷积来捕获跨通道的交互信息, 生成特征图 F_c, 本文中 k 取值为 5; 接着使用 Sigmoid 激活函数处理交互信息生成权重 β; 最后根据式 (6) 加权初始输入特征图 F, 得到加权后的特征 F_o.

$$F_o = \beta \otimes F \tag{6}$$

通过使用双重 ECA 机制, 可以合理设置感受野中标志特征的权重.

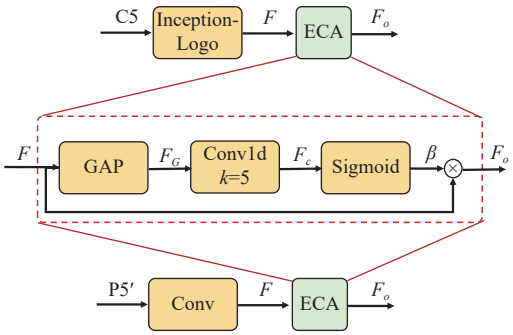


图 7　基于双重 ECA 的特征优化示意图

4　实验结果与分析

4.1　实验设置

本文共使用两个数据集进行实验, 其一为真实数据集 FlickrLogos-32, 共包含 32 类标志, 每类标志 70 张图像, 共 2 240 张图像; 其二为合成数据集, 共 32 类, 对应 FlickrLogos-32 数据集的类别, 每类 100 张图像, 只用于训练. 在训练 AGLDN 及相关模型时, 输入图像的分辨率选择 600×600, 输入图像的标注为 BOX 标注, 使用 Adam 优化器进行训练, 初始学习率为 0.000 01. 实验中采用的机器配置为 NVIDIA GeForce RTX 2080 Ti GPU, 测试时使用 PASCAL VOC 2012 数据集[36]的评价标准. 本文通过 4 组实验来验证标志数据生成方法和 AGLDN 的性能.

4.2　标志数据生成对检测性能的影响

为验证标志数据生成对检测性能的影响, 使用 AGLDN 的基础网络 RetinaNet 进行效果测试, 鉴于 SCL (synthetic context logo)[29]和 CSD (context-based synthetic data)[30]在测试时使用了 Faster R-CNN 网络, 为了客观对比性能, 本文也使用了 Faster R-CNN 网络, RetinaNet 和 Faster R-CNN 使用相同的训练配置. 实验参考 SCL[29]和 CSD[30]对数据合成测试的方法, 使用较少真实数据进行训练, 将 FlickrLogos-32 的每类 10 张图像作为训练集, 其余每类 60 张图像作为测试集. 设计 3 类数据集检验标志数据生成对模型性能的影响: (1) RealImg: 仅使用 FlickrLogos-32 训练集进行训练, 在 FlickrLogos-32 的测试集进行测试; (2) SynImg: 仅使用生成数据 (每类 100 张) 进行训练, 在 FlickrLogos-32 的测试集进行测试; (3) SynImg+RealImg: 首先使用生成数据 (每类 100 张) 进行训练, 然后使用 FlickrLogos-32 训练集进行微调, 最后在 FlickrLogos-32 的测试集进行测试.

性能测试结果如表 2 所示. RetinaNet 使用 RealImg 数据集 mAP 为 54.3%, 而使用 SynImg+RealImg 数据集 mAP 为 64.8%, 提高了 10.5%. Faster R-CNN 使用 RealImg 数据集 mAP 为 53.3%, 而使用 SynImg+RealImg 数据集 mAP 为 63.2%, 提高了 9.9%. SCL 和 CSD 方法均基于 Faster R-CNN 开展研究工作, 两种方法使用 SynImg+RealImg 数据集相较于 RealImg 数据集 mAP 分别提高了 4.4% 和 8.0%. 而使用 Faster R-CNN 时, 所提数据合成方法提高的 mAP 值比 SCL[29]所提方法高 5.55%, 比 CSD[30]所提方法高 1.95%. 这充分说明了使用合成的标志数据可以有效提升样本的多样性, 进而提高标志检测的精度. 同时, 从表 2 中可以看出, 当使用 SynImg 策略进行训练时, CSD 取得了最高的 mAP, 为 32.6%, 而 RetinaNet 的 mAP 仅为 30.1%. 这主要是由于本文所使用合成方法并没有去过度拟合 FlickrLogos-32 的数据内容分布, 包括标志模板的多样化选择、标志模板图像与背景合成的多样性等. 通过所提合成方法能够生成多样化的数据, 从而提高模型的泛化能力. 从表 2 中可以看出, 对比的几种方法均在使用 SynImg+RealImg 训练集时获得了更好的性能.

4.3　消融实验

该实验采用和实验一不同的训练和测试策略以便于对比增加各模块后的性能, 具体地, 在使用生成数据训练

时使用每类 100 张图像, 在使用 FlickrLogos-32 进行训练时使用每类 40 张真实图像, 最终在 FlickrLogos-32 的每类 30 张真实图像上进行测试. 消融实验结果如表 3 所示, 本文对比了使用 RetinaNet、使用生成数据预训练、多尺度特征融合和注意力引导机制等在标志检测与识别任务中的性能.

表 2 使用不同训练数据集的 mAP 对比 (%)

数据集	RetinaNet	Faster R-CNN	SCL[29]	CSD[30]
RealImg	54.3	53.3	50.4	50.5
SynImg	30.1	20.2	27.6	32.6
SynImg+RealImg	64.8	63.2	54.8	58.5

表 3 不同模块对检测精度和检测速度的影响

模型架构	使用合成数据预训练	使用被变形卷积	多尺度特征融合		注意力引导机制		mAP (%)	检测速度 (f/s)
			Inception-Logo	基于FPN的多尺度特征融合	AGM	双重ECA		
RetinaNet	—	—	—	—	—	—	81.37	**21.9**
RetinaNet+S	√	—	—	—	—	—	84.24	21.9
RetinaNet+SI	√	√	√	—	—	—	85.76	19.5
RetinaNet+SF	√	√	—	√	—	—	84.86	21.3
RetinaNet+SIF	√	√	√	√	—	—	85.88	17.3
RetinaNet+SA	√	√	—	—	√	—	86.37	15.6
RetinaNet+SE	√	√	—	—	—	√	86.03	21.3
RetinaNet+SAE	√	√	—	—	√	√	86.52	14.5
AGLDN-D	√	—	√	√	√	√	86.84	12.9
AGLDN	√	√	√	√	√	√	**87.16**	12.7

在表 3 中, RetinaNet 代表使用原始的 RetinaNet, 不添加任何改进措施; RetinaNet+S 代表使用合成数据参与模型的训练; RetinaNet+SI 代表使用合成数据参与训练, 并仅在 RetinaNet 网络的基础上添加 Inception-Logo; RetinaNet+SF 代表使用合成数据参与训练, 并仅对 RetinaNet 网络中的 FPN 特征融合方法进行相关改进; RetinaNet+SIF 代表使用合成数据参与训练, 并在 RetinaNet 中添加完整的多尺度特征融合; RetinaNet+SA 代表使用合成数据参与训练, 并仅在 RetinaNet 网络中添加注意力门控模块 AGM; RetinaNet+SE 代表使用合成数据参与训练, 并仅在 RetinaNet 网络中双重 ECA; RetinaNet+SAE 代表使用合成数据参与训练, 并在 RetinaNet 网络中添加完整的注意力引导机制; AGLDN 代表注意力引导的标志检测与识别网络, 其使用合成数据参与训练, 并在 RetinaNet 的基础上引入多尺度特征融合和注意力引导机制. 此外, 基于完整结构 AGLDN 设计了不使用可变形卷积的 AGLDN-D, 以评估可变形卷积模块的作用. 图 8 列举了使用 RetinaNet、RetinaNet+S、RetinaNet+SIF、RetinaNet+SAE 和 AGLDN 进行检测的示例, 直观展示主要模块的检测性能.

如表 3 所示, 在仅使用基础网络 RetinaNet 时, mAP 为 81.37%, 速度为 21.9 f/s, 从图 8 可以看出, RetinaNet 存在较多检测错误的情况, 同时对于密集标志检测也不具有鲁棒性. 联合使用合成数据训练的 RetinaNet+S mAP 为 84.24%, 相比不使用合成数据提高了 2.87%, 参照图 8 中的检测结果, RetinaNet+S 的检测结果相较于 RetinaNet 有了较大改善, 纠正了 erdinger 图像中的检测错误, 同时也在 rittersport 图像中正确检测出了更多的标志区域, 但是在尺度变换较多的 pepsi 图像上表现不佳, 并没有识别出较多正确的 pepsi 标志. 此外, 表 3 中也展示了添加多尺度特征融合的结果, 其中多尺度特征融合包括 Inception-Logo 以及基于 FPN 的多尺度特征融合, 在仅添加 Inception-Logo 的 RetinaNet+SI mAP 为 85.76%, 相较于未添加的 RetinaNet+S, 提高了 1.52%. 而使用改进后的基于 FPN 的多尺度特征融合模块的 RetinaNet+SF mAP 为 84.86%, 相较于使用原始 FPN 多尺度特征融合的 RetinaNet+S, 其 mAP 提高了 0.62%. 在添加以上两个组件后组成完整的多尺度特征融合模块 RetinaNet+SIF,

mAP 为 85.88%, 相较于 RetinaNet+S, 提高了 1.64%. 如图 8 所示, RetinaNet+SIF 在一定范围内改善了 RetinaNet+S 的检测结果, 例如在尺度较为多变的 pepsi 图像中正确识别出了更多的标志区域, 可得出 RetinaNet+SIF 在尺度变换较多的标志检测与识别任务中具有一定的鲁棒性, 结合表 3 的性能数据, 证明了多尺度特征融合可以有效提升标志检测与识别的精度. 在表 3 展示的添加注意力引导机制结果中, 注意力引导机制主要包括 AGM 以及双重 ECA 机制, 在仅添加 AGM 的 RetinaNet+SA 的 mAP 为 86.37%, 在仅添加双重 ECA 的 RetinaNet+SE 的 mAP 为 86.03%, 在添加以上两个组件后组成完整的注意力引导机制, RetinaNet+SAE 的 mAP 为 86.52%, 速度为 14.5 f/s, 相较于 RetinaNet+S 增长了 2.28% 的精度. 从表中可以看出, 使用注意力引导机制相较于多尺度特征融合有更为明显的精度涨幅, 这主要是由于 RetinaNet 中本来就含有特征融合机制, 所提多尺度特征融合方法是在 RetinaNet 基础上的改进, 同时这也证明了注意力引导机制在提高标志检测与识别精度的有效性. AGLDN-D 的 mAP 为 86.84%, 相比使用可变形卷积的 AGLDN 下降了 0.32, 而速度为 12.9 f/s 相比于 AGLDN 增加了 0.2 f/s. 这一结果表明, 引入可变形卷积可进一步提高模型的检测精度, 且增加的计算开销较小.

(a) RetinaNet (b) RetinaNet+S

(c) RetinaNet+SIF (d) RetinaNet+SAE

(e) AGLDN

图 8 RetinaNet、RetinaNet+S、RetinaNet+SIF、RetinaNet+SAE 和 AGLDN 的结果示例

结合图 8 的检测结果, RetinaNet+SAE 相较于 RetinaNet+SIF 更为准确, 例如, 其纠正了 RetinaNet+SIF 在 milka 图像上的检测错误, 这也符合表 3 中展示的精度数据, 验证了 RetinaNet+SAE 通过注意力引导机制来优化特征质量的效果. 但是在添加注意力引导机制后速度由 21.9 f/s 降低至 14.5 f/s, 降低了 7.4 f/s, 而在添加多尺度特征融合后检测速度从 21.9 f/s 降低至 17.3 f/s, 仅降低了 4.6 f/s, 即注意力引导机制相较于多尺度特征融合提升了较多

精度, 但是也带来了更多的计算开销. 最后, 在 RetinaNet 中添加所有的模块后, 构成了标志检测网络 AGLDN, 测试结果显示 mAP 为 87.16%, 相较于原始的 RetinaNet 提高了 5.79%, 相较于添加多尺度特征融合的 RetinaNet+SIF 提高了 1.38%, 相较于添加注意力引导机制的 RetinaNet+SAE 提高了 0.68%, 这充分证明了 AGLDN 在提高标志检测与识别精度方面的有效性. 从图 8 中也可以看出, AGLDN 获得了更高的召回率, 例如在 pepsi 图像中正确识别出了更多的标志区域.

通过特征热力图可以直观地观察到注意力引导机制有助于标志特征的聚焦. 这里利用 Grad-CAM (梯度加权类激活映射)[37]可视化了注意力权重, 图 9 展示了加入注意力引导机制的权重前后的热力图实例. 从结果中可以看出在没有加入注意力引导机制时, 模型在各个实例中都不能精准的聚焦标志区域. 例如, 其在 a、b、c、d 和 e 等非标志区域有着较大权重, 而对于标志区域 f、g、h、i 无法较为精准地定位标志中心及尺寸, 特别是在具有较大形变的 milka 和 pepsi 实例中, 未加入注意力的模型的权重出现了发散的情况, 无法聚焦标志区域. 而在加入所设计的注意力机制后, 较好地解决了上面的问题. 这一结果表明注意力门控模块 AGM 和双重 ECA 构建的注意力机制, 能够提高标志特征的权重, 帮助模型有效聚焦标志区域.

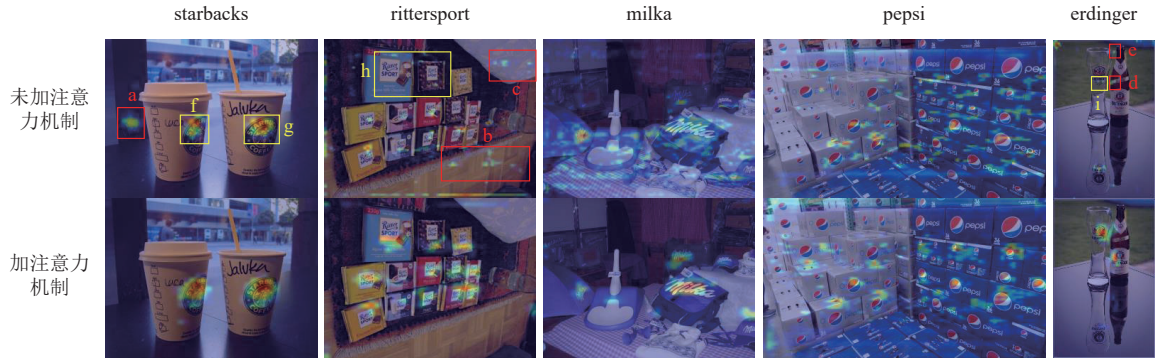


图 9 加入注意力引导机制前后的权重热力图示例

由实验结果可以看出, AGLDN 可以通过多尺度特征融合来有效应对标志尺度多变的情况, 而注意力引导机制的引入一方面可以有效促进可变形卷积特征和普通卷积特征的融合, 提高网络对形变目标的建模能力, 另一方面可以优化目标特征, 提高感受野中标志特征的权重, 进而提高标志检测与识别的精度. 但 AGLDN 在取得较高精度的同时也带来了计算代价, 其检测速度仅为 12.7 f/s, 相较于原始的 RetinaNet 降低了 9.2 f/s.

4.4 AGLDN 和其他方法的性能对比

实验 3 在 FlickrLogos-32 数据集上对比了 AGLDN 和其他方法的性能, 在训练 AGLDN 时, 首先使用合成数据预训练模型, 优化模型参数, 然后使用 FlickrLogos-32 中每类 40 张图像进行训练, 其余每类 30 张图像进行测试.

实验数据如表 4 所示, 本文所提的 AGLDN 方法 mAP 达到 87.16%, 而 DeepLogo[38]、BD-FRCN-M[39]、RealImg[29]、RDSL[31]、Logo-Yolo[40]、MFDNet[5]均为研究者针对标志设计的检测与识别方法, 其中 MFDNet 具有最好的性能优势, mAP 达到 86.20%, 相比之下, AGLDN 使得 mAP 提升了 0.96%. 此外, 实验也对比了两阶段检测器 Faster R-CNN[13]和单阶段检测器 YOLOX[41]的结果, 在训练时使用和 AGLDN 一样的配置. 其中 Faster R-CNN mAP 为 84.86%, 相比之下, AGLDN 提高了 2.30% 的精度. 而 YOLOX 的 mAP 为 86.65%, 其取得了除 AGLDN 外最高的检测精度. 而 Logo-Yolo 因其采用的网络是早期的 YOLOv3, 性能较低.

该实验证明, AGLDN 能够有效解决标志检测识别的难题. 首先所提的标志数据生成方法能有效提高模型的泛化能力, 提高标志检测的鲁棒性; 通过构建多尺度特征融合网络, 有效融合多尺度特征, 能有效应对标志尺度变化问题; 在网络中添加可变性卷积和注意力引导机制能够提高网络对形变目标的建模能力, 降低目标形变对检测性能的影响, 同时优化特征表达能力, 进而提高标志检测精度.

表 4 AGLDN 和其他方法的 mAP 结果 (%)

方法	mAP
DeepLogo[38]	74.40
BD-FRCN-M[39]	73.47
RealImg[29]	81.10
RDSL[31]	82.00
Logo-Yolo[40]	76.11
MFDNet[5]	86.20
Faster R-CNN[13]	84.86
YOLOX[41]	86.65
AGLDN	**87.16**

4.5 AGLDN 和其他方法的检测结果对比

为了更直观展示所提方法的检测性能, 实验四对比了使用 Faster R-CNN、YOLOX 和 AGLDN 方法的检测结果, 如图 10 所示. Faster R-CNN 和 YOLOX 均存在错误识别的情况, 例如在类别为 milka 的标志检测中, 两者都将非标志物体识别为标志, 其中, Faster R-CNN 将非标志物体识别为 pepsi, 而 YOLOX 则将非标志物体识别为 apple, YOLOX 还将变形的 milka 检测成了 adidas. 但是 YOLOX 在 pepsi 图像的检测中表现出相较 Faster R-CNN 更明显的优势, 其正确检测出了更多的标志区域. AGLDN 相较于 Faster R-CNN 和 YOLOX 展现出了更佳的检测结果, 其不仅纠正了两者在 milka 中的检测错误, 还在 rittersport 中正确检测了更多的标志区域. 检测示例符合表 4 中的 mAP 数据, 展现了 AGLDN 在标志检测与识别任务中的优势, 其可以有效提高标志检测与识别的精度.

(a) Faster R-CNN

(b) YOLOX

(c) AGLDN

图 10 使用 Faster R-CNN、YOLOX 和 AGLDN 的检测结果示例

4.6 AGLDN 实测结果

此外, 为了验证本文方法的鲁棒性, 还随机从网络上选取了真实数据, 测试了 AGLDN 在这些数据上的检测效果, 部分测试结果如图 11 所示. 结果表明对于测试集之外的数据, AGLDN 也能较为精准地检测与识别图像中出现的相应标志, 具有较好的鲁棒性.

图 11　使用 AGLDN 在真实数据上的检测结果示例

5 结　语

为了有效应对标志检测面临尺度变化、非刚性形变的问题, 本文提出了一种注意力引导的标志检测与识别网络. 首先使用标志数据集生成方法获取标志数据集, 满足深度学习训练需求, 提高模型的泛化能力; 然后基于 RetinaNet 构建标志检测与识别网络, 通过 Inception-Logo 和基于 FPN 的多尺度特征融合来提取多尺度特征并形成高级语义特征映射; 最后利用注意力门控模块 AGM 和双重 ECA 构建的注意力机制引导机制来关注标志区域, 调整标志特征的权重. 通过统一模型联合优化, 提高标志检测与识别的鲁棒性.

为了进一步提高标志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, 可在合成过程中充分考虑标志和背景图像场景匹配问题, 增加数据真实性和多样性; (2) 本文着重研究了静态标志检测与识别, 动态的标志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标志的检测, 即进一步提高精度; (3) 部分标志包括文字信息, 可添加文字识别分支进一步提升包含字符的标志检测与识别精度.

References:

[1] Gao Y, Wang FL, Luan HB, Chua TS. Brand data gathering from live social media streams. In: Proc. of the 2014 Int'l Conf. on Multimedia Retrieval. Glasgow: ACM, 2014. 169–176. [doi: 10.1145/2578726.2578748]

[2] Zhang GP. Visual logo detection and recognition technology based on deep learning [MS. Thesis]. Beijing: Beijing University of Technology, 2022 (in Chinese).

[3] Yuan Y, Xiong ZT, Wang Q. VSSA-NET: Vertical spatial sequence attention network for traffic sign detection. IEEE Trans. on Image Processing, 2019, 28(7): 3423–3434. [doi: 10.1109/TIP.2019.2896952]

[4] Gandhi S, Kokkula S, Chaudhuri A, Magnani A, Stanley A, Ahmadi B, Kandaswamy V, Ovenc O, Mannor S. Scalable detection of offensive and non-compliant content/logo in product images. In: Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision. Snowmass: IEEE, 2020. 2236–2245. [doi: 10.1109/WACV45572.2020.9093454]

[5] Hou Q, Min WQ, Wang J, Hou SJ, Zheng YJ, Jiang SQ. FoodLogoDet-1500: A dataset for large-scale food logo detection via multi-scale feature decoupling network. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. New York: ACM Press, 2021. 4670–4679. [doi: 10.1145/3474085.3475289]

[6] 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]

[7] Lin TY, Dollár P, Girshick R, He KM, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944. [doi: 10.1109/CVPR.2017.106]

[8] Zhu XZ, Hu H, Lin S, Dai JF. Deformable convnets v2: More deformable, better results. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9300–9308. [doi: 10.1109/CVPR.2019.00953]

[9] Liu WX, Song YB, Chen DS, He SF, Yu YL, Yan T, Hancke GP, Lau RWH. Deformable object tracking with gated fusion. IEEE Trans. on Image Processing, 2019, 28(8): 3766–3777. [doi: 10.1109/TIP.2019.2902784]

[10] Zou ZX, Chen KY, Shi ZW, Guo YH, Ye JP. Object detection in 20 years: A survey. Proc. of the IEEE, 2023, 111(3): 257–276, [doi: 10.1109/JPROC.2023.3238524]

[11] 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]

[12] 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]

[13] 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]

[14] Hoi SCH, Wu XW, Liu HT, Wu Y, Wang HQ, Xue H, Wu Q. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. arXiv:1511.02462, 2015.

[15] He KM, Zhang XY, Ren SQ, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]

[16] Eggert C, Zecha D, Brehm S, Lienhart R. Improving small object proposals for company logo detection. In: Proc. of the 2017 ACM on Int'l Conf. on Multimedia Retrieval. Bucharest: ACM Press, 2017. 167–174. [doi: 10.1145/3078971.3078990]

[17] Li YY, Shi QY, Deng JF, Su F. Graphic logo detection with deep region-based convolutional networks. In: Proc. of the 2017 IEEE Visual Communications and Image Processing. St. Petersburg: IEEE, 2017. 1–4. [doi: 10.1109/VCIP.2017.8305065]

[18] 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]

[19] 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]

[20] Yang QP, Cheng ML, Zhou WM, Chen Y, Qiu MH, Lin W. IncepText: A new inception-text module with deformable PSROI pooling for multi-oriented scene text detection. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI, 2018. 1071–1077. [doi: 10.24963/IJCAI.2018/149]

[21] Dai JF, Qi HZ, Xiong YW, Li Y, Zhang GD, Hu H, Wei YC. Deformable convolutional networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 764–773. [doi: 10.1109/ICCV.2017.89]

[22] 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]

[23] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2017. 4278–4284.

[24] Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar VR, Lu SJ, Shafait F, Uchida S, Valveny E. ICDAR 2015 competition on robust reading. In: Proc. of the 13th Int'l Conf. on Document Analysis and Recognition. Tunis: IEEE, 2015. 1156–1160. [doi: 10.1109/ICDAR.2015.7333942]

[25] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

[26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

[27] Wang QL, Wu BG, Zhu PF, Li PH, Zuo WM, Hu QH. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11531–11539. [doi: 10.1109/CVPR42600.2020.01155]

[28] Zhang C, Kim J. Object detection with location-aware deformable convolution and backward attention filtering. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9444–9453. [doi: 10.1109/CVPR.2019.00968]

[29] 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]

[30] 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]

[31] Song J, Kurniawati H. Exploiting trademark databases for robotic object fetching. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation. Montreal: IEEE, 2019. 4946–4952. [doi: 10.1109/ICRA.2019.8793829]

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

[33] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2315–2324. [doi: 10.1109/CVPR.2016.254]

[34] Montserrat DM, Lin Q, Allebach J, Delp EJ. Logo detection and recognition with synthetic images. Electronic Imaging, 2018, 30: art00018. [doi: 10.2352/issn.2470-1173.2018.10.imawm-337]

[35] Wang XL, Girshick R, Gupta A, He KM. Non-local neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803. [doi: 10.1109/CVPR.2018.00813]

[36] Everingham M, Eslami SMA, van Gool L, Williams CKI, Winn J, Zisserman A. The PASCAL visual object classes challenge: A retrospective. Int'l Journal of Computer Vision, 2015, 111(1): 98–136. [doi: 10.1007/s11263-014-0733-5]

[37] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626.

[38] Iandola FN, Shen AT, Gao P, Keutzer K. DeepLogo: Hitting logo recognition with the deep neural network hammer. arXiv:1510.02131, 2015.

[39] Oliveira G, Frazão X, Pimentel A, Ribeiro B. Automatic graphic logo detection via fast region-based convolutional networks. In: Proc. of the 2016 Int'l Joint Conf. on Neural Networks. Vancouver: IEEE, 2016. 985–991. [doi: 10.1109/IJCNN.2016.7727305]

[40] Wang J, Min WQ, Hou SJ, Ma SN, Zheng YJ, Jiang SQ. LogoDet-3K: A large-Scale image dataset for logo detection. ACM Trans. on Multimedia Computing, Communications, and Applications, 2022, 18(1): 21. [doi: 10.1145/3466780]

[41] Ge Z, Liu ST, Wang F, Li ZM, Sun J. YOLOX: Exceeding YOLO series in 2021. arXiv:2107.08430, 2021.

附中文参考文献:

[2] 张广朋. 基于深度学习的视觉标志检测与识别技术研究 [硕士学位论文]. 北京: 北京工业大学, 2022.

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

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.

靳国庆(1988－), 男, 博士, 副研究员, 主要研究领域为视频编码, 多媒体内容检索.

张勇东(1973－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为网络安全, 多媒体内容分析与处理.

鲁鼎煜(1999－), 男, 硕士生, 主要研究领域为人工智能, 视频处理, 人脸伪造检测.