计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(6):192−201 [doi: 10.15888/j.cnki.csa.008539]

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

基于变分对抗与强化学习的行人重识别①

陈　莹[1,2], 夏士雄[1,2], 赵佳琦[1,2], 周　勇[1,2], 姚　睿[1,2], 朱东郡[1,2]

[1](中国矿业大学 计算机科学与技术学院, 徐州 221116)
[2](矿山数字化教育部工程研究中心, 徐州 221116)
通信作者: 夏士雄, E-mail: shixiongxia.cumt@outlook.com

摘　要: 行人重识别技术在实际应用中易受行人姿态变化的干扰, 由于行人姿态的变化不仅丢失部分行人信息, 而且还会引起大于身份差异的外观变化, 导致现有工作难以学到鲁棒的行人特征. 为了解决上述问题, 本文提出一种基于变分对抗与强化学习的生成式对抗网络 (RL-VGAN) 用于多姿态行人重识别任务. 该方法的核心思想是在不受姿态变化干扰的情况下通过外观编码器和姿态编码器将行人属性分解为外观特征和姿态特征, 用以学习鲁棒的身份视觉特征. 首先, 设计的变分生成网络利用 Kullback-Leibler 散度损失促进外观编码器推断与身份信息相关的连续隐变量. 其次, 为了使生成式对抗网络逐步收敛到稳定状态, 采用强化学习策略平衡变分生成网络和判别网络的性能. 此外, 针对基于姿态引导图像生成任务, 提出一种新的 Inception Score 损失用于规范变分生成网络生成图像质量的过程. 实验结果证明, 所提出的 RL-VGAN 方法在多个基准数据集上优于其他方法.

关键词: 行人的重识别; 图像生成; 强化学习; 生成对抗网络; 变分学习

引用格式: 陈莹,夏士雄,赵佳琦,周勇,姚睿,朱东郡.基于变分对抗与强化学习的行人重识别.计算机系统应用,2022,31(6):192–201. http://www.c-s-a.org.cn/1003-3254/8539.html

Person Re-identification Based on Variational Adversarial and Reinforcement Learning

CHEN Ying[1,2], XIA Shi-Xiong[1,2], ZHAO Jia-Qi[1,2], ZHOU Yong[1,2], YAO Rui[1,2], ZHU Dong-Jun[1,2]

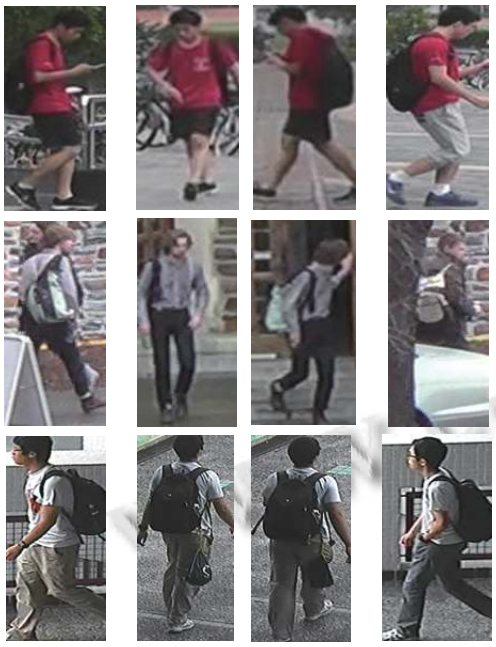
[1](School of Computer Science & Technology, China University of Mining and Technology, Xuzhou 221116, China)
[2](Engineering Research Center of Mine Digitization, Xuzhou 221116, China)

Abstract: Person re-identification (ReID) technology is easily disturbed by the pose variation which causes loss of person information and appearance changes exceeding identity differences. It is a challenging task for existing ReID methods to learn robust person features. For such problems, we propose the generative adversarial network (GAN) based on variational inference and reinforcement learning (RL-VGAN). The core idea of the proposed method is to disentangle person attributes into appearance features and pose features via appearance and pose encoders, which learns robust identity-related features without interference from pose changes. Firstly, the designed variational generative network leverage the Kullback-Leibler divergence loss to strengthen the appearance encoder for inferring identity-related continuous latent variables. Secondly, we use reinforcement learning to balance the performance of the generative and discriminative networks during the training process. Thirdly, for the pose-guided generative task, a novel Inception Score loss is designed for evaluating the image synthesis quality in the variational generative network. Experimental results demonstrate the superiority of the proposed RL-VGAN over other methods for the benchmark datasets.

Key words: person re-identification (ReID); image synthesis; reinforcement learning; generative adversarial network (GAN); variational learning

① 基金项目: 国家自然科学基金 (61806206, 62172417); 江苏省自然科学基金 (BK20180639, BK20201346); 江苏省六大人才高峰计划 (2015-DZXX-010, 2018-XYDXX-044)
收稿时间: 2021-09-11; 修改时间: 2021-10-14; 采用时间: 2021-10-24; csa 在线出版时间: 2022-02-21

行人重识别技术 (person re-identification, ReID)[1] 是在行人检测的基础上利用计算机视觉方法判断图像或者视频序列中是否存在特定行人的技术, 被认为是图像检索的子问题. 行人重识别技术与行人检测技术相结合, 可广泛应用于智能视频监控、智能商业、智能安防等领域. 在实际的视频监控环境中, 由于目标尺寸变化、姿态变化、非刚体目标形变等目标自身变化的多样性和光照变化、背景复杂、相似行人干扰、遮挡等应用环境的复杂性, 使得鲁棒、高效的行人重识别是一个极具挑战性的课题, 也是当前国内外的研究热点. 其中, 摄像机视角不同和多姿态行人是导致 ReID 任务识别精度低的主要原因. “多姿态” (例如正身与侧身匹配) 是指当目标发生运动时引起身体几何形变或者角度变化, 从而导致不同姿态下同一行人图像在像素级别的差别大于不同行人在相同姿态下的图像, 如图 1 所示. 针对上述问题, ReID 方法的核心在于如何设计鲁棒的行人视觉特征和如何得到最优的行人图像特征相似性度量.



(a) 锚节点 (b) 同一行人 (c) 不同行人

图 1 行人重识别任务中“多姿态”样本示例

卷积神经网络 (convolutional neural networks, CNNs) 作为深度学习的一个重要组成部分, 它可以从大规模数据集中自动学习鲁棒的行人特征, 基于深度学习的 ReID 方法能够自动学习较好的视觉特征和最优的相似性度量, 因此基于深度学习的行人重识别技术得到迅速发展[2]. 人体姿态的变化会引起识别漂移或者失败, 其原因是当人体发生形变或者角度变化时, 行人的表观特征也会发生变化, 与初始跟踪时的目标有较大外观差异. 行人姿态多变仍然是 ReID 方法提取有效行人特征的一大挑战, 现有深度学习领域主要有 3 类方法针对该问题: 行人图像对齐[3–6], 局部特征学习[7–11] 和行人姿态转换[12–16].

行人图像对齐方法解决的是由于姿态或者视角变化以及不可靠的关键点检测引起的身体部件缺失和图像背景冗余问题, 通过将非对准图像数据进行人体结构对齐来学习图像对的相似度. 局部特征学习方法针对姿势变化引起的人体不对准问题, 采用关键点定位技术生成多个区域, 从而学习易于判别行人身份的局部特征. 行人姿态转换方法利用生成式对抗网络生成身份一致的规范姿态图像达到学习与身份相关特征的目的. 尽管这些方法获得了较好的 ReID 性能, 但行人图像对齐和局部特征学习方法在识别阶段需要辅助的姿态信息, 这限制了 ReID 方法的泛化能力. 尤其是基于行人姿态转换的 ReID 方法, 它们忽略了生成任务对识别精度的影响.

针对行人重识别数据集的姿态多样性带来的挑战, 在不进行行人对齐或学习基于人类区域表示的情况下, 本文提出一种基于变分对抗与强化学习 (RL-VGAN) 的行人重识别方法来提取仅与身份相关的视觉特征. 一方面提升网络生成多样性样本的能力, 另一方面提升行人重识别方法对相似样本干扰的鲁棒性. 具体而言, RL-VGAN 在孪生网络结构中嵌入设计的变分生成式对抗网络 (variational generative network, VG-Net), VG-Net 中变分生成网络由外观编码器和图像解码器组成, 图像解码器将外观编码器编码的外观特征和姿态编码器编码的姿态特征解码为新的行人图像; 姿态判别器用以判断生成的行人图像是否与原始的目标姿态一致. 除了 VG-Net 外, 还包括一个身份验证分类器实现行人身份的判断. 特别地, 变分生成网络将行人图像分解为两个基本特征: 与内在身份信息相关的外观特征和可变化的姿态特征 (包括位置、体型、形状等). 大量定性和定量实验证明 RL-VGAN 方法在基准数据集上取得显著效果. 本文的主要贡献包括以下 3 点.

(1) 设计了一个新的变分生成网络将行人特征解耦为外观特征和姿态特征, 有效地缓解姿态变化带来识别精度低的问题. 特别地, 通过采用 Kullback-Leibler

(KL) 散度损失促进编码网络学习潜在空间变量和真实图像之间的关系, 保证编码的空间变量包含更多与行人身份相关的信息.

(2) 采用强化学习策略能够处理变分生成式对抗网络在方向传播中不可微分的问题, 通过限制生成网络迭代的梯度调整判别网络的参数, 保证生成网络和判别网络的协调工作.

(3) 针对基于姿态引导图像生成任务生成图像质量差的问题, 设计新的 inception score (IS) 损失, IS 是评估 GAN 生成图像真实性和多样性的指标, 因此提出新的 IS 损失使变分生成网络生成具有真实性和多样性的行人图像.

本文的其余部分组织如下: 第 1 节讨论了行人重识别方法的相关工作; 第 2 节详细地介绍基于变分对抗与强化学习的行人重识别方法; 第 3 节描述了实验细节和分析了实验结果; 第 4 节概括了本文的结论以及提出未来研究工作的方向.

1 相关工作

行人 ReID 技术通常包含 3 个环节: 特征提取、相似度量和特征匹配. 首先利用行人特征表示方法提取行人图像的视觉特征; 然后对提取到的行人图像视觉特征进行训练, 学习合适的相似性度量方法; 最后将待检索的行人图像视觉特征与其他行人图像视觉特征进行相似度排序, 找到与其相似度高的行人图像. ReID 方法的核心在于如何设计鲁棒的行人视觉特征和如何得到最优的行人图像特征相似性度量. 由于目标在不同的角度和距离拍摄下, 其形状、姿态和相对大小都有变化, 行人姿态多变仍然是 ReID 方法提取有效行人特征的一大挑战, 现有深度学习领域主要有 3 类方法针对该问题: 行人图像对齐[3–6], 局部特征学习[7–11] 和行人姿态转换[12–16].

基于行人图像对齐的行人重识别方法通过把人体分解成几块区域后获取每个区域的特征表示, 计算两幅图像对应区域之间相似度和作为它们的匹配得分. 王金等人[3] 利用行人图像的图像块集合, 提取每个图像块特征表示获取行人图像的局部信息, 对局部信息进行聚类处理建立两幅行人图像块之间的对应关系以获得姿态对齐后的图像块序列. 基于深度学习的部件表示 (deeply-learned part-aligned representations, DPR)[4] 方法针对人体空间分布不一致问题, 采用注意力机制提取一个更具区分性的三维特征向量, 其通道对应人体部位, 在不借助人体部件标注的情况下采用最小化三元损失训练网络模型. 这些行人图像对齐方法要么简单地把人体分为几个部分, 要么通过姿态估计器估计人体骨架信息来实现对齐, 而行人对齐网络 (pedestrian alignment network, PAN)[5] 采用深度学习方法来矫正行人姿态, 学习一个二维的变换把行人对齐好后再做识别, 该方法包含基本网络分支和对齐网络分支这两个 CNN 分类网络和一个放射估计网络. 基本分类网络由 ResNet-50 作为骨干网络, 执行识别预测任务; 对齐网络定位行人关节点以便放射估计网络学习一个能够对齐人体结构的二维变换. Zheng 等人[6] 提出位姿不变嵌入 (pose invariant embedding, PIE) 作为行人描述符, 首先利用姿态估计和仿射变换产生将行人与标准姿势对齐的 PoseBox 结构; 其次设计 PoseBox Fusion 网络融合输入图像、PoseBox 和姿态估计误差, 在姿态估计失败时提供了一种后退机制.

上述方法利用人体结构来增强识别能力, 通过人体部件对齐表示来处理身体部件不对齐导致的局部距离过大问题. 而基于局部特征学习的行人重识别方法通过区分人体区域精准地识别行人, 因为人体具有高度的结构[17]. Chen 等人[7] 提出了可以提取人体整体和区域特征的集成模型, 该集成模型包括提取整体特征的卷积神经网络和提取区域特征的双向高斯混合模型. 为了提高模型的泛化性, 在特征匹配时采用距离归一化提取的特征. 另一个解决此类问题的有效方法是将长短期记忆网络嵌入到孪生网络中[8], 利用上下文信息以序列的方式对人体部件进行处理, 提高局部特征的判别能力实现识别行人的任务. Spindle Net[9] 是 ReID 任务中第 1 个考虑人体结构信息的方法, 它利用 14 个定位的人体关节来检测感兴趣区域, 产生 7 个身体区域: 头-肩、上体和下体宏观区域以及双腿、双臂微观区域与 Spindle Net 相似, 姿态驱动的深卷积方法 (pose-driven deep convolutional, PDC)[10] 也采用了同时学习全局和局部信息的方式, 但将 14 个关键点分成 6 个区域. 而全局局部对齐描述符方法 (global local alignment descriptor, GLAD)[11] 在提取人体关键点后将人体分为头部、上半身和下半身 3 部分, 采用 4 个子网络组成的 CNN 对全局区域和局部区域进行特征表示学习, 结合全身输入到网络中进行特征融合.

尽管这些方法获得了较好的 ReID 性能, 但由于需

要辅助姿态信息增加了计算复杂度. 近年来, 许多学者对 Goodfellow 等人[18] 首次提出的生成式对抗网络 (generative adversarial network, GAN) 产生了兴趣, 一些工作致力于研发基于 GAN 的 ReID 任务. Zheng 等人[19] 利用深度卷积生成式对抗网络 (deep convolutional GAN, DCGAN) 生成无类标样本, 这是利用 GAN 完成 ReID 任务的第一个工作. 同时也有很多 ReID 方法利用 GAN 来指导姿态转换的行人图像生成. Ge 等人提出 FD-GAN (feature distilling GAN)[12] 仅学习和身份信息有关的视觉特征, 去除冗余的姿态特征表示. 在网络学到行人视觉特征后, 在测试阶段不需要辅助的姿态信息, 因此减少了计算成本. 为了解决在跨摄像机下对姿态多变训练数据的差异特征和不变特征的鲁棒性学习, Ho 等人[13] 提出一种端到端的稀疏时态学习框架用以解决姿态时序变化问题. Qian 等人[14] 提出一种基于姿态归一化图像生成的方法 (pose-normalization GAN, PN-GAN), 该方法可以生成身份一致和姿态可控的行人图像. 而基于姿态生成的方法 (pose transferrable GAN, PT-GAN)[15] 是一个实现转移行人姿态的模型, 将 MARS 数据集中的多姿态行人图像迁移到目标数据集以扩充训练样本, 设计引导子网络模型使生成的新姿态图像更好地适应 ReID 任务.

2 基于变分对抗与强化学习的行人重识别方法

本文提出的 RL-VGAN 模型以姿态引导图像生成的思想解决 ReID 易受姿态变化影响和相似行人干扰的问题. 整体的网络模型结构如图 2 所示. RL-VGAN 模型采用孪生网络结构, 该结构的每个分支嵌入由变分生成网络G和姿态判别器D_p组成的变分生成式对抗网络. 以孪生网络一个分支的训练过程为例, 条件行人图像x_i被G中的外观编码器E_a编码成外观特征f_a, 目标姿态图像p_k被姿态编码器E_p编码为姿态特征f_p, 图像解码器D根据外观特征f_a、姿态特征f_p和随机噪声n拼接的特征z生成拥有x_i外观以及姿态p_k的行人图像x_i^k. 接下来, 姿态判别器D_p通过判别样本姿态的真实性来规范图像解码器D生成姿态变化样本的能力. 此外, 身份验证分类器V监督外观编码器E_a学习仅与身份相关的视觉特征.

2.1 变分生成网络

给定序列$(X,Y)=(\{x_1,\cdots,x_N\},\{y_1,\cdots,y_M\})$, x_i表示有M个类别和N张图像数据集中的一张行人图像, y_j表示x_i的身份标签. 为了生成真实的行人图像, 本节设计变分生成网络学习与图像相关的连续隐变量分布以便进行采样和插值. 一方面利用变分推理保留条件行人图像的细节信息, 另一方面采用最近邻损失保证生成的图像在外观和纹理上与条件行人图像一致.

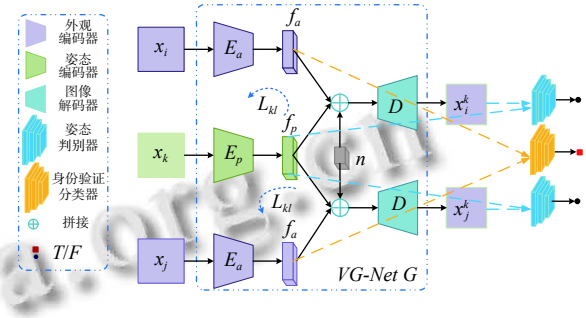


图 2 RL-VGAN 网络结构示意图

在孪生网络分支中, 变分生成网络G由外观编码器E_a和图像解码器D组成. 外观编码器E_a采用的是 ResNet-50 网络, x_i经过外观编码器E_a处理后得到 2 048 维外观特征f_a. 为了实现姿态迁移任务还需要借助姿态编码器E_p编码得到的 128 维姿态特征f_p, E_p通过以下卷积神经网络序列来实施: C_{64}^2 - C_{128}^2 - C_{256}^2 - C_{512}^2 - C_{512}^2, 其中, C_f^s表示由f个卷积核和s个滤波器组成的卷积层. 除了第 1 层卷积以外, 其他 4 层卷积操作都是在激活函数 ReLU 之后以及批归一化 (batch normalization, BN) 之前执行. 对外观特征f_a、姿态特征f_p以及 256 维服从高斯分布的噪声向量n进行拼接操作得到 2 432 维的向量z, 作为图像解码器D的输入以生成姿态迁移后的行人图像x_i^k, 图像解码器D采用反卷积序列D_{512}^2 - D_{256}^2 - D_{128}^2 - D_{64}^2 - D_3^2结构.

在 RL-VGAN 模型的行人样本生成网络中, 利用 KL 散度学习由外观编码器E_a处理得到的连续隐变量f_a与条件行人图像x_i之间的潜在关系来保证外观编码器E_a学到x_i鲁棒的外观特征. 即, 利用近似后验分布$q(f)$学习生成图像分布$p(x|\cdot)$, 如式 (1) 所示.

$$\begin{aligned}\ln p(x) &= \ln p(x,f) - \ln p(f|x)\\ &= \ln\frac{p(x,f)}{q(f)} - \ln\frac{p(f|x)}{q(f)}\\ &= \ln\int_f p(x,f)q(f)df - \ln\int_f q(f)q(f)df\\ &\quad - \ln\int_f \frac{p(f|x)}{q(f)}q(f)df\\ &= L(q) + KL(q\|p)\end{aligned} \tag{1}$$

其中, $L(q)=\int_f \ln p(x,f)q(f)df-\int_f \ln q(f)q(f)df$ 是根据 Jensen 散度不等式得到的证据下界 (evidence lower bound, ELBO), $KL(q\|p)$表示$-\int_f \ln \frac{p(x|f)}{q(f)}q(f)df$, 其值大于 0. 因此式 (1) 可以简化为:

$$\begin{aligned}\ln p(x)&=\ln p(x,f)-\ln p(f|x)\geqslant L(q)\\&=E_{q(f)}\ln\frac{p(x,f)}{q(f|x)}\\&=E_{q(f)}\ln\frac{p(x|f)p(f)}{q(f|x)}\end{aligned}\tag{2}$$

根据以上分析, G模型损失函数定义如下:

$$\begin{aligned}L_{\mathrm{VG}}&=L(x,\theta,\phi)\\&=-KL(q_\phi(f|x)\|p_\theta(f))+E_{q_\phi(f|x)}[\log p_\theta(x|f)]\end{aligned}\tag{3}$$

其中, ϕ是外观编码器E_a学习的参数, θ是图像解码器D需要学习的参数. G根据拼接后的向量z生成新的行人图像x_i^k要与其对应的真实图像x_k相似. 因此重构损失为:

$$L_{RE}=\sum\left\|x_k-x_i^k\right\|_1\tag{4}$$

为了保证生成图像x_i^k在外观和纹理上与给定图像x_k一致, 使用最近邻损失代替重构损失L_{RE}保留细节信息:

$$L_{\mathrm{NN}}=\sum_{p\in\bar{x}}\min_{q\in N(p)}\left\|C_{x_k}(q)-C_{x_i^k}(p)\right\|_1\tag{5}$$

其中, p表示生成图像x_i^k的像素点, $N(p)$是对应的局部邻域点, $C_{x_i^k}(p)$表示对应p特征图$C_{x_i^k}(\cdot)$的信道值. 在保证生成清晰的行人图像情况下, 还要关注变分生成网络生成的图像是否具有多样性, 因此本文设计了基于评估图像生成质量的 IS[20] 损失函数L_{IS}:

$$L_{\mathrm{IS}}=-\exp\left(E_{x'\sim p_G}KL\left(p\left(v\left|x'\right.\right)\|p(v)\right)\right)\tag{6}$$

其中, $x'\sim p_G$表示x'从变分生成网络G生成样本中进行采样的图像, v是图像x'的类别向量, $p(v|x')$的每个维度值表示图像属于某类的概率, $p(v)$是图像中所包含类别的分布.

外观编码器E_a仅仅学习与身份相关的行人特征, 那么当同一行人的两张不同图像和一个目标姿态被输入到整个孪生网络中, 两个图像解码器从理论上来说生成的行人具有相同的姿态. 定义相同姿态损失L_{sp}[12] 用来减小两个分支网络生成图像姿态之间的差异. 相同姿态损失表述为:

$$L_{sp}=E\left\|x_i^k-x_j^k\right\|\tag{7}$$

借助姿态编码器, 孪生网络中两个图像解码器生成的行人图像姿态一致, 保证一个分支中的外观编码器可以学习仅用身份相关与姿态无关的特征.

2.2 姿态判别器

变分生成式对抗网络通过变分推理和对抗学习生成较为真实的图像, 编码网络通过隐变量和真实图像之间的 KL 损失保持了外观特征的一致性. 在对抗性学习阶段, RL-VGAN 模型将变分生成网络和姿态判别器嵌入到孪生网络模型中, 通过生成样本对抗学习提升 RL-VGAN 模型学习身份特征以及生成相似样本的能力. GAN 的基本思想来源于极小极大博弈, 变分生成网络试图通过生成更自然的图像“欺骗”判别器以获得高匹配置信度, 姿态判别器D_p用来判别变分生成网络G生成的行人图像是否能完成姿态迁移的任务.

将外观特征f_a、姿态特征f_p和服从正态分布的随机噪声n统一到相同空间维度z, 加入噪声n目的是提高模型鲁棒性. 在基于变分生成对抗网络模型中, 图像解码器D根据z生成具有p_k姿态和x_i外观的新图像x_i^k, 姿态判别器D_p判别生成的图像x_i^k与相同分支输入图像x_k的姿态特征是否保持一致, 保证D在姿态转移上的生成能力. D_p的损失函数如式 (8) 所示.

$$L_p^G=\frac{1}{2}\sum_{m=1}^{2}E\left[\log D_p(x_k[m])-\log\left(1-\log D_p\left(x_i^k[m]\right)\right)\right]\tag{8}$$

其中, m表示孪生网络的分支数.

2.3 基于强化学习的变分生成式对抗网络算法

深度强化学习 (reinforcement learning, RL) 将深度学习的强大感知能力及表征能力与强化学习的决策能力相结合, 通过最大化奖励函数的学习方式使学习器从环境中获取行为. 具体而言就是通过一系列动作策略与环境交互, 学习器产生新的参数, 再利用新的参数去修改自身的动作策略, 经过数次迭代后, 学习器就会学习到完成任务所需要的动作策略. 在基于姿态指导行人图像生成任务中, 采用强化学习的方法训练变分生成网络G和姿态判别器D_p中的参数, 对它们的参数进行调整保证两个网络协调工作来学习行人的几何特征. 基于强化学习的变分生成式对抗网络 (RL-VGAN) 模型如图 3 所示.

在 RL-VGAN 网络模型中, 变分生成网络G作为学习器在更新网络参数生成新的样本过程中, 与姿态判

别器D_p环境进行交互, 产生新的状态S, S表示在当前姿态判别器D_p的状态下是否需要对G进行状态更新. G生成图像的质量通过强化学习决策产生动作a影响D_p. 同时环境给出反馈即由标量奖励信号r组成, 通过达到最大奖赏值来提高生成网络生成图像的能力, 以及通过学习器和环境不断交互来更新网络. G将生成的图像送入D_p计算奖励信号Q_r, 根据得到的奖励信号进行策略梯度下降优化模型. 采用$D_p(\cdot)$作为奖励函数一方面促使变分生成网络G和姿态判别器D_p协同工作, 另一方面保证生成的图像具有目标姿态特征. 奖励信号Q_r定义如下:

$$Q_r = D_p(G(x_i, p_k), p_k) \tag{9}$$

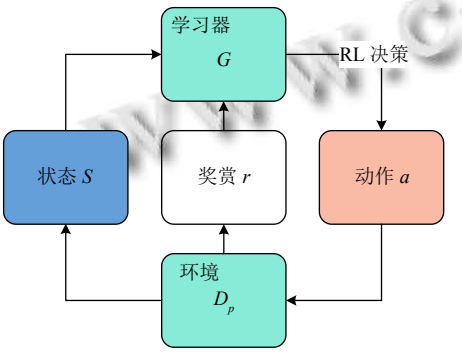


图 3 强化变分生成式对抗网络结构示意图

一个分支网络的 D_p 试图最小化以下损失函数:

$$L_p^1 = Q_r \times D_p(p_k, x_k) \tag{10}$$

其中, x_k是具有姿态p_k的行人图像, 因此整个网络的姿态判别损失为$L_p = L_p^G + \frac{1}{2}\sum_{m=1}^{2} L_p^m$. 算法 1 详细地给出了上述过程的算法描述.

算法 1. 基于强化学习的变分生成式对抗网络算法流程

输入: 学习器G, 环境D_p, 行人样本x和姿态数据p, 起始状态S_0=$G(x,p)$
输出: 学习器G生成的图像x'

1 **for** $t_{epoches}$<$maxepoches$ **do**
2 使用变分生成网络G根据姿态和行人图像生成一张行人图像x'
3 根据x'质量来产生是否更新G的状态S以及动作a
4 使用式 (9) 计算奖励信号Q_r
5 根据奖励信号Q_r判断当前是否对G执行更新网络参数的决策
6 根据Q_r进行策略梯度下降优化模型
7 **end for**

2.4 训练

为了完成识别行人身份任务, 需要借助身份验证分类器V进行行人身份的识别, V根据两个分支外观编码器E_a编码的特征识别输入的图像是否属于同一个行人, 因此验证分类损失 Lve 可以由式 (11) 表示:

$$L_{ve} = -\theta\log\left(\left(x_i - x_j\right)^{\mathrm{T}}\left(x_i - x_j\right)\right) - (1-\theta)\left(1 - \log\left(\left(x_i - x_j\right)^{\mathrm{T}}\left(x_i - x_j\right)\right)\right) \tag{11}$$

若$y_i = y_j$, 则$\theta = 1$; 否则$\theta = 1$. 将上述定义的损失函数合并到最终损失函数来优化所提出的 RL-VGAN 模型, 包括变分生成网络G、姿态编码器E_p、身份验证分类器V和姿态判别器D_p, 如下所示:

$$L_{\mathrm{RL-SVGAN}} = \lambda_{\mathrm{VG}}L_{\mathrm{VG}} + \lambda_{\mathrm{NN}}L_{\mathrm{NN}} + \lambda_{\mathrm{IS}}L_{\mathrm{IS}} + \lambda_{sp}L_{sp} + \lambda_{ve}L_{ve} + \lambda_p L_p \tag{12}$$

其中, $\lambda_{(\cdot)}$是权重因子, 它们的值分别设置为$\lambda_{\mathrm{VG}} = \lambda_{\mathrm{IS}} = \lambda_p = 0.5$, $\lambda_{\mathrm{NN}} = 10$, $\lambda_{VE} = \lambda_{sp} = 1$.

3 实验结果与分析

本节对所提出的 RL-VGAN 模型在 3 个基准数据集上进行实验验证, 证明 RL-VGAN 模型在 ReID 任务中的优越性. 首先对本文使用的数据集和评价指标进行介绍; 其次针对图像生成任务, 与基于姿态指导行人图像生成方法进行对比; 最后对 RL-VGAN 模型与先进的行人重识别方法在姿态变化问题上进行比较.

3.1 数据集和评价指标

基于卷积神经网络的行人重识算法依赖于大规模的数据集, 本文在大型数据集 CUHK03[21], Market-1501[22] 和 DukeMTMC[23] 上进行 ReID 算法验证, 通过 3 个指标: IS[20], structural similarity (SSIM)[24] 和 Frechet inception distance (FID)[25] 评价图像生成质量, 采用平均准确度 (mean average precision, mAP) 和累计匹配特征 (cumulative match characteristics, CMC) 曲线评估 ReID 算法的性能.

采用的数据集详细信息如表 1. CUHK03 数据集是由香港中文大学从 2 个摄像头上采集的, 包含 1 476 个行人的 14 097 张图像, 每个行人平均有 9.6 张训练数据. 由 1 367 个行人作为训练集和 100 个行人作为测试集组成, 且提供人工标注的行人检测框和机器检测的行人检测框. Market-1501 数据集的采集地点是清华大学校园, 使用 6 个摄像头采集了 1 501 个行人的 32 668 张图像, 其中训练集有 751 个行人和 12 936 张图像, 平均每人有 17.2 张训练数据; 测试集包含 750 个行人的 19 732 张图像, 平均每人拥有 26.3 张测试数据. Duke-

MTMC 数据集是在杜克大学由 8 个摄像头采集, 该数据集由 16 522 张行人图像的训练集和 17 661 张图像的测试集组成. 训练集中有 702 个行人, 平均每人有 23.5 张训练数据; 测试数据集中有 702 个行人, 平均每人有 25.2 张测试数据, 该数据集提供了行人属性 (性别/长短袖/是否背包等) 的标注信息.

表 1 行人重识别图像数据集信息

数据集	公布时间	行人总数	训练集行人数量/身份	测试集行人数量/身份	摄像头
CUHK03	2014	1 467	13 132/1367	965/100	2
Market-1501	2015	1 501	12 936/751	19281/750	6
DukeMTMC	2016	1 404	16 522/702	17 661/702	8

由于各种概率标准, 评估不同模型生成图像的质量是一项艰巨的任务. 使用 3 个标准: 可辨别性, 多样性和真实性来量化 FD-GAN, RL-VGAN(w/IS) (w/IS 表示 RL-VGAN 模型在 FD-GAN 的基础上仅用 IS 损失) 和 RL-VGAN 生成模型. IS 度量标准表示生成图像的质量和多样性之间的合理相关性, 这也是 IS 广泛用于度量生成图像的原因. SSIM 作为感知度量, 经常用来衡量由于数据压缩或数据传输中丢失而导致的图像质量恶化程度. FID 在判别生成图像真实性方面表现良好, 因此它被认为是对带有标记数据集样本质量评估的标准. FID 值越低表示两个样本分布越近, 生成的图像越接近真实图像, 而 IS 和 SSIM 值越高表示生成的图像质量越好.

现有的 ReID 算法采用 CMC 曲线评估算法中分类器的性能, 即匹配给定目标行人图像在大小为r的行人图像库中出现的概率. CMC 曲线将行人匹配结果的高低进行排序, 通过*rank-r*的形式给出, 即查找r次即可找到目标行人的概率. CMC 曲线能够检验 ReID 算法的查准率, 此外还要考虑算法的查全率, 因此采用 mAP 对算法的性能进行评估. mAP 是对 ReID 算法中准确率和召回率的综合考量, 其计算方式是对每个检索目标求 AP (average precision) 并取平均. 将准确率和召回率作为横纵坐标时, AP 的值是曲线下的面积.

3.2 实现细节

与面向多姿态行人重识别的变分对抗与强化学习网络模型和传统的 ReID 模型相比, 模型的任务更复杂, 故采用多阶段的学习方法来训练本文提出的 RL-VGAN 模型, 实现多个任务的协同学习: 一方面实现高质量样本生成, 另一方面提升行人重识别方法的泛化性能. 使用 PyTorch 环境实现代码编写, 采用一张 Geforce RTX 2080Ti 卡训练所提方法. 在训练过程中, 3 个基准数据集的图像大小设置为256×128, 与 FD-GAN[12] 一样, 整个网络的训练分为 3 个阶段. 第 1 阶段利用损失函数L_{ve}在数据集上训练变分生成网络中的外观编码器E_a和身份验证分类器V, 采用随机梯度下降法 (stochastic gradient descent, SGD)[26] 优化两个神经网络, 动量因子大小为 0.9, 初始学习率设为 0.01. 第 1 阶段*batch_size*设为 128, 共训练 100 个迭代次数. 第 2 阶段是针对生成任务, 在固定外观编码器E_a和身份验证分类器V网络参数的情况下训练图像解码器D和姿态判别器D_p, 即式 (12) 中$\lambda_{ve}=0$. 图像解码器D采用 Adam 优化器[27] ($\beta_1=0.5$, $\beta_2=0.999$), 姿态判别器D_p采用 SGD 进行优化, 其中β_1和β_2是矩估计的指数衰减率, 两个网络的初始学习率分别是10^{-3}、10^{-2}, 第 2 阶段的*batch_size*设为 16, 共训练 100 个迭代次数. 第 3 阶段, 整个行人重识别网络以端到端的方式联合微调进行模型参数的学习, *batch_size*设为 16, 共训练 50 个迭代次数.

3.3 实验结果分析

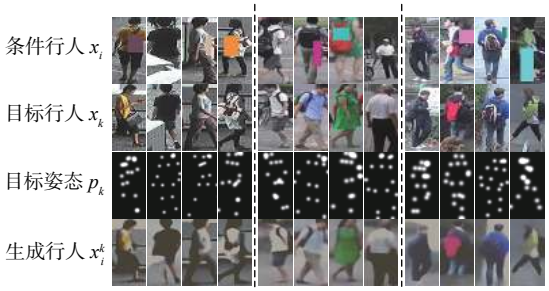
为了证明在本小节中, 我们首先在 3 个基准数据集上, 展示所提方法生成图像的视觉效果, 其次使用 IS, SSIM 和 FID 三种评价指标评估 RL-VGAN 方法生成图像的效果. 最后采用 mAP 和 rank-1 准确率对比 RL-VGAN 方法和其他行人重识别方法.

3.3.1 基于姿态指导行人图像生成结果

图 4 展示了 RL-VGAN 生成图像示例, 从上到下依次为条件行人图像、目标行人图像、目标姿态图像和生成行人图像. RL-VGAN 方法在大多数情况下能够生成真实和多样的图像, 由于数据集中图像存在遮挡以及清晰度低的问题, 因此生成的图像中存在一些噪点, 但整体上比较好的保留了原图像的细节信息.

为了定量地分析方法的有效性, 选用 IS、SSIM 和 FID 作为分析和评估本文方法与基准方法的客观评价指标, 如表 2 所示. 其中, RL-VGAN(w/IS) 表示 RL-VGAN 只采用 IS 损失. 与基线 FD-GAN 相比, 在 CUHK03 数据集上, RL-VGAN(w/IS) 分别在 IS 和

SSIM 评估指标提高了 3.86%、3.45%, 在 FID 指标上下降了 4.77%. 表明 IS 损失能够促进生成网络很好地保留更多外观信息. 而且, RL-VGAN 得到的 IS 准确率相比于 RL-VGAN(w/IS), 分别提高了 9.83%、6.81% 和 1.21%. 其原因在于结合强化学习的生成式对抗网络有效地规范了生成网络生成图像的过程, 从而进一步提高行人图像的姿态转移能力. 针对本文提出的 IS 损失, 我们评估了其在不同数据集上的收敛性, 如图 5 所示. 我们可以看出 IS 损失收敛值约为 0.02.

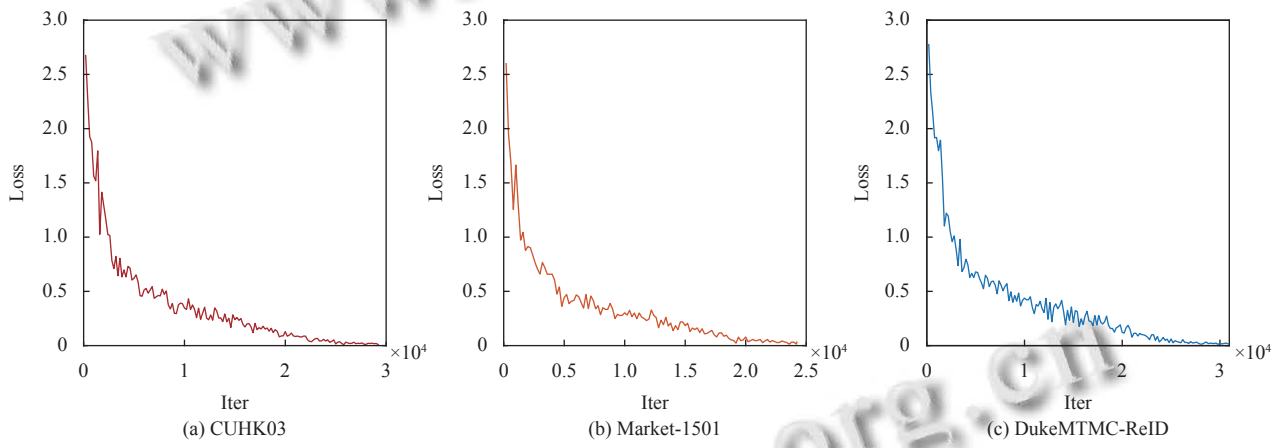


(a) CUHK03 (b) Market-1501 (c) DukeMTMC-ReID

图 4 在 3 个数据集上的生成图像示例

表 2 3 个基准数据集上生成图像的 IS、SSIM 和 FID 值

方法	CUHK03			Market-1501			DukeMTMC-ReID		
	IS↑	SSIM↑	FID↓	IS↑	SSIM↑	FID↓	IS↑	SSIM↑	FID↓
FD-GAN[12]	2.125	0.261	376.733	2.310	0.317	328.744	**2.425**	0.337	337.034
RL-VGAN (w/ IS)	2.207	0.270	358.746	2.334	0.326	**317.167**	2.391	0.270	353.346
RL-VGAN	**2.424**	**0.281**	**348.008**	**2.493**	**0.335**	324.399	2.420	**0.349**	**321.154**



(a) CUHK03 (b) Market-1501 (c) DukeMTMC-ReID

图 5 训练阶段, IS 损失随着迭代次数在 3 个数据集上的变化说明

3.3.2 与现有行人重识别方法的结果比较

为了公平起见, 我们选择的 ReID 对比方法是解决 ReID 任务中行人姿态变化导致识别精度差的问题, 包括基于行人图像对齐的 ReID 方法[5] 和基于行人姿态转换的 ReID 方法[12–16], 如表 3 所示.

表 3 中“*”表示本文复现结果, CMC 包括 rank-1 正确率, 即预测的标签取最后概率向量里面最大的作为预测结果, 若预测结果中概率最大的分类正确则预测正确, 否则预测错误. 值得注意的是, 采用不同的 GPU 卡和不同数量的卡都会严重影响实验结果, 比如 FD-GAN 结果与原论文相比下降严重, mAP 在 CUHK03、Market1501 和 DukeMTMC-ReID 分别下降 2.85%、3.86% 和 12.25%. 因为 GPU 卡的好坏会影响浮点运算, 以及 *batch_size* 大小. 实验数据表明, 在数据集 CUHK03 和 Market1501 上, 本文提出的方法表现均优于其他行人重识别方法. 与基准方法 FD-GAN 相比, RL-VGAN 分别提高了 1.35%、0.67% 和 8.66% (mAP 指标), 0.76%、0.11% 和 3.44% (rank-1 指标). 在 DukeMTMC 数据集上, 所提方法取得了与 GLAD 方法相当的结果. 实验结果表明, 本文提出的方法不仅可以有效地生成高质量的行人样本, 而且还可以缓解行人姿态变化带来的干扰.

4 结论与展望

本文构建了基于变分对抗与强化学习 (RL-VGAN) 的行人重识别模型, 在变分生成式对抗网络中, 利用

变分推理促进生成网络生成相似行人图像的同时学习鲁棒的身份信息. 此外, 提出一种新的 IS 损失提升变分生成网络生成图像的质量, 从而解决行人重识别系统易受相似行人干扰以及行人姿态变化的问题. 由于采用交替迭代方式会导致生成式对抗网络训练过程不稳定, 因此本文采用强化学习策略促进变分生成网络和判别网络收敛到稳定状态. 本文提出的 RL-VGAN 将姿态指导行人图像生成任务与行人重识别任务相结合, 在 3 个基准数据集上进行的大量实验证明, RL-VGAN 不仅能够生成高质量的行人图像还能够有效地完成 ReID 的任务. 基于变分对抗与强化学习的行人重识别方法具有极高的准确性, 但该网络模型容易存在网络参数过拟合的问题. 针对该问题, 将进一步研究基于多目标优化的生成式对抗网络参数学习和结构修剪方法, 提升生成式对抗网络学习的稳定性和泛化性能.

表 3 RL-VGAN 与其他方法在 3 个基准数据集下的 mAP 和 rank-1 准确率 (%)

方法	CUHK03		Market-1501		DukeMTMC	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
PN-GAN[14]	—	79.8	72.6	89.4	53.2	73.6
PT-GAN[15]	45.1	42.0	68.9	87.7	56.9	78.5
*FD-GAN[12]	88.7	90.3	74.7	89.9	56.6	75.6
PAN[5]	35.0	36.9	63.4	82.8	51.5	71.6
SGAN[16]	86.8	83.5	64.3	84.0	—	—
*ST-ReID[13]	79.3	83.1	73.2	89.5	**62.9**	77.8
RL-VGAN	**89.9**	**91.0**	**75.2**	**90.0**	61.5	78.2

参考文献

1 陈丹, 李永忠, 于沛泽, 等. 跨模态行人重识别研究与展望. 计算机系统应用, 2020, 29(10): 20–28. [doi: 10.15888/j.cnki.csa.007621]

2 冯霞, 杜佳浩, 段仪浓, 等. 基于深度学习的行人重识别研究综述. 计算机应用研究, 2020, 37(11): 3220–3226, 3240.

3 王金, 刘洁, 高常鑫, 等. 基于姿态对齐的行人重识别方法. 控制理论与应用, 2017, 34(6): 837–842. [doi: 10.7641/CTA.2017.60647]

4 Zhao LM, Li X, Zhuang YT, *et al*. Deeply-learned part-aligned representations for person re-identification. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 3239–3248.

5 Zheng ZD, Zheng L, Yang Y. Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(10): 3037–3045. [doi: 10.1109/TCSVT.2018.2873599]

6 Zheng L, Huang YJ, Lu HC, *et al*. Pose-invariant embedding for deep person re-identification. IEEE Transactions on Image Processing, 2019, 28(9): 4500–4509. [doi: 10.1109/TIP.2019.2910414]

7 Chen SC, Lee YG, Hwang JN, *et al*. An ensemble of invariant features for person re-identification. 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP). Xiamen: IEEE, 2015. 1–6.

8 Varior RR, Shuai B, Lu JW, *et al*. A siamese long short-term memory architecture for human re-identification. European Conference on Computer Vision (ECCV). Amsterdam: Springer, 2016. 135–153.

9 Zhao HY, Tian MQ, Sun SY, *et al*. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 1077–1085.

10 Su C, Li JN, Zhang SL, *et al*. Pose-driven deep convolutional model for person re-identification. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 3980–3989.

11 Wei LH, Zhang SL, Yao HT, *et al*. Glad: Global-local-alignment descriptor for scalable person re-identification. IEEE Transactions on Multimedia, 2019, 21(4): 986–999. [doi: 10.1109/TMM.2018.2870522]

12 Ge YX, Li ZW, Zhao HY, *et al*. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: NeurIPS, 2018. 1230–1241.

13 Ho HI, Shim M, Wee D. Learning from dances: Pose-invariant re-identification for multi-person tracking. ICASSP 2020 –2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 2113–2117.

14 Qian XL, Fu YW, Xiang T, *et al*. Pose-normalized image generation for person re-identification. European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 661–678.

15 Liu JX, Ni BB, Yan YC, *et al*. Pose transferrable person re-identification. Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 4099–4108.

16 刘仁春, 孟朝晖. 基于孪生对抗 SGAN 的行人重识别研究. 电子测量技术, 2019, 42(15): 155–160, 166.

17 Yao HT, Zhang SL, Hong RC, *et al*. Deep representation learning with part loss for person re-identification. IEEE Transactions on Image Processing, 2019, 28(6): 2860–2871. [doi: 10.1109/TIP.2019.2891888]

18 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: NIPS, 2014. 2672–2680.

19 Zheng ZD, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 3774–3782.

20 Salimans T, Goodfellow I, Zaremba W, *et al*. Improved techniques for training GANs. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: NIPS, 2016. 2234–2242.

21 Li W, Zhao R, Xiao T, *et al*. Deepreid: Deep filter pairing neural network for person re-identification. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus: IEEE, 2014. 152–159.

22 Zheng L, Shen LY, Tian L, *et al*. Scalable person re-identification: A benchmark. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1116–1124.

23 Ristani E, Solera F, Zou R, *et al*. Performance measures and a data set for multi-target, multi-camera tracking. European Conference on Computer Vision (ECCV). Amsterdam: Springer, 2016. 17–35.

24 Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

25 Heusel M, Ramsauer H, Unterthiner T, *et al*. GANs trained by a two time-scale update rule converge to a local nash equilibrium. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: NIPS, 2017. 6629–6640.

26 Yazan E, Talu MF. Comparison of the stochastic gradient descent based optimization techniques. 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). Malatya: IEEE, 2017. 1–5.

27 Kingma DP, Ba J. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations (ICLR). San Diego: ICLR, 2015. 1–15.