计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(6):192−200 [doi: 10.15888/j.cnki.csa.009520]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于历史信息及改进 SimSiam 的道路目标检测①

姜世豪, 朱　明

(中国科学技术大学 信息科学技术学院, 合肥 230026)
通信作者: 姜世豪, E-mail: jsh0@mail.ustc.edu.cn

摘　要: 视觉导航旨在通过环境中的视觉信息提供导航依据, 其中关键任务之一就是目标检测. 传统的目标检测方法需要大量的标注, 且只关注图像本身, 并未充分利用视觉导航任务中的数据相似性. 针对以上问题, 本文提出一种基于历史图像信息的自监督训练任务. 该方法聚合同一位置的多时刻图像, 通过信息熵区分前景与背景, 将图像增强后传入 SimSiam 自监督范式进行训练. 并改进 SimSiam 投影层和预测层中的 MLP 为卷积注意力模块和卷积模块, 改进损失函数为多维向量间损失, 以提取图像中的多维特征. 最后, 将自监督预训练所得模型用于下游任务的训练. 实验表明, 在处理后的 nuScenes 数据集上, 本文提出的方法有效提高了下游分类及检测任务的精度, 在下游分类任务上 Top5 准确率达到 66.95%, 检测任务上 mAP 达到 40.02%.

关键词: 历史信息; 自监督学习; 目标检测

引用格式: 姜世豪,朱明.基于历史信息及改进 SimSiam 的道路目标检测.计算机系统应用,2024,33(6):192–200. http://www.c-s-a.org.cn/1003-3254/9520.html

Road Object Detection Based on Historical Information and Improved SimSiam

JIANG Shi-Hao, ZHU Ming

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

Abstract: Visual navigation uses the visual information in the environment as the navigation basis, and one of the key tasks of visual navigation is object detection. Traditional object detection methods require a large number of annotations and only focus on the image itself, failing to fully utilize the data similarity in visual navigation tasks. To solve the above problem, this paper proposes a self-supervised training task based on historical image information. In this method, multi-moment images at the same location are aggregated. Furthermore, the foreground and the background are distinguished by information entropy, and the images are enhanced and then sent into the simple siamese (SimSiam) self-supervised paradigm for training. In addition, the multi-layer perception (MLP) networks in the projection and prediction layers of the SimSiam paradigm are upgraded into a convolutional attention module and a convolution module, and the loss function is improved into one of the losses among multi-dimensional vectors, thereby extracting multi-dimensional features from the images. Finally, the model pre-trained by the self-supervised paradigm is used to train the model for downstream tasks. Experiments reveal that the proposed method effectively improves the precision of downstream classification and detection tasks on the processed nuScenes dataset. Its Top5 precision on downstream classification tasks reaches 66.95%, and its mean average precision (mAP) on downstream detection tasks reaches 40.02%.

Key words: historical information; self-supervised learning; object detection

① 基金项目: 科技创新特区计划 (20-163-14-LZ-001-004-01)
收稿时间: 2023-12-14; 修改时间: 2024-01-17; 采用时间: 2024-01-23; csa 在线出版时间: 2024-04-30
CNKI 网络首发时间: 2024-05-07

1 引言

随着深度学习的不断发展, 视觉导航任务应用场景也不断增加, 对环境的感知需求也愈发提升. 目标检测便是视觉导航任务中极为基础和重要的一项任务, 旨在通过图像框选出其中存在的物体, 从而提供环境感知所需要的语义信息. 传统的目标检测方法基于有监督学习, 需要大量的标注, 极其耗费人力与物力, 且并未充分考虑视觉导航任务中数据之间的关联性. 为此, 学者们提出自监督学习方法. 相较于传统的有监督学习, 自监督不需要昂贵的数据标记, 可以通过大量未标注的数据来训练模型, 从而得到更为通用和丰富的上游特征表达. 在计算机视觉、自然语言处理及语音处理等领域, 自监督学习已经变得越来越重要, 图灵奖得主 LeCun 在 AAAI 上也称自监督学习为未来的大势所趋, 并称之为智能的暗物质. 图片是环境信息的综合, 潜在地包含了物理世界的各种规律和规则, 自监督学习正是通过这些规律获取图像中的一致性信息, 从而学习到较好的上游表征. 图像中自监督表征学习主要包括对比式和掩码重建式两种模式, 都是在单张图片上进行变换操作, 生成两种样本并最小化差异. 视频自监督学习则包括视频速度预测、帧序列验证、帧序列掩码重建等模式, 企图在时间维度上获取更丰富的信息. 在视觉导航任务中, 多次收集的数据之间亦有相似性, 如图 1 所示, 可以在同一位置获取历史图像数据, 其中交通参与者及图像所处天气皆不同. 我们可以通过 GPS 及 SIFT [1]图像特征进行时间维度上的聚合, 充分利用历史信息, 构造质量更高的样本, 使模型学习到更优秀的特征, 从而在下游检测任务上达到更高的精度.

图 1 同一位置的历史图像数据

具体来说, 本文提出了一种基于历史图像信息的图像增强算法, 关注时间及空间上下文信息, 通过多次经过同一区域的图像提取时空特征信息. 这种图像序列在很多场景下都可以轻易地获得, 如定轨巡检车辆、每日上下班的驾车路线等. 对于多次经过的同一场景, 其内背景物体 (如楼房、道路) 具有一致性, 交通参与者 (如车辆、行人) 则动态变化, 可依据该特点设计图像增强方法, 使模型提取到更加精确的上游特征表示. 同时, 本文改进了 SimSiam 自监督学习模型, 并将上述图像增强算法应用在正样本的构建上. 最后, 本文在 nuScenes 数据集[2]上进行实验. 在原始数据集的基础上, 根据 GPS 信息对样本进行聚合, 提取场景中的历史数据并构建上游分类任务样本集和下游检测任务样本集, 验证了方法的有效性.

2 相关工作

2.1 图像自监督学习算法

在图像处理领域, 目前主流的方法可分为两大类, 即基于 Generative 的方法和基于 Constrative 的方法[3]. 其中基于 Generative 的方法对输入图像进行掩膜重建, 主要关注模型的重建能力, 如 BEiT[4]通过 dVAE 和 encoder 分别对掩膜前后图像编码计算重构损失, MAE[5]和 SimMIM[6]则直接对输入图像进行掩膜重建. 基于 Costrative 的方法采用对比学习策略, 使用增强后的图像作为正样本, 其余图像作为负样本进行训练, 如 SimCLR[7–9]使用数据增强处理后的原图像和其余图像分别作为正负样本进行训练, MoCo[9]则设计了移动平均更新模型权重的方法增加负样本数量. BYOL[10]设计了不需要负样本的模型范式, 而 SimSiam[11]则使用梯度停止方法进一步去除了滑动编码器, 使模型训练更加简单. 不同于以上两种方法, PIC[12]则提出了一种单分支结构, 使用滑动窗口、最近负样本采样等策略达到较好的性能.

2.2 目标检测方法

基于深度学习的目标检测方法可分为一阶段、二阶段等算法. 二阶段算法如 RCNN[13]、Fast R-CNN[14]、Faster R-CNN[15]等, 将目标检测分为生成 proposals、生成物体边框两个阶段. 一阶段目标检测算法如 YOLO 系列[16–20]、SSD[20]等, 直接产生物体的类别概率和位置坐标值. 而上述两种方法都基于 Anchor, 带来大量的超参数和复杂的计算量. 诸如 CenterNet[21]、FCOS[22]等网络则摈弃 Anchor, 通过关键点的方式进行检测.

2.3 历史信息聚合方法

视觉导航任务的数据具有背景一致性及交通参

与者多样性等特点, 通过特定处理可以预提取出物体特征. 文献[23]提出了使用过去遍历经验指导当前三维感知任务的方法, 提出一种称为“hindsight optimization”的方法, 旨在为当前任务提供先验信息. 它通过重新构建过去路径上的状态和行动序列, 预测当前感知任务中未知物体的位置和属性. 在实际应用中, 该方法使用了一种基于强化学习的方法来优化过去路径的重构, 以使得重构后的路径尽可能地符合当前感知任务的要求. 文献[24]提出了一种自监督学习的方法, 通过学习对抗干扰物来提高单目视觉里程计在城市环境下的鲁棒性. 该方法通过合成训练数据和干扰物识别网络的优化, 能够有效应对城市环境中的干扰, 提高里程计的准确性和稳定性. MODEST 方法[25]结合了瞬时点及自训练方法, 该方法使用多次通过同一区域附近的雷达点云信息计算瞬时点, 聚类移动物体并自训练物体检测器. 整个过程无需标注, 且取得较好的检测精度.

2.4 本文工作

为了实现对图像历史信息聚合并使用自监督学习算法提取特征, 本文贡献如下: 1) 基于历史信息序列, 使用 GPS、SIFT 特征聚合同一位置的历史图像信息, 计算信息熵以区分前后景, 构建正样本对; 2) 使用 SimSiam 进行自监督训练, 修改其预测层中的 MLP 为卷积注意力模块 CBAM[26], 投影层中的 MLP 为卷积结构, 并更改损失函数, 以提取更丰富的特征; 3) 将自监督训练所得骨干模型在下游分类及检测任务上进行微调. 实验证明, 改进后的算法在下游场景分类任务上可达到 20.83% 的 Top1 准确度与 66.95% 的 Top5 准确率, 高于原 SimSiam 算法的 17.29% 与 60.07%. 在 nuScenes 检测任务上, 改进后的算法可达到 40.02% 的 mAP, 亦高于原始算法的 38.88%.

3 算法设计与实现

主流的自监督学习算法大多分为对比式和掩膜重建式两种, 本文提出的方法在对比式自监督的基础上进行. 考虑到本文算法中历史信息聚合后样本的相似性, 以 SimSiam 为基线模型进行实验. 本节从 SimSiam 模型介绍, 基于历史信息的图像增强方式, SimSiam 模型优化几个部分介绍所做工作.

3.1 SimSiam 模型介绍

在自监督学习中, 常使用最大化同一图像不同增强方式所的样本的相似度来训练特征提取器. 而单纯的最大化相似度可能会出现崩溃解 (collapse solution), 即无论模型输入什么, 都会得到一个恒定的输出, 从而最小化网络损失. 为了解决这一情况, 诸如 MoCo、SimCLR 方法使用同一 batch 中的其余样本当做负样本对, 优化损失函数的设计. 而 BYOL 则是使用孪生网络和动量编码器的方式代替负对, SimSiam 则在其基础上进一步去除了动量编码器, 仅使用梯度停止方法便可得到优秀的结果.

SimSiam 模型结构如图 2 所示, 分为两个分支. 输入图像 x 经过一系列图像增强后得到两个版本 x_1, x_2. 接下来, x_1, x_2 通过同一个编码器 f (此处为 ResNet-50[27]) 编码提取特征, 得到两个表征向量 h_1, h_2. 然后, 两者会通过同一投影层 (包含 3 个全连接层, 每层 2 048 维且最后一个全连接层中不使用激活函数) 经过处理后会得到新的表征向量 z_1, z_2. 最后, 将 z_1 输入预测层 (包含两个全连接层, 输入输出为 2 048 维, 隐藏层为 512 维, 输出层中仅由线性层) 映射得到 v_1, 计算 v_1 和 z_2 的相似度. 在训练中, 将来自同一图片的 x_1, x_2 看作正样本对, 因此需要对称地计算 v_2 和 z_1 的相似度, 并将其最大化. 最终的损失函数如式 (1), 其中 $\mathcal{L}$ 为总损失函数, 具有对称性. $D(p_i,z_i)$ 为余弦损失, 用于度量两个向量之间的相似度.

$$\mathcal{L}=\frac{1}{2}D(p_1,z_2)+\frac{1}{2}D(p_2,z_1) \tag{1}$$

$$D(p_1,z_2)=-\frac{p_1}{\|p_1\|_2}\cdot\frac{z_2}{\|z_2\|_2} \tag{2}$$

在 SimSiam 中, 为了避免崩溃解的出现, 采用了梯度停止的方法, 也即只计算并更新包含预测层通路的梯度, 而不更新另一分支部分权重. 实际训练过程中 f、g、h 皆更新两次, 损失函数为:

$$\mathcal{L}=\frac{1}{2}D(p_1,s.g(z_2))+\frac{1}{2}D(p_2,s.g(z_1)) \tag{3}$$

3.2 基于历史信息的图像增强方式

SimSiam 的正样本生成方式与对比式自监督相同, 都是同一张图片经过一系列不同的图像增强所得到. 而该方式得到的正样本关联信息过多, 模型很容易只学到简单特征, 从而导致下游任务的精度较低. 因此, 如何构建高质量的正样本正是影响自监督模型性能的关键之一.

在视觉导航任务中, 可以天然地依据 GPS 信息区分出当前图像附近及较远区域的图像集合, 从而构建

更精确的正负样本集. 而在正样本中, 多时刻采集到的图像可能含有不同的交通参与者, 从而造成正样本的特征不一致. 大多数情况下, 同一交通参与者不会出现在多个时刻的图像中. 我们可以利用此特性, 统计同一位置在多时刻图像中不相同的区域, 降低其在正样本中的占比, 从而提高正样本一致性.

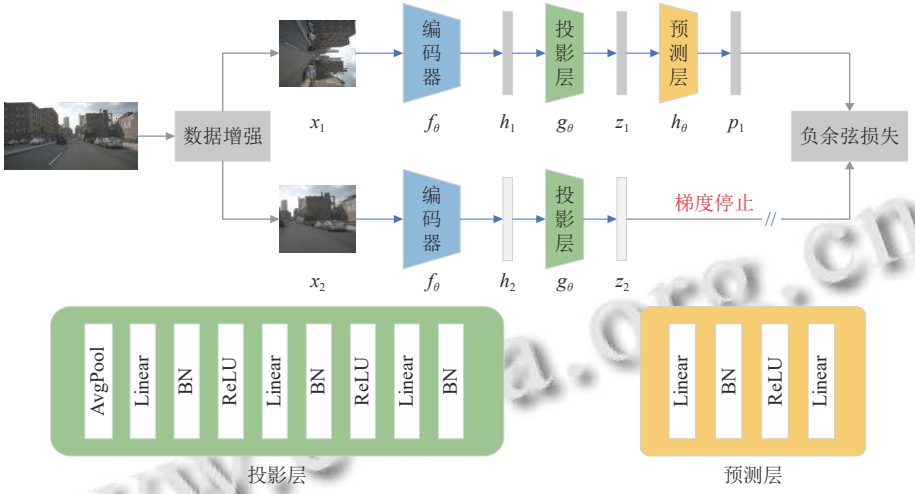

图 2 SimSiam 模型结构

具体来说, 对同一位置c的N次遍历, 使用 GPS 聚合c位置前后r距离内的采集到的图像, 随后通过 SIFT 特征匹配筛选出最相近的N张. 这N张图片可看作同一位置在不同时刻的多次拍摄结果. 对这N张图片, 其中每张图片尺寸为$m\times n$, 在x,y两个方向以k为间距均匀取点, 得到图像上的一组点集$\{(ak,bk),a<m/k,b<n/k\}$, 可以对每个点分别计算其为交通参与者的概率. 对点集中的q点, 计算以q为中心, 周围$r\times r$个点中, 与q灰度差异小于d的点数量. 对于第i次遍历的点q与第j次遍历的点集S_j, 该值为:

$$N_i^j(q)=\left|\left\{p_i\left|\begin{array}{l} dis(p_i,q)<d,\\ p_i\subseteq S_j \text{ and } \|p-q\|_2<r\end{array}\right.\right\}\right| \tag{4}$$

若点q为静态背景的一部分, 那么该点的局部邻域灰度分布在所有遍历中都应相似. 因此, 可以通过$N_i^j(q)$的分布一致性推测出q是否为静态背景点. 为此, 计算:

$$P(i,q)=\frac{N_i^j(q)}{\sum_{j'=1}^{T}N_i^{j'}(q)} \tag{5}$$

$P(i,q)$代表每次遍历邻域点的占比, 并计算归一化后的信息熵.

$$H(q,i)=-\sum_{i=1}^{T}\frac{p(i,q)\log(p(i,q))}{\log(T)} \tag{6}$$

$H(q,i)$越大, 意味着$P(i,q)$与均匀分布间的 KL 散度越小, 也就更接近于均匀分布. 设置一个阈值t, 则低于t的点可认为是前景点. 这些点将图像划分为一个个$k\times k$的分片, 每个分片对应一个熵值. 为了构建更高质量的自监督训练样本, 设计如下基于历史信息的图像增强方法的算法 1.

算法 1. 基于历史信息的图像增强算法

1) 输入当前图像x.
2) 根据当前图像对应 GPS 信息, 加载对应的历史图像列表h.
3) 在图像上以k为步距均匀取点, 根据历史图像信息周围$r\times r$个点计算每个点对应的归一化信息熵, 差异度小于 d 的认为是同一类型点.
4) 对每个点, 若归一化信息熵小于阈值t, 说明该点属于前景点, 则将该点k邻域替换为历史图像的均值:

$$x[-k{:}k+1,-k{:}k+1]=\text{avg}(h[-k{:}k+1,-k{:}k+1])$$

若归一化信息熵大于 t, 说明该点属于背景点, 不做处理.
5) 输出增强后的图像x.

经过上述算法增强后, 图像相似的部分被保留, 不同的部分被均值替代, 从而使样本间具有更高的一致性. 流程结构如图 3.

(1) 输入原始图像 (2) 加载历史图像序列 (3) SIFT 特征筛选匹配 (4) 计算归一化熵值 (5) 增强原始图像并输出

图 3 基于历史信息的图像增强过程

3.3 SimSiam 模型改进

视觉导航任务图像集中, 不同场景图像间差异较小 (如都有成片的天空与道路), 且交通参与者多, 场景复杂, 使用 SimSiam 算法学习时无法学习到很好的特征. 因此, 本文将 SimSiam 的投影层、预测层加以改进, 以提取更优异的骨干特征. 具体包括: 1) 将投影层的 MLP 改进为卷积注意力模块 CBAM, 预测层的 MLP 改为卷积层, 并通过自适应平均池化下采样到$S \times S$维度; 2) 修改损失函数为负余弦相似度矩阵, 每维取最小计算平均损失. 接下来将具体介绍两部分工作.

SimSiam 投影层和预测层皆由全连接层组成, 属于 MLP 结构, 并且在特征提取前使用自适应平均池化将 ResNet-50 输出的特征进行了降维, 使得网络整体架构更加适配分类任务. 而对于本文中的对象目标检测任务, 网络需要更多地关注图像的局部特征, 而不是全局特征.

为了解决这一问题, 本文使用卷积注意力模块 CBAM 构建投影层, 使用卷积层构建预测层, 并将修改损失函数以匹配更多维度的特征向量. 卷积注意力模块如图 4 所示, 包含连续的通道注意力模块及空间注意力模块. 其中, 通道注意力模块如图 5, 将输入特征分别进行全局平均池化及全局最大池化, 再使用共享的 MLP 进行处理, 最后将所得结果进行相加并使用 Sigmoid 函数得到特征层每一个通道的权重, 从而使网络关注权重更高的通道信息. 空间注意力模块如图 6, 将输入特征沿通道方向进行全局平均池化及全局最大池化, 并通过通道为 1 的卷积层及 Sigmoid 函数得到特征向量每个空间位置上的权重, 提升网络对图像局部特征的捕捉能力. 结合这两者, 卷积注意力模块可以更好地建模目标之间的空间关系, 对于目标检测任务更为有效. 同时, 卷积注意力模块可以直接接受骨干网络的多维输出特征, 而不需要降维, 使得特征提取更加充分.

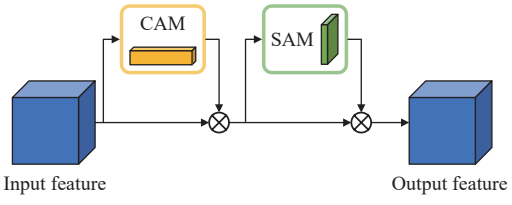

图 4 卷积注意力模块 CBAM 结构图

SimSiam 中使用的损失函数也需要进行相应的更改. 在原始模型中, 作者计算两个 2 048 维度向量内积空间的余弦值, 以度量两个样本的相似度. 在改进后的模型中, 骨干网络输出特征向量个数为$S \times S$, 无法直接使用负余弦相似度计算. 为此, 本文通过自适应平均池化将特征向量下采样为$N \times N$, 计算两个样本之间的负余弦相似度矩阵, 并对每一维度取最小值. 这代表将一个样本的每个特征向量与另一个样本中最相似的特征向量相匹配. 最后, 将所得值求平均输出, 当作总体的损失.

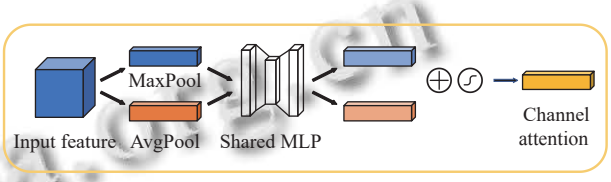

图 5 通道注意力模块 CAM 结构图

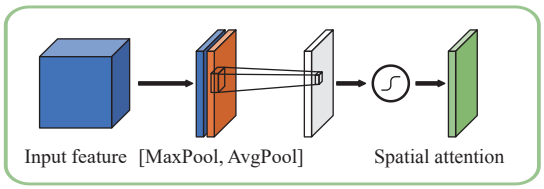

图 6 空间注意力模块 SAM 结构图

改进后的 SimSiam 网络结构如图 7, 其中虚线框内为改进后的部分, 包括基于历史信息的增强、投影层、预测层及损失函数.

4 实验分析

4.1 数据集

实验使用 nuScenes 数据集进行模型训练及验证. nuScenes 是一个公开的自动驾驶数据集, 主要在波士顿和新加坡地区进行数据采集. 与大多数以 camera 为主的公开数据集不同, nuScenes 注重多传感器, 提供了每个位姿的传感器数据. 为了保证客观性及泛化性, 实验中对该数据集做了以下处理.

(1) 截取波士顿地区的数据, 共 15 695 张标注样本. 使用 GPS 数据提取多次经过同一区域片段的样本, 并丢弃少于两个样本的位置.

(2) 在每张标注图像前后各取 15 张未标注图像组成一个样本, 在地理位置上的采样间隔为 2 m, 由 GPS 数据划分, 用于后续历史信息的处理.

(3) 将提取出的标注图像与非标注图像按场景聚合, 得到 227 个场景种类, 并划分训练测试集, 用于下游分类任务评估.

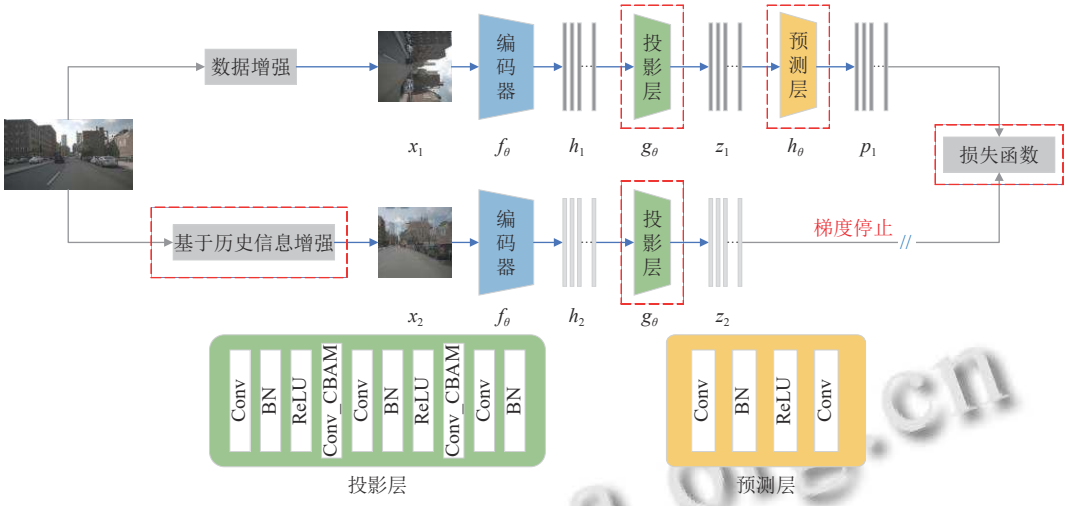

图 7 改进后的 SimSiam 结构图

(4) 将提取出的标注图像组合, 得到 2 590 张标注图像, 组合为增强部分数据集, 并划分训练测试集, 与整体数据集一起用于下游检测任务评估.

本实验中包括自监督预训练及下游分类任务评估、目标检测任务性能评估几个部分. 在自监督预训练过程中, 将上述得到的所有图像放在一起进行训练; 下游分类任务中, 使用场景分类数据集评估分类任务性能; 目标检测任务中, 分别使用增强部分数据集与全部标注图像数据集进行检测任务性能评估.

4.2 实现细节

实验中使用的硬件平台如表 1. 在数据集处理方面, 将1600×900的图像下采样为800×450. 在图像增强部分, 仿照 SimSiam 原始增强方法, 将基于历史信息的增强方式、随机裁剪、随机灰度等数据变化组合成增强管道进行不同分支的图像处理, 其中基于历史信息增强方式中取 k 为 3, r 为 5, d 为 20, t 为 0.6. 在网络模型部分, 以 SimSiam 为主体, 骨干网络采用 ResNet-50, 输入为经过裁剪后的224×224大小图像, 输出为 $batch_size\times2048\times n\times n$维度特征; 投影、预测层改为前文提出的卷积注意力模块及卷积模块; 损失函数改为前文提到的损失函数. 训练时, 设置批次大小为 32, 使用衰减权重为 0.000 1 的 SGD 优化器. 初始学习率大小为 0.062 5, 在余弦衰减策略下调整学习率, 训练 300 个轮次.

在分类任务中, 提取预训练得到的骨干网络, 并将最后一层改为全连接层, 输出维度$1\times class_num$ (在本任务中该维度为 227), 随后使用 MMSelfSup 框架进行下游 ImageNet 分类任务评估. 训练时, 设置批次大小为 512, 使用 LARS 优化器, 初始学习率大小为 1.6, 在余弦衰减策略下调整学习率, 训练 100 个轮次.

表 1 实验所用硬件配置

类型	型号	参数
系统	Ubuntu	16.04.12LTS
CPU	Intel(R) Core(TM) i7-7800X	6核12线程
显存	NVIDIA GeForce GTX 1080Ti	11 GB×2
内存	DDR4	32 GB

在目标检测任务中, 提取预训练得到的骨干网络, 使用 MMDetection 框架搭建 Faster R-CNN C4 目标检测框架, 进行下游目标检测任务评估. 所用数据集包括增强部分数据集和全部数据, 由于数据集中各种类物体标注框的数量差异较大, 从中选取占比较高的 4 个种类进行训练, 具体类别及数量如表 2.

表 2 数据集实例分布 (个)

种类	增强部分数据集		全部数据	
	训练	测试	训练	测试
车辆	4 741	829	45 759	4 986
行人	2 494	311	17 210	1 959
交通牌	2 837	398	11 214	1 175
路障	1 806	297	13 064	1 317

同时, 为了评估模型在不同目标尺寸上的检测性能, 本文对检测目标大小进行划分, 分别计算 3 种尺寸目标的 AP. 其中, 目标框面积小于 32×32 像素的为小目标, 大于 96×96 的为大目标, 之间的为中等目标, 具体数量如表 3.

训练时, 设置批次大小为 4, 使用衰减权重为 0.000 1 的 SGD 优化器. 初始学习率大小为 0.02, 在余弦衰减

策略下调整学习率, 迭代 72 000 个轮次. 其中, 在前 100 个轮次使用预热策略, 48 000 及 64 000 轮次时学习率会下降为原来的 0.1.

4.3 实验结果

4.3.1 下游分类及检测任务

本文在前述数据集的基础上, 以 ResNet-50 为骨干网络, 分别对随机初始化、SimSiam 预训练、BYOL 预训练及本文方法进行训练评估, 结果如表 4.

表 3 数据集标签大小分布 (个)

类别	增强部分数据集		全部数据	
	训练	测试	训练	测试
大目标	1 238	209	10 268	1 123
中等目标	4 697	724	35 803	3 901
小目标	5 943	902	41 176	4 413
总计	11 878	1 835	87 247	9 437

表 4 对比实验-下游任务结果 (%)

预训练模型	分类任务 (Acc)		检测任务									
			增强部分数据					全部数据				
	Top1	Top5	AP50	mAP	AP_s	AP_m	AP_l	AP50	mAP	AP_s	AP_m	AP_l
随机初始化	14.47	56.60	64.20	36.80	16.97	45.86	**38.68**	65.90	38.88	14.76	43.30	56.65
SimSiam	17.29	60.07	65.70	38.49	20.19	47.54	37.64	66.10	39.37	15.05	44.25	56.04
BYOL	17.68	58.50	65.50	38.21	19.96	46.06	35.69	**66.60**	39.44	14.96	44.31	**58.15**
本文算法	**20.83**	**66.95**	**66.90**	**39.86**	**22.09**	**48.15**	36.37	66.40	**40.02**	**15.39**	**44.76**	56.04

可以看出, 在下游分类任务上, 相较于随机初始化, 经过自监督算法 SimSiam 提取特征的骨干网络在分类任务上有了明显的性能提升, 达到了 17.29% 的 Top1 准确率与 60.07% 的 Top5 准确率, BYOL 算法同样达到了相似的效果, 这说明了自监督算法的有效性. 且本文提出的算法在更为有效, Top1 及 Top5 准确率更高, 达到了 20.83% 及 66.95%, 证明了所提出模块在分类任务上的有效性. 同时, 在下游检测任务上, 经过本文算法预训练后的模型在增强部分数据集上达到了 38.49% 的 mAP 与 65.7% 的 AP50 精度, 相较于仅使用随机初始化模型及 SimSiam 算法和 BYOL 算法更高; 在全部数据集上也达到了 40.02% 的 mAP, 高于其余方法. 这说明了模型在下游检测任务上的有效性, 即改进后的模型可以提取更丰富的特征, 使下游检测任务精度更高. 在不同尺寸的检测结果上, 本文算法对小尺寸目标更加有效, 在增强部分数据集上达到了 22.09% 的 AP_s, 在全部数据上也达到了 15.39% 的 AP_s, 为所有方法中最高. 这说明模型可以有效地提取图像细节信息, 对小尺寸目标有更好的检测效果. 检测任务中, 部分检测结果如图 8.

4.3.2 消融实验

为了进一步验证改进模块的有效性, 本文进行了模块之间的消融实验. 以 SimSiam 模型作为基线, 分别进行以下改动: (1) SimSiam+GPS: 将图像按照 GPS 划分, 构建正样本时分别从同类图像中选取不同的两张. (2) SimSiam+历史信息增强: 在 (1) 的基础上, 将图像按照基于历史信息的方法进行增强. (3) SimSiam+卷积: 将 SimSiam 的投影和预测层的 MLP 改为卷积注意力模块和卷积模块. (4) 本文算法: 将本文提出方法全部部署.

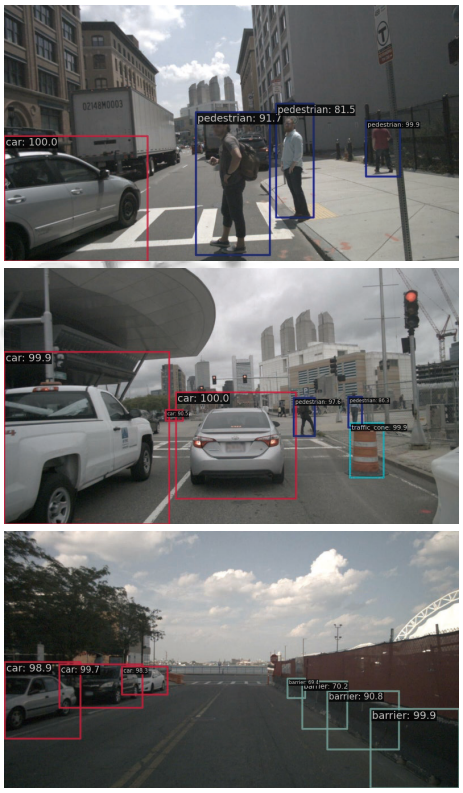

图 8 部分检测结果图

消融实验结果如表 5, 每个方法在基线的基础上都有所提升. 在分类任务上, 仅使用 GPS 聚类就达到了

70.98% 的 Top5 准确率, 而使用历史信息增强后达到了 72.00%, 高于基线的 60.07%, 这说明基于 GPS 信息及历史信息的图像增强效果对骨干网络的特征提取更加有效, 这也与 MoCo[9]中的观点一致, 即有效的数据增强方式可以提升自监督预训练模型的性能. 在检测任务上, 仅使用卷积代替 MLP 也可以提高下游检测任务的性能, 在全部数据上达到 39.74% 的 mAP 与 66.20% 的 AP50 精度, 高于基线的 38.88% 与 65.90%, 这说明卷积注意力模块可以提取更多局部信息, 加强了模型的表征能力, 使其提取出的特征在下游检测任务上更加有效. 本文算法则结合了图像增强及模型改进两个部分, 在全部数据上达到了 40.02% 的 mAP, 为所有实验中最高. 在不同尺寸的检测结果上, 仅使用 GPS 聚类对大目标更加有效, 在增强部分数据集和全部数据集上分别达到了 39.16% 和 58.14% 的 AP_l. 添加了本文提出模块的方法则对中小目标检测精度更高, 其中, 仅使用卷积可以在全部数据集上分别达到 45.02% 的 AP_m 和 16.04% 的 AP_s, 为所有实验中最高. 本文所提出方法则在两种数据集上分别达到 22.09% 和 15.39% 的 AP_s. 这说明本文所提出的模块对图像的局部特征信息提取较为充分, 在下游任务上对小目标的检测更加有效.

表 5 消融实验-下游任务结果 (%)

预训练模型	分类任务 (Acc)		检测任务									
			增强部分数据					全部数据				
	Top1	Top5	AP50	mAP	AP_s	AP_m	AP_l	AP50	mAP	AP_s	AP_m	AP_l
SimSiam	17.29	60.07	65.70	38.49	20.19	47.54	37.64	65.90	38.88	15.05	44.25	56.04
+GPS	20.78	70.98	66.70	38.94	20.84	47.77	**39.16**	66.60	39.53	15.04	44.04	**58.14**
+历史信息	21.52	**72.00**	66.40	39.05	21.50	**48.22**	36.24	**66.70**	38.58	15.27	43.94	55.26
+卷积模块	17.19	55.40	66.80	38.78	21.62	47.54	34.73	66.20	39.74	**16.04**	**45.02**	57.28
本文算法	**20.83**	66.95	**66.90**	**39.86**	**22.09**	48.15	36.37	66.40	**40.02**	15.39	44.76	56.04

5 结论与展望

本文出一种基于历史信息及改进 SimSiam 的自监督学习算法, 基于历史信息计算归一化熵值进行数据增强, 以构建更高质量的样本对; 改进投影层的 MLP 为卷积注意力模块、改进预测层的 MLP 为卷积模块, 以感知局部特征, 训练更适合下游检测任务的骨干网络; 改进损失函数为最小化负余弦特征矩阵行的平均损失, 使其可以对密集特征向量计算损失. 本文方法在处理后的 nuScenes 数据集上进行实验, 在分类任务上达到 20.83% 的 Top1 分类精度和 66.95% 的 Top5 分类精度, 在增强部分数据集的检测任务上达到 39.86% 的 mAP 与 66.90% 的 AP50 精度, 在全部数据的检测任务上达到 40.02% 的 mAP 和 66.40% 的 AP50 精度. 且相较于其他方法, 本文方法在小尺寸目标上的检测精度更高, 在增强部分数据集和全部数据集上分别达到了 22.09% 和 15.39% 的 AP_s. 虽然本文算法在精度上高于现有算法, 但整体仍有不足之处, 如历史信息聚合时无法识别一直静止的物体 (如停在路边的车), 且主要关注上游自监督任务, 而对下游检测任务改进较小. 因此, 后续工作将在本文基础上探讨如何区分静态背景与静态交通参与者, 并改进目标检测部分性能, 使其达到更高的检测性能.

参考文献

1 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

2 Caesar H, Bankiti V, Lang AH, *et al*. nuScenes: A multimodal dataset for autonomous driving. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2022. 11618–11628. [doi: 10.1109/CVPR42600.2020.01164]

3 Jing LL, Tian YL. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(11): 4037–4058. [doi: 10.1109/TPAMI.2020.2992393]

4 Bao HB, Dong L, Piao SH, *et al*. BEiT: BERT pre-training of image transformers. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022. 2.

5 He KM, Chen XL, Xie SN, *et al*. Masked autoencoders are scalable vision learners. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15979–15988. [doi: 10.1109/CVPR52688.2022.01553]

6 Xie ZD, Zhang Z, Cao Y, *et al*. SimMIM: A simple framework for masked image modeling. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 9643–9653. [doi:

10.1109/CVPR52688.2022.00943]
7 Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2022. 1597–1607.
8 Chen T, Kornblith S, Swersky K, *et al*. Big self-supervised models are strong semi-supervised learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1865.
9 He KM, Fan HQ, Wu YX, *et al*. Momentum contrast for unsupervised visual representation learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9726–9735. [doi: 10.1109/CVPR42600.2020.00975]
10 Grill JB, Strub F, Altché F, *et al*. Bootstrap your own latent a new approach to self-supervised learning. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1786.
11 Chen XL, He KM. Exploring simple Siamese representation learning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15745–15753. [doi: 10.1109/CVPR46437.2021.01549]
12 Cao Y, Xie ZD, Liu B, *et al*. Parametric instance classification for unsupervised visual feature learning. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1310.
13 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.
14 Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.
15 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]
16 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.
17 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6517–6525.
18 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.
19 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934, 2020.
20 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.
21 Duan KW, Bai S, Xie LX, *et al*. CenterNet: Keypoint triplets for object detection. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6568–6577. [doi: 10.1109/ICCV.2019.00667]
22 Tian Z, Shen CH, Chen H, *et al*. FCOS: Fully convolutional one-stage object detection. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9626–9635. [doi: 10.1109/ICCV.2019.00972]
23 You YR, Luo KZ, Chen XY, *et al*. Hindsight is 20/20: Leveraging past traversals to aid 3D perception. arXiv: 2203.11405, 2022.
24 Barnes D, Maddern W, Pascoe G, *et al*. Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments. Proceedings of the 2018 IEEE International Conference on Robotics and Automation. Brisbane: IEEE, 2018. 1894–1900. [doi: 10.1109/ICRA.2018.8460564]
25 You YR, Luo KZ, Phoo CP, *et al*. Learning to detect mobile objects from LiDAR scans without labels. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 1120–1130. [doi: 10.1109/CVPR52688.2022.00120]
26 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]
27 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

(校对责编: 张重毅)