计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(1):219−230 [doi: 10.15888/j.cnki.csa.009362]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向 RGB-D 语义分割的多模态任意旋转自监督学习①

李鸿宇[1,2], 张宜飞[1,2], 杨东宝[1,2]

[1](中国科学院 信息工程研究所, 北京 100085)
[2](中国科学院大学 网络空间安全学院, 北京 100049)
通信作者: 杨东宝, E-mail: yangdongbao@iie.ac.cn

摘　要: 基于 RGB-D 数据的自监督学习受到广泛关注, 然而大多数方法侧重全局级别的表示学习, 会丢失对识别对象至关重要的局部细节信息. 由于 RGB-D 数据中图像和深度具有几何一致性, 因此这可以作为线索来指导 RGB-D 数据的自监督特征表示学习. 在本文中, 我们提出了 ArbRot, 它可以无限制地旋转角度并为代理任务生成多个伪标签用于自监督学习, 而且还建立了全局和局部之间的上下文联系. 本文所提出的 ArbRot 可以与其他对比学习方法联合训练, 构建多模态多代理任务自监督学习框架, 以增强图像和深度视图的特征表示一致性, 从而为 RGB-D 语义分割任务提供有效的初始化. 在 SUN RGB-D 和 NYU Depth Dataset V2 数据集上的实验结果表明, 多模态任意旋转自监督学习得到的特征表示质量均高于基线模型. 开源代码: https://github.com/Physu/ArbRot.

关键词: 自监督学习; 代理任务; 对比学习; RGB-D; 多模态

引用格式: 李鸿宇,张宜飞,杨东宝.面向 RGB-D 语义分割的多模态任意旋转自监督学习.计算机系统应用,2024,33(1):219–230. http://www.c-s-a.org.cn/1003-3254/9362.html

Self-supervised Learning Based on Multi-modal Arbitrary Rotation for RGB-D Semantic Segmentation

LI Hong-Yu[1,2], ZHANG Yi-Fei[1,2], YANG Dong-Bao[1,2]

[1](Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China)
[2](School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Self-supervised learning on RGB-D datasets has attracted extensive attention. However, most methods focus on global-level representation learning, which tends to lose local details that are crucial for recognizing the objects. The geometric consistency between image and depth in RGB-D data can be used as a clue to guide self-supervised feature learning for the RGB-D data. In this study, ArbRot is proposed, which can not only rotate the angle without restriction and generate multiple pseudo-labels for pretext tasks, but also establish the relationship between global and local context. The ArbRot can be jointly trained with contrastive learning methods for establishing a multi-modal, multiple pretext task self-supervised learning framework, so as to enforce feature consistency within image and depth views, thereby providing an effective initialization for RGB-D semantic segmentation. The experimental results on the datasets of SUN RGB-D and NYU Depth Dataset V2 show that the quality of feature representation obtained by multi-modal, arbitrary-orientation rotation self-supervised learning is better than the baseline models.

Key words: self-supervised learning; pretext task; contrastive learning; RGB-D; multi-modal

① 基金项目: 国家自然科学基金面上项目 (62376266); 中国科学院基础前沿科学研究计划从 0 到 1 原始创新项目 (ZDBS-LY-7024)
收稿时间: 2023-06-29; 修改时间: 2023-07-27; 采用时间: 2023-08-18; csa 在线出版时间: 2023-11-24
CNKI 网络首发时间: 2023-11-27

© 中国科学院软件研究所 http://www.c-s-a.org.cn

RGB-D 场景的语义分割[1−3]是计算机视觉和多媒体领域的基本任务, 在增强现实、室内场景理解和机器人环境感知等领域得到了广泛的应用. RGB-D 数据[4]结合了 RGB (红色、绿色、蓝色) 彩色信息和 Depth 深度信息, RGB-D 数据得益于颜色、纹理和几何结构间的互补信息, 可以提供更多信息用于分析. 预训练学习得到的特征提取模型作为下游任务网络的权重初始化可以提高任务的准确率和训练效率, 减少收敛到最佳性能所需的时间. 因此, 选择适当的初始化权重对于 RGB-D 语义分割模型十分重要.

深度学习的训练过程需要大量的标注数据 (如 ImageNet[5]) 以获得满意的性能表现. 多数情况下, 数量众多且高质量的标注数据是难以获得的. 自监督学习利用数据本身的特点来设计代理任务, 可以在没有数据标注的情况下进行自监督训练, 获得特征提取模型用于下游任务权重初始化. 例如图片缺失区域恢复 (图片修复)[6], 将图片切块打乱顺序然后恢复图片 (拼图问题)[7], 对图片加入不同的噪声后去噪修复 (图片去噪)[8], 对图片进行灰度处理后恢复图片颜色 (图片上色)[9]等自监督代理任务. 近年来, 关于 RGB-D 数据自监督代理任务的研究得到广泛关注, 例如通过彩色图和深度图互相恢复[10], 彩色图和深度图经过数据增广后进行对比学习[11]以及通过掩膜自编码器恢复图像[12]等自监督代理任务.

旋转预测是自监督学习中常见的代理任务[13,14], 通过让模型学习识别应用于输入图像的旋转角度, 进而获得有效的特征表示. 现有工作受限于旋转方式, 可供旋转的角度有限, 代理任务难度较低. 然而, 提升代理任务的难度有益于提高自监督学习的性能[15]. 相对于单代理任务自监督学习方法, 多代理任务自监督学习可以提供多样的监督信号, 得到更鲁棒的特征表示, 提高模型的泛化能力, 使其能够更好地适应在实际应用场景中遇到的不同变换和干扰[16]. 然而, 代理任务也存在着设计难度大和生成伪标签不准确等一系列问题. 基于数据特征设计的代理任务需要对数据有着深入的了解, 使得模型可以根据数据自身的特性来进行特征表示学习.

因此, 本文针对旋转预测代理任务中旋转角度受限和代理任务单一的问题, 提出任意角度旋转代理任务, 克服旋转角度的限制, 提升了代理任务的难度同时, 还可以同时产生旋转角度和位置伪标签用于多代理任务自监督学习. 在基于任意角度旋转代理任务上, 我们构建了多模态多代理任务自监督表示学习框架, 辅助自监督学习得到更好的特征表示. 实验结果表明, 与以往方法相比, 本文提出的基于任意旋转代理任务的多模态多代理任务自监督表示学习框架可以提高模型在 RGB-D 数据上的特征表示能力, 在 RGB-D 语义分割下游任务中取得更好的结果.

1 相关工作

1.1 基于数据特征设计的代理任务

目前多数自监督学习方法通过定义一个代理任务来辅助编码器学习有意义的视觉特征. 用于自监督训练的伪标签信息可以根据数据的特性自动生成, 无需引入人工标注. 当训练完成后, 将编码器的权重提取出来作为特征提取模型, 用于下游任务训练时的权重初始化.

基于数据特征设计的代理任务需要对数据有深入了解, 根据数据特点设计生成伪标签来训练模型. 一些工作设计代理任务来识别数据的变换类型, 比如图像旋转角度预测代理任务. 干扰后恢复原始输入[16]. 旋转预测代理任务首先由 RotNet[13]提出, 该代理任务首先将输入图像整体旋转 $90\times N$, $N\in\{0, 1, 2, 3\}$, 即随机选取 $\{0, 90°, 180°, 270°\}$ 中一个角度, 旋转角度类别即作为伪标签. 卷积神经网络为了识别出图片上的旋转角度需要学习图片中物体结构和上下文信息. 训练时, 生成伪标签只需要极少的时间消耗, 可以做到实时训练. 然而这种方法旋转角度类别少, 代理任务较为简单, 因此不利于学习细粒度特征表示. 尽管如此, 旋转代理任务得到了广泛的应用, Hendrycks 等人[17]发现添加旋转代理任务作为辅助后, 可以提升特征提取模型的鲁棒性和显著改善目标检测性能. PIRL (pretext invariant representation learning)[14]发现加入旋转代理任务不会改变图像的语义表征, 而多代理任务联合训练可以改善特征表示的质量. Chen 等人[18]将旋转代理任务引入对抗生成网络中, 辅助编码器获取在训练中容易被遗忘的特征表示. SelfAugment[19]发现旋转代理任务避免了训练时产生特征混杂的问题. 以上工作表明旋转代理任务对于自监督学习有显著的促进意义, 然而旋转代理任务存在 3 个问题: (1) 无法实现任意角度旋转, 否则会出现边缘碰撞 (图片补丁部分区域超出了图像的边框) 的情况; (2) 旋转预测代理任务仅通过全局的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

自监督训练任务来学习场景层面的特征, 忽视了彩色图像和深度信息在全局和局部之间的互补关系, 难以学习到更细粒度特征; (3) 现有旋转数据增广方法比较简单, 无法为特征表示学习提供多样的监督信号.

1.2 对比学习代理任务

对比学习 (contrastive learning) 是基于实例判别的自监督方法, 旨在拉近相似样本的特征距离, 推远不同样本间的特征距离. 对比学习具有出色的扩展性和泛化能力, 可应用于不同领域的任务. 对比学习代理任务在训练时, 首先对输入样本进行数据增广 (如随机裁剪、随机旋转、颜色抖动等) 操作得到正样本对, 同一批其他输入样本作为负样本. 然后将样本输入孪生网络, 将其映射到低维向量空间中, 通过最大化正样本与负样本之间的距离, 最小化正样本之间的距离, 完成特征表示的学习.

近年来有许多基于实例判别的对比学习自监督学习方法被提出. MoCo (momentum contrast)[20]通过孪生网络将不同的数据样本映射到相同的特征空间, 利用实例判别从无标签数据中获得有意义的特征表示. MoCo 使用了梯度和动量两种权重更新策略, 网络能够更好地捕捉到数据的长期特征, 从而得到更加鲁棒和语义丰富的特征表示. SimCLR (simple framework for contrastive learning of visual representations)[21]同样使用孪生网络结构, 表明不同的数据增强组合对于获得高质量的特征表示非常重要. 此外, 在进行对比损失之前引入一个可学习的非线性变换可以提高特征表示的质量. 不同于 MoCo 方法, SimCLR 未使用缓冲区存储向量特征, 因此训练时需要较大的数据批作为输入. BYOL (Bootstrap your own latent)[22]与上述两种对比学习不同之处在于训练时无需负样本, 只采用正样本进行对比学习. 由于训练集中不存在负样本, 如果上下网络分支结构完全相同, 训练就有可能出现“捷径解”问题, “捷径解”指的是对不同的输入, 输出的特征向量完全相同. 为了解决这一问题, BYOL 采用了非对称网络架构, 并将对比学习的实例判别代理任务替换为近似任务, 即孪生网络最终得到的特征要尽可能地相似. BYOL 表明无需负样本, 网络依然可以学习到有效的特征表示, 而且这种方式对数据批处理大小不敏感, 适合实际应用. SimSiam (simple Siamese)[23]在孪生网络架构的基础上, 提出梯度停止操作是防止模型坍塌的关键. 在不使用负样本和动量更新策略情况下, 训练时直接最大化一张图片的两个视图的相似性来进行特征表示学习. SimSiam 使用了两个相同的神经网络, 完全共享网络参数, 并且将一个数据增广后的两个不同样本作为输入. 与之前对比学习方法不同, SimSiam 采用交替停止孪生网络梯度回传, 再交替使用梯度更新策略, 侧面证实了孪生网络架构是对比学习方法成功的关键性因素. 得益于对比学习良好的扩展性, 本文利用任意旋转代理任务结合对比学习构建了多代理任务自监督学习框架.

2 任意旋转多代理任务自监督框架介绍

本文提出的多模态多代理任务自监督学习框架主要包括 3 个组件: (1) 多模态数据增强方法 ArbRot, 负责生成彩色图像、深度图像正样本对和伪标签信息, 详见图 1(a) 下侧. (2) 一个孪生网络作为编码器提取不同层次和尺度的特征信息, 促进特征表示收敛到统一的特征空间. (3) 通过多代理任务联合训练实现自监督学习, 从而提高特征提取模型的鲁棒性, 总体框架如图 2 所示. RGB-D 数据首先经过任意旋转数据增广操作得到一对 RGB-D 数据和相应的旋转角度以及位置两个伪标签, 然后分别将这一对 RGB-D 分别输入到编码器中. 其中图 2 上方编码器和解码器采用梯度更新权重, 编码器的输出会输入到解码器当中, 二者采用 U-Net[24]中应用的跳跃连接完成不同层之间信息的传递, 解码器的输出作为图像恢复头和深度估计头的输入. 此外, 编码器的输出经过平均池化后经过全连接层后输入到旋转预测头、位置预测头和对比学习头中计算损失. 图 2 中下方编码器采用动量更新权重, 编码器的输出经过平均池化后输入全连接层用于对比学习损失计算.

2.1 RGB-D 任意旋转

RotNet 对整图进行 90°倍数的旋转, 即选取{0, 90°, 180°, 270°}中角度进行旋转, 旋转角度作为伪标签用于自监督学习. 我们还设置了 PatchRot 旋转方法, 按照设定的尺寸均匀分割图片得到图片补丁, 此时可以对补丁进行旋转, 可选择的旋转角度和 RotNet 相同, 否则在图片边缘可能会产生边缘碰撞问题 (即补丁的边缘超出了图像边框), 如图 1 红圈位置所示. ArbRot 首先将输入的 RGB-D 数据进行圆形裁剪然后缩放到指定尺寸, 随后对圆形补丁任意旋转并在候选位置中随机选择一个作为贴回位置. 与矩形旋转相比, ArbRot

© 中国科学院软件研究所 http://www.c-s-a.org.cn

可以实现在 RGB-D 数据上进行任意角度旋转, 并且不会产生边缘碰撞, 同时得到位置和旋转角度两个伪标签用于自监督学习. ArbRot 增强样本多样性的同时也提升了代理任务的难度.

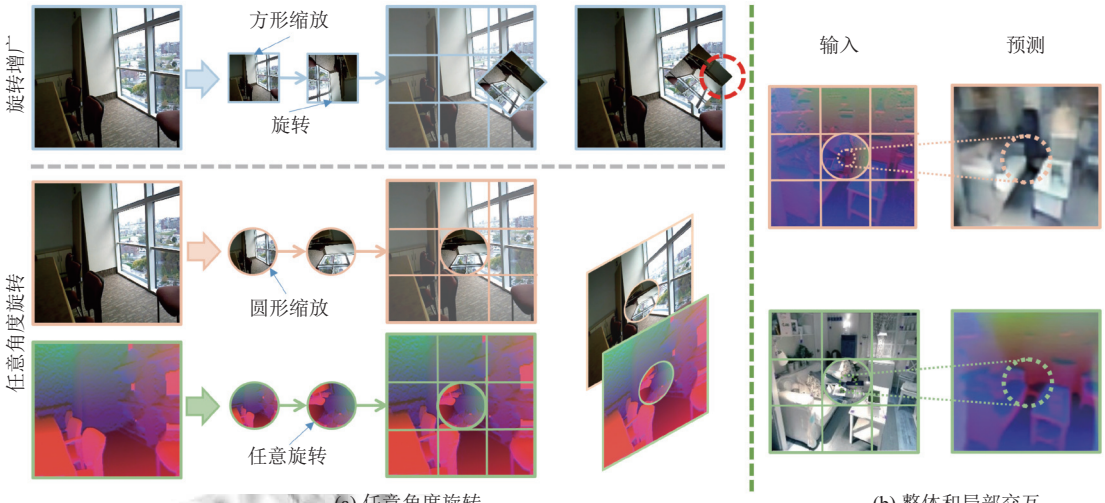


图 1 不同旋转方式对比图

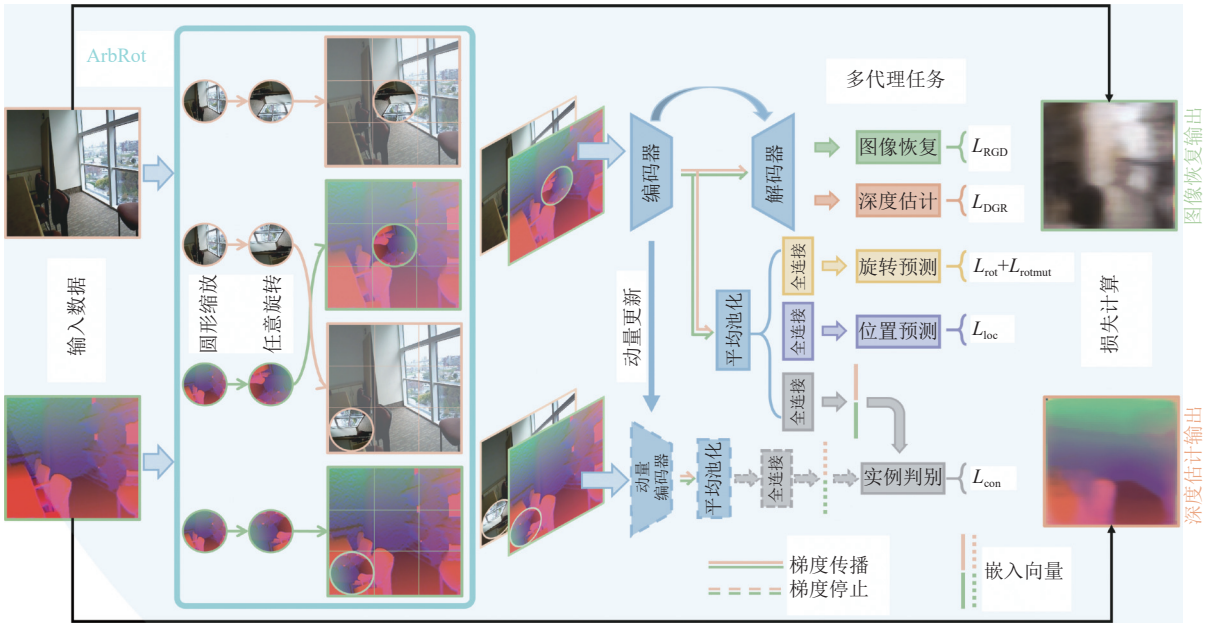


图 2 多模态多代理任务自监督表示学习框架图

2.2 多模态多代理任务头的具体构成

虽然多代理任务训练时不同代理任务之间共享编码器, 但是对应的代理任务头相互独立. 在训练时, 它们会根据伪标签和对应代理任务头的输出计算损失, 完成网络的权重更新. 它们的具体组成如下.

旋转预测头: 旋转预测头负责预测补丁相对于整体图像的旋转角度. 旋转预测代理任务促进了整体和局部、彩色图像和深度信息不同模态数据间的信息交互. 预测结果输出尺寸为$N\times D$, N表示输入样本数量, D表示预测的旋转类别概率和角度残差, 本文采用 3DSSD[25] 中设置, D 设置为 24, 其中前 12 维向量表示旋转类别概率, 后 12 维向量表示预测的角度残差, 具体损失设

© 中国科学院软件研究所 http://www.c-s-a.org.cn

置参见本文第 3.1 节.

位置预测头: 位置预测头负责预测补丁在整体图片中的位置, 加强模型对空间上下文的理解. 首先将编码器的输出特征进行自适应平均池化, 然后将特征调整为指定维度输入到全连接层 (FC) 得到位置预测. 位置预测代理任务相对简单, 可以为模型训练时提供良好的初始权重调整.

图像恢复和深度估计预测头: 图像恢复和深度估计同为像素级别的预测任务, 因此我们同时介绍这两部分预测头的组成. 我们采用 U-Net[24]编码器-解码器架构, 其中包含两个部分: 1) 共享的解码器和编码器, 2) 不共享的代理任务预测头分别对应图像恢复和深度估计代理任务. 图像恢复和深度估计代理任务让模型学习两种模态之间有意义的语义信息和形状线索.

对比学习头: 一个 RGB-D 数据经过任意旋转 (ArbRot) 数据增广方法可以直接生成正样本对用于对比学习, 辅助编码器获得不同模态间一致的特征表示. 本文中使用不同对比学习方法, 除非特别说明, 否则使用官方设置.

3 损失函数的具体构成

3.1 旋转损失

旋转损失由旋转预测损失L_{rot}和多模态旋转损失L_{rotmut}组成. 旋转预测损失L_{rot}由负责旋转角度分类的交叉熵 (CrossEntropy) 损失和负责缩小角度残差的平滑 L1 损失组成. 其中, 交叉熵损失用于计算旋转方向的类别差异, 平滑 L1 损失用于预测旋转角度残差. 这两部分损失共同优化旋转预测任务, 使模型能够准确地预测物体的旋转角度. 受 3DSSD[25]启发, 我们将 360°分为 12 个角度区间 (即每 30°为一个区间), 并以此区间中心分为 [–15°, 15°], 把角度残差标准化到 [–0.5, 0.5] 范围内, 以进行损失计算.

$$L_{\text{rot}} = \frac{1}{N}\sum_{i=1}^{N}(CE(d_c,t_c)+SmoothL1(d_r,t_r)) \tag{1}$$

$$L_{\text{rotmut}} = \frac{1}{N}\sum_{i=1}^{N}\left(\left|Rot(d_c^{\text{RGB}},d_r^{\text{RGB}})-Rot(d_c^{\text{Dep}},d_r^{\text{Dep}})\right|\Big/360\right) \tag{2}$$

其中, d_c和d_r表示旋转类别和残差的预测值, t_c和t_r是目标值. 例如旋转角度 278°, 则$t_c=9$, $t_r=-0.23$. L_{rotmut}负责解决两种模态下旋转角度预测值不一致的问题, 促进不同模态间的信息交互. N表示输入样本数量.

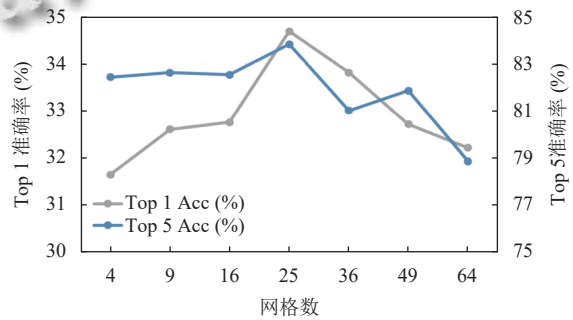
3.2 位置预测损失

我们对整体图像按照网格进行均匀划分, 并对网格进行编号, 选择其中某个网格贴回补丁图片. 如此位置预测转换为一个分类问题, 因此我们采用交叉熵损失来监督物体位置预测. 位置预测损失L_{loc}:

$$L_{\text{loc}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}x_j\log(y_j) \tag{3}$$

其中, x_j表示位置类别标签, y_j表示位置预测的概率. 当只用位置预测代理任务来进行自监督训练时, 我们将不同网格划分设置下得到的特征提取模型在下游 CIFAR10[26]分类任务上进行测试, 结果如图 3(b) 所示. 在 CIFAR10[26]上冻结主干网微调全连接层训练 50 轮时, 发现 5×5 网格划分时分类性能最好.

(a) 不同特征提取模型的余弦相似度



(b) 不同网格数在 CIFAR10 上分类性能

图 3 代理任务间相似度对比图和下游分类任务性能图

3.3 图像恢复和深度估计采用的损失

我们使用感知损失和风格损失来衡量生成的图像与真实图像之间的差异. 感知损失衡量生成图像和真实图像在语义特征上的差异. 风格损失衡量生成图像

© 中国科学院软件研究所 http://www.c-s-a.org.cn

和真实图像在纹理和风格上的差异, 二者保证生成的图像与真实图像保持一致.

我们的 RGB 生成深度 (深度生成 RGB) 的损失函数为:

$$L_{\text{per}}=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{H_iW_iC_i}\left|\Phi_i^{\text{gt}}-\Phi_i^{\text{pred}}\right|_1\right) \tag{4}$$

$$L_{\text{sty}}=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{C_i\times C_i}\left|\frac{1}{H_iW_iC_i}\left(\Phi_i^{\text{gt}}\Phi_i^{\text{gt}^{\text{T}}}-\Phi_i^{\text{pred}}\Phi_i^{\text{pred}^{\text{T}}}\right)\right|_1\right) \tag{5}$$

$$L_{\text{RGD(DGR)}}=\lambda_{\text{per}}L_{\text{per}}+\lambda_{\text{sty}}L_{\text{sty}} \tag{6}$$

其中, Φ^{gt}, Φ^{pred} 表示 RGB 彩色图像 (Depth 深度图像) 真实值和预测值; H_i, W_i, C_i 表示第 i 个输出的高, 宽和通道数, 数值分别为 256, 256, 3. λ_{per} 和 λ_{sty} 表示感知损失和风格损失的权重, 参考文献[27]分别设置为 0.05 和 120. 真实值和预测值可视化结果如图 4 所示.

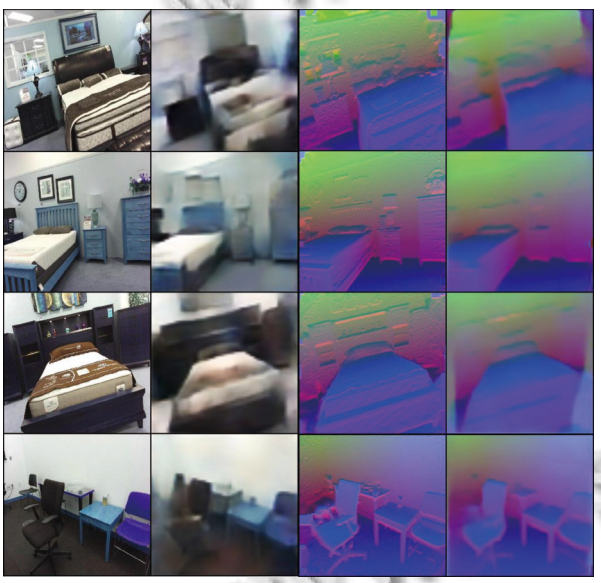


图 4 图像恢复和深度估计的可视化结果

3.4 对比学习损失

我们采用 InfoNCE[20]作为对比学习的损失函数. X 表示训练数据集合, x_0 表示当前给定输入数据, 通过在其上应用 ArbRot 数据增强方法, 我们可以生成与其相似度较高的正样本 x^+, $y^+\sim\text{ArbRot}(x_0)$, 同时输入数据批中其他数据 (或者缓冲队列中的向量) 作为负样本 $x_1^-,\cdots,x_N^-\sim p(X)$. 对比学习 L_{con} 损失函数可以表示如下:

$$L_{\text{con}}=-\frac{1}{N}\log\frac{\exp(x^+\cdot y^+/\tau)}{\sum_i^N\exp(x^+\cdot x_i^-/\tau)} \tag{7}$$

其中, τ 表示温度超参数, 设置为 0.2.

多模态多代理任务自监督的总体训练损失函数 L 可以表示为:

$$L=\lambda_1L_{\text{rotrot}}+\lambda_2L_{\text{mut}}+\lambda_3L_{\text{loc}}+\lambda_4L_{\text{RGD}}+\lambda_5L_{\text{DGR}}+\lambda_6L_{\text{con}} \tag{8}$$

4 实验分析

4.1 数据集

我们在 SUN RGB-D 数据集[28]和 NYU Depth Dataset V2 数据集[29]上进行了实验.

SUN RGB-D 数据集[28]一共包含 10 335 张 RGB-D 图像, 整个数据集包含 146 617 个 2D 多边形和 64 595 个带有准确物体方向的 3D 边界框, 以及每个场景的 3D 房间布局和场景类别, 图片中像素均有类别标签, 共有 37 类别. 我们利用训练集 (5 285 张图像) 和验证集 (5 050 张图像) 进行自监督训练, 未使用任何标注信息.

NYU Depth Dataset V2 数据集[29]由纽约大学创建, 通过微软 Kinect 体感外设采集了各种室内场景视频序列, 其中包含了同一室内场景的彩色信息和深度信息. 其中, 数据集包括 1 449 张标注的 RGB 图片和深度图, 来自于 3 个城市, 464 个场景, 407 024 张无标注图片, 每个对象都有一个类和一个实例号码. 数据集分为 795 张图片用于训练, 654 张用于测试. 我们采用 40 类像素级别标签进行下游任务训练. 上述两个数据集中的深度图使用参考文献[2]中的方法将其编码为 HHA 图像.

4.2 实现细节

我们将每张 RGB 彩色图像和 HHA 深度图像的高度和宽度调整为 256 像素. 在 ArbRot 数据增强方法中, 网格采用 5×5 划分, 还采用了随机裁剪和颜色抖动两种数据增强方法. 我们采用 ResNet-50[30]作为编码器, 预训练时权重采用随机初始化. 解码器使用了 U-Net[24]中解码器设置, 由 3 个上采样卷积模块 (UpConvBlock) 组成. 上采样卷积模块由一个上采样模块和一个卷积模块组成. 上采样模块将高层低分辨率特征图进行上采样, 卷积模块融合来自编码器的低层高分辨率特征图和上采样的高层低分辨特征图. 解码器的输出作为图像恢复和深度估计预测头的输入. 此外, 编码器的输出特征经过平均池化后, 输入 3 个独立的全连接层 (FC), 得到 3 个不同特征向量 (维度分别为 N×24, N×25,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

N×128, N 表示数据批大小) 用于旋转预测, 位置预测和实例判别任务.

我们在 SUN RGB-D 数据集上进行多模态多代理任务自监督训练. 对于优化器, 我们采用随机梯度下降 (SGD) 优化器, 学习率设置为 0.1, 动量设置为 0.9, 权重衰减为 0.000 1, 并使用步长 (step learner) 学习率调节器. 网络训练开始阶段采用热身 (warm-up) 策略缓慢增大学习率到指定大小. 自监督训练 300 轮 (epoch), 在训练到 180 和 270 轮时乘以 0.1 降低学习率, 并将整体损失函数中的超参数λ_1, λ_2, λ_3, λ_4, λ_5, λ_6设置为 0.8、0.8、1.0、0.8、0.2、0.2. 我们在 2 个 NVIDIA 2080 进行预训练, 每个 GPU 的批处理大小为 8.

我们将预训练得到的特征提取模型作为 ShapeConv 和 DeepLabv3+模型的主干网络的权重初始化, 然后进行有监督训练微调下游任务性能. 二者的优化器选择随机梯度下降 (SGD), 学习率设置为 0.007, 动量设置为 0.9, 权重衰减系数为 1E–4, 采用 PolyLR[31] 调节学习率, 一共训练 500 轮. 我们在 2 个 NVIDIA 2080 上进行微调, 每个 GPU 的批处理大小为 4, 具体设置参考文献[2].

4.3 评价指标及基准模型

在 RGB-D 语义分割任务中, 我们采用了与 FCN[32] 相同的评估标准, 即像素准确率 (Pixel Acc)、平均准确率 (Mean Acc)、平均区域交并比 (Mean IoU) 和加权交并比 (f.w. IoU) 这 4 个指标.

表 1 中我们对其他的旋转代理任务进行了测试, 并和我们提出的任意旋转代理任务进行了对比. 可以发现和整图矩形旋转的 RotNet 相比, PatchRot 性能更好. 而 ArbRot 的性能又优于 RotNet 和 PatchRot 两个代理任务. 对于任意旋转代理任务, 我们发现整体补丁优于局部补丁. 其中局部补丁是指补丁所在位置图像直接旋转, 整体补丁是指将整图缩小到补丁尺寸后旋转.

我们把从 SimCLR、SimSiam、BYOL 和 MoCo 预训练得到的 4 个特征提取模型分别在 SUN RGB-D 和 NYU Depth Dataset V2 数据集上使用 ShapeConv 语义分割方法微调, 作为基线性能. 然后, 将 ArbRot 用于上述 4 个对比学习方法中构建多模态多代理任务自监督学习框架在 SUN RGB-D 进行预训练, 将得到的特征提取模型用于 SUN RGB-D 和 NYU Depth Dataset V2 数据集中语义分割下游任务, 结果见表 2 和表 3, 各行上方为基线性能, 下方为我们提出框架的性能. 表 2 中在 SUN RGB-D 数据集上的平均交并比 (mIoU) 结果分别比 SimCLR、SimSiam、BYOL 和 MoCo 这 4 个基线性能提高了 1.75%、6.39%、4.92% 和 6.99%. 在 SUN-RGB-D 数据集上的结果表明了我们方法的有效性. 表 3 中在 NYU Depth Dataset V2 数据集上的平均交并比 (mIoU) 结果分别比 SimCLR、SimSiam、BYOL 和 MoCo 这 4 个基线性能提高了 6.82%、10.38%、10.32% 和 11.22%. 在 NYU Depth Dataset V2 数据集上的结果表明了我们方法在其他数据集上展现出良好的泛化性.

表 1 不同旋转预测代理任务在 NYU Depth Dataset V2 数据集上语义分割性能 (自监督训练 300 轮) (%)

旋转代理任务	旋转对象	矩形	圆形	旋转角度	Pixel Acc	Mean Acc	Mean IoU	f.w. IoU
RotNet	整图	√	—	4°	64.79	42.45	32.32	49.74
PatchRot	局部补丁	√	—	4°	66.01	44.71	34.01	50.97
	整体补丁	√	—	4°	66.13	44.86	34.38	51.03
ArbRot	局部补丁	—	√	任意角度	67.13	45.71	35.22	51.82
	整体补丁	—	√	任意角度	67.18	46.17	36.11	52.18

表 2 使用 ShapeConv 模型在 SUN RGB-D 数据集上语义分割性能 (自监督训练 300 轮) (%)

对比学习方法	负样本	梯度停止	动量更新	Pixel Acc	Mean Acc	Mean IoU	f.w. IoU
SimCLR	√	—	—	74.33	43.06	32.26	60.98
				74.86 **(+0.53)**	45.46 **(+2.40)**	34.01 **(+1.75)**	61.71 **(+0.73)**
SimSiam	—	√	—	74.75	46.96	34.46	61.61
				78.59 **(+3.84)**	54.16 **(+7.20)**	40.85 **(+6.39)**	66.81 **(+5.20)**
BYOL	—	√	√	74.69	45.85	33.56	61.67
				77.89 **(+3.20)**	52.20 **(+6.35)**	38.48 **(+4.92)**	65.35 **(+3.68)**
MoCo	√	√	√	74.65	44.50	33.53	61.49
				78.64 **(+3.99)**	52.86 **(+8.36)**	40.52 **(+6.99)**	66.57 **(+5.08)**

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 3 使用 ShapeConv 模型在 NYU Depth Dataset V2 数据集上语义分割性能 (自监督训练 300 轮) (%)

对比学习方法	负样本	梯度停止	动量更新	Pixel Acc	Mean Acc	Mean IoU	f.w. IoU
SimCLR	√	—	—	61.17 65.84 **(+4.67)**	38.25 45.85 **(+7.60)**	27.90 34.72 **(+6.82)**	45.86 51.30 **(+5.44)**
SimSiam	—	√	—	63.27 70.97 **(+7.70)**	41.26 52.95 **(+11.69)**	30.71 41.09 **(+10.38)**	48.17 56.51 **(+8.34)**
BYOL	—	√	√	62.51 70.38 **(+7.87)**	40.57 52.13 **(+11.56)**	29.97 40.29 **(+10.32)**	46.92 56.05 **(+9.13)**
MoCo	√	√	√	61.97 70.01 **(+8.04)**	39.19 52.29 **(+13.10)**	28.95 40.17 **(+11.22)**	46.72 55.58 **(+8.86)**

5 消融实验设置

为了揭示多模态多代理任务学习范式的潜力, 我们使用 MoCo 作为对比学习的代表在 SUN RGB-D 数据集上进行了多模态多代理任务的消融实验. 优化器采用随机梯度下降 (SGD), ArbRot 中网格采用 5×5 划分. 我们设置每个 GPU 上批处理大小为 8、学习率为 0.1, 自监督训练 100 轮, 在训练到 60 和 90 轮的时候乘以 0.1 减少学习率; 动量系数设置为 0.9, 权值衰减系数设置为 0.000 1. 为了分析组合不同代理任务对下游任务影响, 我们将整体损失函数中的超参数 λ_1, λ_2, λ_3, λ_4, λ_5, λ_6 都设置为 1, 即每个代理任务对模型的贡献相同, 观察不同任务组合对语义分割下游任务性能的影响.

5.1 消融实验及模型分析

为了验证基于任意旋转的多模态多代理任务自监督学习的有效性, 我们研究了以下两个问题.

RQ1: 多代理任务是否可以获得比对比学习单代理任务更好的结果?

通过表 4、表 5 中前 5 行的结果 (表 4、表 5 中各代理任务损失权重均设置为 1, 自监督训练 100 轮), 表明 MoCo+ArbRot 双代理任务在 DeepLabv3+[33]和 ShapeConv[2]两个语义分割方法在像素准确率上结果优于单代理任务, 证明 ArbRot 与对比学习方法进行联合训练可以得到更细粒度的语义特征. 通过表 4、表 5 中第 5–11 行结果, MoCo+loc+ArbRot 三代理任务组合优于双代理任务组合. 通过表 4、表 5 中第 12–15 行结果, MoCo+loc+ArbRot+RGD 四代理任务组合获得的特征提取模型优于三代理任务. 表 4、表 5 最后一行结果表明随着代理任务的增加, 性能总体呈现上升趋势, 但是联合训练过程中不同代理任务之间的干扰也愈发严重.

综上, 多代理任务可以获得比对比学习单代理任务更好的结果, 相对于现有的对比学习单代理任务自监督学习更具实用价值.

RQ2: 多代理任务之间是否存在互相干扰以及如何克服这些干扰?

通过表 4, 表 5 中的实验结果可以发现, 如果各个代理任务产生的损失权重都置为 1, 即 λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, 权重均取值 1.0, 此时随着代理任务的增加, 虽然像素准确率、平均交并比指标有所上升, 但是性能开始不稳定. 说明多代理任务之间存在着相互干扰导致特征提取模型表征能力下降.

表 4 使用 DeepLabv3+模型在 NYU Depth Dataset V2 数据集上比较不同代理任务组合对性能的影响 (%)

代理任务					NYU Depth Dataset V2			
MoCo	loc	rot	RGD	DGR	Pixel Acc	Mean Acc	Mean IoU	f.w. IoU
√	—	—	—	—	60.80	38.87	28.28	45.98
√	√	—	—	—	62.81	41.04	30.44	47.91
√	—	√	—	—	**62.87**	**41.66**	**30.52**	**48.16**
√	—	—	√	—	62.76	40.94	30.11	47.71
√	—	—	—	√	62.09	39.75	29.23	47.06
√	√	√	—	—	**64.38**	**43.46**	**32.12**	**49.41**
√	√	—	√	—	63.69	42.85	31.66	48.97
√	√	—	—	√	63.47	42.63	31.44	48.81
√	—	√	√	—	63.38	41.78	30.98	47.96
√	—	√	—	√	62.33	38.84	28.98	46.51
√	—	—	√	√	63 37	42.07	31.24	48.08
√	√	√	√	—	**64.64**	**44.06**	**32.74**	**49.59**
√	√	√	—	√	64.01	43.32	32.04	48.97
√	√	—	√	√	63.42	42.28	30.93	48.27
√	—	√	√	√	63.09	41.97	30.62	47.86
√	√	√	√	√	64.90	44.31	33.19	49.97

我们统计了各个代理任务的特征提取模型每一层神经元权重之间的平均余弦相似度, 如图 3(a) 所示. 可以得知彩色图像生成深度图像 (RGD) 更接近于深度图像生成彩色图像 (DGR) 代理任务, 而 MoCo 则更接近于 ArbRot 代理任务. 我们发现, 当两个代理任务越相似, 它们被组合在一起时性能越差. 我们猜测相似的任务有相似的收敛空间, 在训练过程中它们将受到其对应的相似代理任务的干扰而陷入局部最优. 此外我们注意到, 不相似的任务组合到一起后将会提高视觉表

© 中国科学院软件研究所 http://www.c-s-a.org.cn

示的质量. 我们将这种提升归因于不相似的代理任务有助于获得不同的收敛空间, 减少表征学习中的干扰.

表 5 使用 ShapeConv 模型在 NYU Depth Dataset V2 数据集上比较不同代理任务组合对性能的影响 (%)

代理任务					NYU Depth Dataset V2			
MoCo	loc	rot	RGD	DGR	Pixel Acc	Mean Acc	Mean IoU	f.w. IoU
√	—	—	—	—	61.97	39.19	28.95	46.72
√	√	—	—	—	63.80	41.10	31.39	49.07
√	—	√	—	—	**64.03**	**42.21**	**31.66**	**48.77**
√	—	—	√	—	62.87	40.36	30.18	47.47
√	—	—	—	√	61.56	38.93	28.73	46.31
√	√	√	—	—	**65.72**	**44.38**	**33.55**	**50.73**
√	√	—	√	—	63.96	42.18	31.56	49.14
√	√	—	—	√	63.21	41.74	30.91	48.63
√	—	√	√	—	64.74	42.72	32.01	49.71
√	—	√	—	√	62.77	39.32	29.47	46.92
√	—	—	√	√	63.42	41.19	30.91	48.63
√	√	√	√	—	**66. 01**	**45.01**	**34.01**	**50.97**
√	√	√	—	√	64. 98	43.22	32.44	50.02
√	√	—	√	√	64.75	43.01	32.08	49.85
√	—	√	√	√	64.19	42.64	31.86	48.93
√	√	√	√	√	65.38	43.14	32.70	49.94

为了克服不同代理任务间的干扰, 我们根据下游任务的性能确定主要和辅助任务, 并将二者的损失权重之和设置为 1. 因此, 我们将 ArbRot 和 RGD 作为主要任务, 将 MoCo 和 DGR 作为辅助任务. loc 任务与其他 4 个代理任务均不相似且训练时损失较早收敛, 因此将其损失权重设置为 1. 如表 6 所示, 整体损失函数中的超参数λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, 分别设置为 0.8、0.8、1.0、0.8、0.2、0.2, 可以在下游任务中获得最佳性能.

从表 4, 表 5 中发现, 当λ_1, λ_2, λ_3, λ_4, λ_5, λ_6均设置为 1 时, 原有任务组合上额外加入 DGR 代理任务, 下游任务性能就会下降. 我们采用λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, 分别设置为 0.8、0.8、1.0、0.8、0.2、0.2 时测试对下游语义分割任务的影响. 从表 7 中的结果可以看出降低损失权重后的 DGR 代理任务可以辅助提升下游任务性能, 如果去除 DGR 代理任务会导致性能下降.

表 6 使用 ShapeConv 模型在 NYU Depth Dataset V2 数据集上比较不同自监督代理任务权重对性能的影响 (自监督训练 300 轮) (%)

代理任务权重						NYU Depth Dataset V2			
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	Pixel Acc	Mean Acc	Mean IoU	f.w. IoU
1.0	1.0	1.0	1.0	1.0	1.0	65.38	43.14	34.01	50.97
0.9	0.9	1.0	0.9	0.1	0.1	68.80	47.54	36.76	53.24
0.8	0.8	1.0	0.8	0.2	0.2	**70.01**	**52.29**	**40.17**	**55.58**
0.7	0.7	1.0	0.7	0.3	0.3	69.46	49.44	38.69	54.81
0.6	0.6	1.0	0.6	0.4	0.4	69.34	49.60	38.75	54.61

表 7 使用 ShapeConv 模型在 NYU Depth Dataset V2 数据集上比较深度恢复彩色图片任务对性能的影响 (%)

代理任务					训练轮数	NYU Depth Dataset V2			
MoCo	loc	rot	RGD	DGR	(epoch)	Pixel Acc	Mean Acc	Mean IoU	f.w. IoU
					100	68.65	48.96	37.98	53.81
√	√	√	√	—	200	69.22	49.79	38.75	54.68
					300	69.72	49.98	39.07	55.10
					100	69.60	50.15	38.73	54.88
√	√	√	√	√	200	69.73	50.73	39.45	55.04
					300	70.01	52.29	40.17	55.58

5.2 不同增广方式对性能的影响

图片恢复和深度估计的预测质量很难直接观察, 为了更直观地观察其他不同数据增广方式对自监督训练的影响, 我们使用 ArbRot 数据增强技术在 SUN RGB-D 数据集上构建了一个新数据集, 称为 ArbRot RGB-D. 为了提高 ArbRot RGB-D 数据集的多样性, 位置有 9 个候选位置, 旋转角度任意. ArbRot RGB-D 由一个训练集 (5 285 张图像对) 和一个验证集 (5 050 张图像对) 组成, 我们使用训练集来训练模型, 使用验证集来验证性能. 旋转统计结果见表 8, 我们尽量覆盖所有旋转角度. 图 5 呈现了 ArbRot RGB-D 数据集中的部分图例, 交替为彩色图和深度图.

表 8 数据集 ArbRot RGB-D 中旋转方向的分布表

旋转类别	0	1	2	3	4	5	6	7	8	9	10	11
角度区间	0–29°	30°–59°	60°–89°	90°–119°	120°–149°	150°–179°	180°–209°	210°–239°	240°–269°	270°–299°	300°–329°	330°–359°
数量	902	88	872	841	844	828	883	885	892	859	812	829

我们测试了其他 5 种数据增强方法对 RGB-D 表示学习的影响, 分别是随机剪裁 (random crop, RC)、颜色扰动 (color jitter, CJ)、随机灰度 (random gray, RG)、高斯模糊 (Gaussian blur, GB) 和随机翻转 (random flip, RF). 我们结合表 9 发现, 当使用 RC 和 CJ 数据增广时, 旋转分类的性能可以达到较好的效果; 当同时使用 RC, CJ, RG, GB 这 4 种数据增强方法时, 位置分类的性能得到了进一步提升. 旋转预测代理任务比位置分类代理任务难度更高, 因此我们使用 RC 和 CJ 数据增强方法用于多模态多代理任务的自监督

© 中国科学院软件研究所 http://www.c-s-a.org.cn

学习.

5.3 多模态多代理任务自监督训练损失变化

多模态多代理任务自监督训练过程中, 不同损失的变化见图 6. 可以发现结合 SimCLR 方法时总体损失下降最慢. 而当结合 SimSiam、BYOL 和 MoCo 的方法时, 总体损失下降较快. 位置预测损失最快收敛; 图片恢复、深度估计和旋转预测对应损失抖动下降, 表明三者为难代理任务.

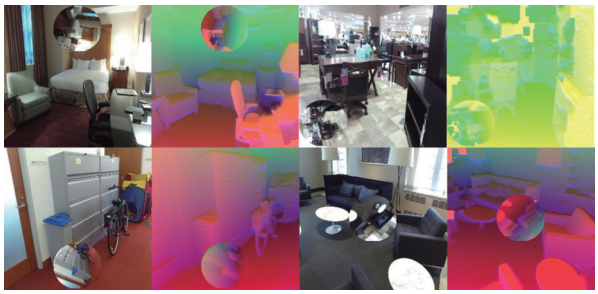
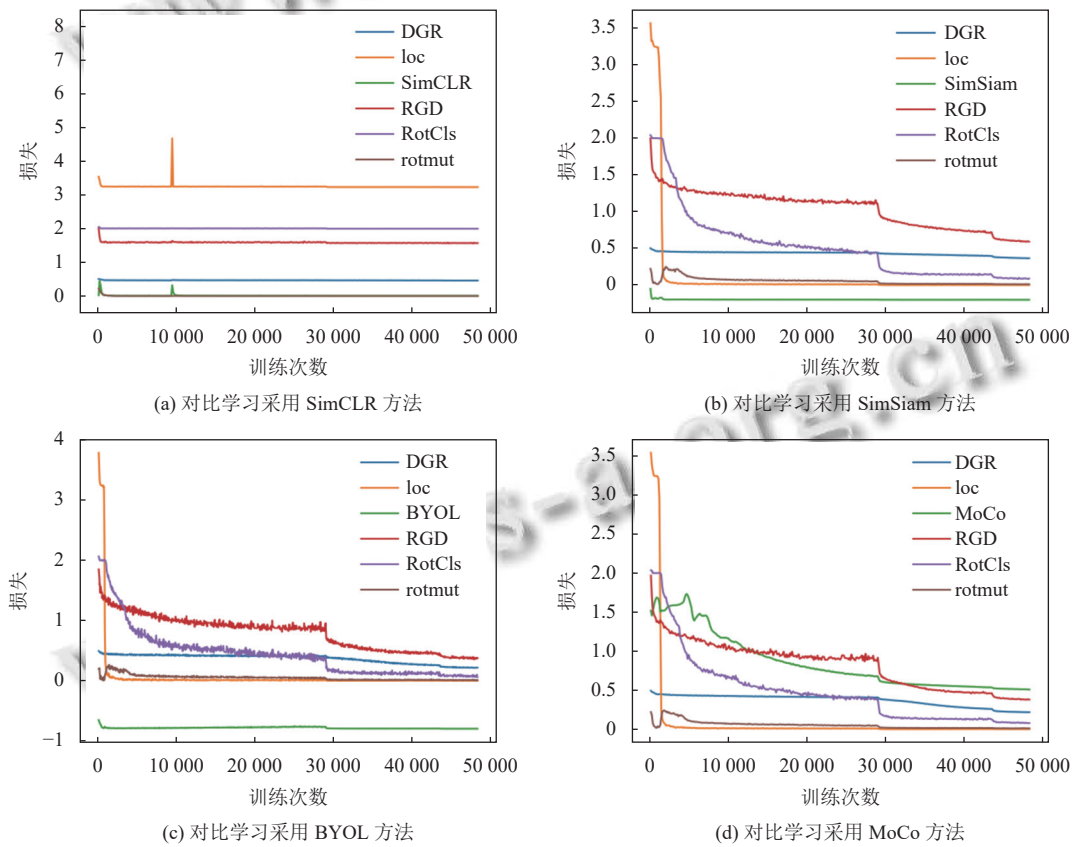


图 5 ArbRot RGB-D 数据集图例

表 9 数据集 ArbRot RGB-D 中使用不同数据增广方法后旋转和位置预测准确率 (%)

数据增广方法					RGB图像上准确率			HHA图像上准确率			总体准确率		
RC	CJ	RG	GB	RF	类别	角度残差	位置	类别	角度残差	位置	类别	角度残差	位置
√	—	—	—	—	56.75	0.60	82.65	80.22	0.60	97.88	68.89	0.60	90.27
√	√	—	—	—	64.46	0.56	94.08	80.75	0.58	98.57	72.60	0.57	96.33
√	√	√	—	—	52.81	0.58	86.63	77.58	0.56	94.30	65.20	0.57	90.47
√	√	√	√	—	56.91	0.56	97.70	78.89	0.54	98.44	67.90	0.55	98.07
√	√	√	√	√	60.65	0.58	88.28	80.59	0.58	97.96	70.62	0.58	93.12



(a) 对比学习采用 SimCLR 方法

(b) 对比学习采用 SimSiam 方法

(c) 对比学习采用 BYOL 方法

(d) 对比学习采用 MoCo 方法

图 6 多模态多代理任务和对比学习结合后训练过程中损失的变化

6 结论

本文提出了一种新颖的 RGB-D 数据增强方法, 可以在训练中通过任意旋转补丁图片创建彩色图像和深度图像数据对, 增广后的数据可以促进对 RGB-D 数据中整体和局部关系的理解, 有助于提升特征提取模型的表征能力. 我们设计了一个多模态多代理任务自监督学习框架, 用于完成 RGB-D 多模态表示学习, 并通过下游语义分割任务证明了有效性. 广泛的实验表明,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

我们提出的基于任意旋转代理任务的多模态多代理自监督表示学习框架, 可以得到通用有效的特征表示, 提升下游任务的性能. 我们提出的多模态多代理任务进行自监督表示学习不需要昂贵且费时的人工标注信息, 而且可以融合不同代理任务的优势, 具有较强的实用性.

参考文献

1 Lopes A, Souza R, Pedrini H. A survey on RGB-D datasets. Computer Vision and Image Understanding, 2022, 222: 103489. [doi: 10.1016/j.cviu.2022.103489]

2 Cao JM, Leng HC, Lischinski D, *et al*. ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 7068–7077. [doi: 10.1109/ICCV48922.2021.00700]

3 李梦怡, 朱定局. 基于全卷积网络的图像语义分割方法综述. 计算机系统应用, 2021, 30(9): 41–52. [doi: 10.15888/j.cnki.csa.008078]

4 Zhao XQ, Zhang LH, Pang YW, *et al*. A single stream network for robust and real-time RGB-D salient object detection. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 646–662. [doi: 10.1007/978-3-030-58542-6_39]

5 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

6 Pathak D, Krähenbuhl P, Donahue J, *et al*. Context encoders: Feature learning by inpainting. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2536–2544. [doi: 10.1109/CVPR.2016.278]

7 Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 69–84. [doi: 10.1007/978-3-319-46466-4_5]

8 Vincent P, Larochelle H, Bengio Y, *et al*. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th International Conference on Machine Learning. Helsinki: ACM, 2008. 1096–1103. [doi: 10.1145/1390156.1390294]

9 Larsson G, Maire M, Shakhnarovich G. Colorization as a proxy task for visual understanding. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 840–849. [doi: 10.1109/CVPR.2017.96]

10 Zhao XQ, Pang YW, Zhang LH, *et al*. Self-supervised pretraining for RGB-D salient object detection. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 3463–3471. [doi: 10.1609/aaai.v36i3.20257]

11 Chen YJ, Nießner M, Dai A. 4DContrast: Contrastive learning with dynamic correspondences for 3D scene understanding. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 543–560. [doi: 10.1007/978-3-031-19824-3_32]

12 Yang JG, Guo S, Wu GS, *et al*. CoMAE: Single model hybrid pre-training on small-scale RGB-D datasets. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 3145–3154. [doi: 10.1609/aaai.v37i3.25419]

13 Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. Proceedings of the 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.

14 Misra I, van der Maaten L. Self-supervised learning of pretext-invariant representations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6706–6716. [doi: 10.1109/CVPR42600.2020.00674]

15 Chen PG, Liu S, Jia JY. Jigsaw clustering for unsupervised visual representation learning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 11521–11530. [doi: 10.1109/CVPR46437.2021.01136]

16 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]

17 Hendrycks D, Mazeika M, Kadavath S, *et al*. Using self-supervised learning can improve model robustness and uncertainty. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1403.

18 Chen T, Zhai XH, Ritter M, *et al*. Self-supervised GANs via auxiliary rotation loss. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12146–12155. [doi: 10.1109/CVPR.2019.01243]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

19 Reed CJ, Metzger S, Srinivas A, *et al*. SelfAugment: Automatic augmentation policies for self-supervised learning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2673–2682. [doi: 10.1109/CVPR46437.2021.00270]

20 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 (CVPR). Seattle: IEEE, 2020. 9726–9735. [doi: 10.1109/CVPR42600.2020.00975]

21 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 (ICML). PMLR, 2020. 1597–1607.

22 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.

23 Chen XL, He KM. Exploring simple siamese representation learning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 15745–15753. [doi: 10.1109/CVPR46437.2021.01549]

24 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI). Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

25 Yang ZT, Sun YN, Liu S, *et al*. 3DSSD: Point-based 3D single stage object detector. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11037–11045. [doi: 10.1109/CVPR42600.2020.01105]

26 Krizhevsky A. Learning multiple layers of features from tiny images [Master's Thesis]. Toronto: University of Toronto, 2009.

27 Li JY, Wang N, Zhang LF, *et al*. Recurrent feature reasoning for image inpainting. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 7757–7765. [doi: 10.1109/CVPR42600.2020.00778]

28 Song SR, Lichtenberg SP, Xiao JX. SUN RGB-D: A RGB-D scene understanding benchmark suite. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 567–576. [doi: 10.1109/CVPR.2015.7298655]

29 Silberman N, Hoiem D, Kohli P, *et al*. Indoor segmentation and support inference from RGB-D images. Proceedings of the 12th European Conference on Computer Vision (ECCV). Florence: Springer, 2012. 746–760. [doi: 10.1007/978-3-642-33715-4_54]

30 He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016. 770–778.

31 Mishra P, Sarawadekar K. Polynomial learning rate policy with warm restart for deep neural network. Proceedings of the 2019 IEEE Region 10 Conference (TENCON). Kochi: IEEE, 2019. 2087–2092. [doi: 10.1109/TENCON.2019.8929465]

32 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 3431–3440. [doi: 10.1109/CVPR.2015.7298965]

33 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.

(校对责编: 孙君艳)

© 中国科学院软件研究所 http://www.c-s-a.org.cn