重庆大学本科学生毕业设计（论文）

基于生成对抗网络的多姿态行人图像合成算法

学　　生：李光睿

学　　号：20141802

指导教师：葛永新

校外指导老师：杨易

专　　业：软件工程

重庆大学大数据与软件学院

2018 年 6 月

Graduation Design(Thesis) of Chongqing University

Generative Adversarial Networks based Multi-Pose Person Image Synthesize

Undergraduate: Li Guangrui

Supervisor: Ge Yongxin

External Supervisor: Yang Yi

Major: Software Engineering

School of Big Data & Software Engineering

Chongqing University

June 2018

摘　　要

这篇文章提出了一种新颖的基于姿态引导的行人图片生成网络，基于该网络，通过输入原图片和姿态信息，我们能够生成任意姿态的行人图片。该网络通过经典的基于编码器-解码器网络以及该网络与鉴别器间的对抗学习，使得我们生成器网络有着惊人的学习与模拟能力，从而使得大规模的数据增广成为可能。

不同于当前的只能生成图像质量低且训练不稳定的传统的生成模型，本文提出的方法通过创新的目标函数，极大地稳定了生成器与判别器间的对抗学习，是两者的训练进度一致，从而提升生成图片质量；为了解决解空间过大导致的生成图像质量差的问题，通过在个体、低层信息、高层信息上添加对应的监督，进一步缩小模型的解空间，并强化部分所需要强调的特征，从而生成理想的保留了行人个体信息与个体间差异以及高低层信息的图像，进一步提升了生成图像的质量。

关键词：深度神经网络，生成对抗网络，数据增强，行人重识别

ABSTRACT

This paper proposed a novel pose guided person generation network, which could synthesize person images in arbitrary poses, based on the combination of original image and a target pose as input. Based on the classical Encoder-Decoder architecture and the adversarial training between the generator and discriminator, the person generation network shows its surprising potential in mimic and learn the feature of existing person image, which makes it possible for data augmentation in person re-id task.

Different from existing Generative adversarial networks that generate images with low quality, the proposed method use novel objectives to significantly stabilize the adversarial train between the generator and discriminator, which improve the quality of synthesized image; to address the poor quality of synthesized image due to the too wide solution space. through the supervision added on the identity information, lower and higher information, the proposed method greatly narrow the solution space and augment the information or feature which are needed, so generate the multi-pose person image while preserve the identity and inter-identity information, lower and higher information respectively.

Key words： Deep Neural Network, Generative Adversarial Network, Person re-Identification, image to image translation

目录

1 绪论

本文将着重于对基于深度神经网络的行人图像合成研究中的生成对抗网络方法进行进一步的发掘，尝试通过设计更加稳定、具备更强泛化性能的生成对抗网络，进一步发掘生成对抗网络潜力，提升合成的行人图片的质量。本章将对行人重识别问题及图像合成的研究背景和国内外研究现状进行详细介绍，并简要说明本文主要工作。

1.1 研究背景和意义

如图 1.1，行人重识别技术是指在由多个摄像头组成的视频监控系统中，当在指定了某个摄像头下出现目标行人时，行人重识别技术能够将该行人在该监控系统下任意摄像头下任意时间出现的历史查询。该任务的出现主要是由于：①公共安全的强烈需求；②大规模监控网络的铺设使之具备必要的前提条件，同时也使得通过人力去高效、精确地定位和查询目标行人或者跨摄像头跟踪目标行人的代价变得十分昂贵。

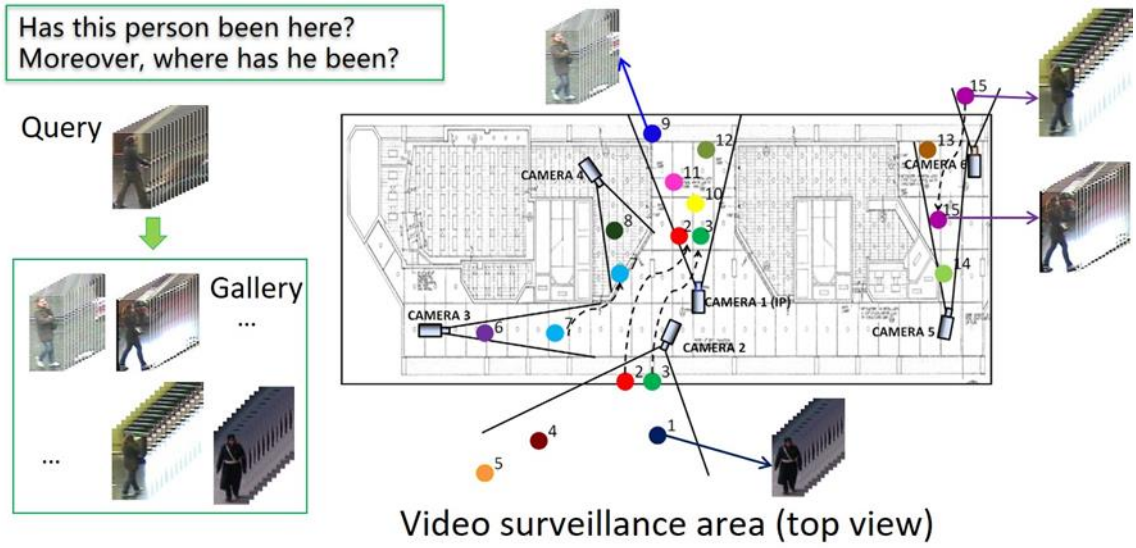


图 1.1 视频监控网络中的行人再识别

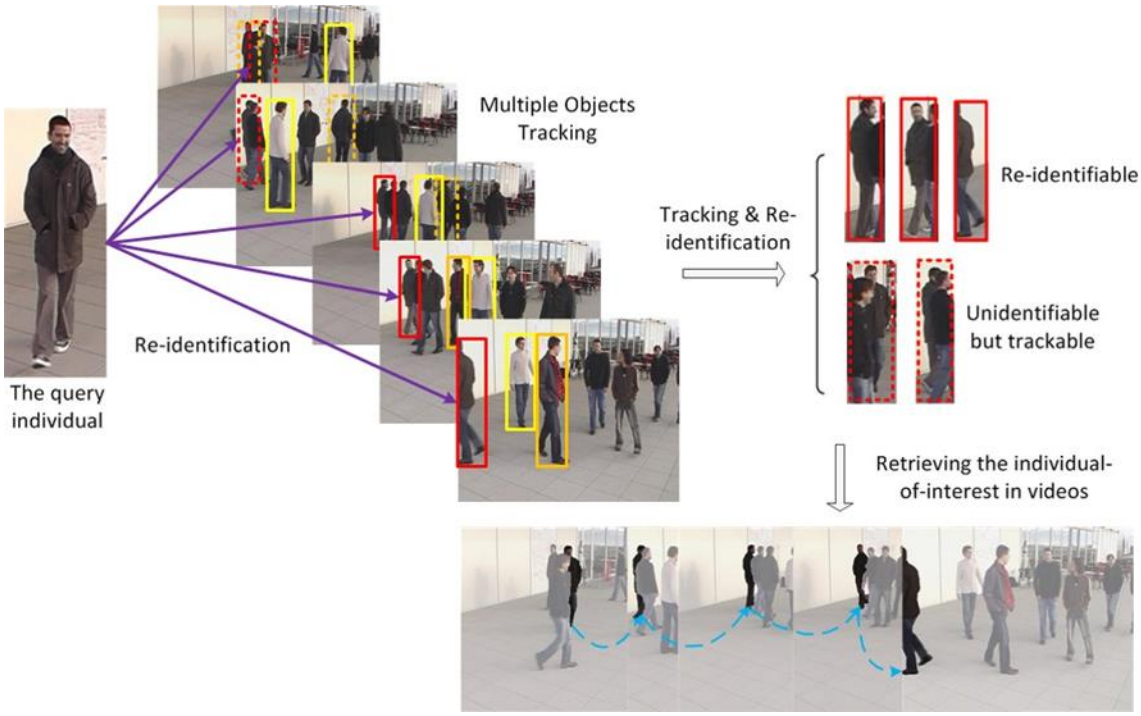


图 1.2 完整的行人再识别系统

技术上来讲，一个完整的大规模的行人重识别系统由以下三部分组成：行人检测(person detection)，行人跟踪(person tracking)和行人检索(person retrieval)。如图1.2，行人检测任务即如图中黄框、红框所示，视频监控系统中的一帧图像通过紧致的方框将行人边界明确地界定，并作为后续系统的输入；行人跟踪系统则是在某单个摄像头下，识别、并跟踪行人在该摄像头下从进入到离开的运动踪迹；行人检索则是在通过检索算法通过计算不同摄像头下不同时间下的截取道德不同行人间的特征距离，根据特定度量距离的远近，从而检索出某行人在监控系统下的出现的轨迹。通常认为前两个任务由于与跟传统的物体检测、追踪任务有着极大的相似性，故被认为是独立的计算机视觉任务，当前行人重识别的主要工作集中在最后一个模块，即行人检索。本文中，除非特别说明，行人重识别即指行人检索模块。

行人重识别旨在弥补目前固定的摄像头的视觉局限，并可与行人检测/行人跟踪技术相结合 ，可广泛应用于智能视频监控、智能安保等领域，故该技术在刑侦、安防等多个重要的领域有着十分广阔的应用前景，使得该技术的研究有着极强的现实意义和很好的应用价值，因此受到了越来越多的关注，如图 1.3，近年来的计算机视觉顶级会议中的行人重识别领域的论文数量有着较为明显的增长趋势。

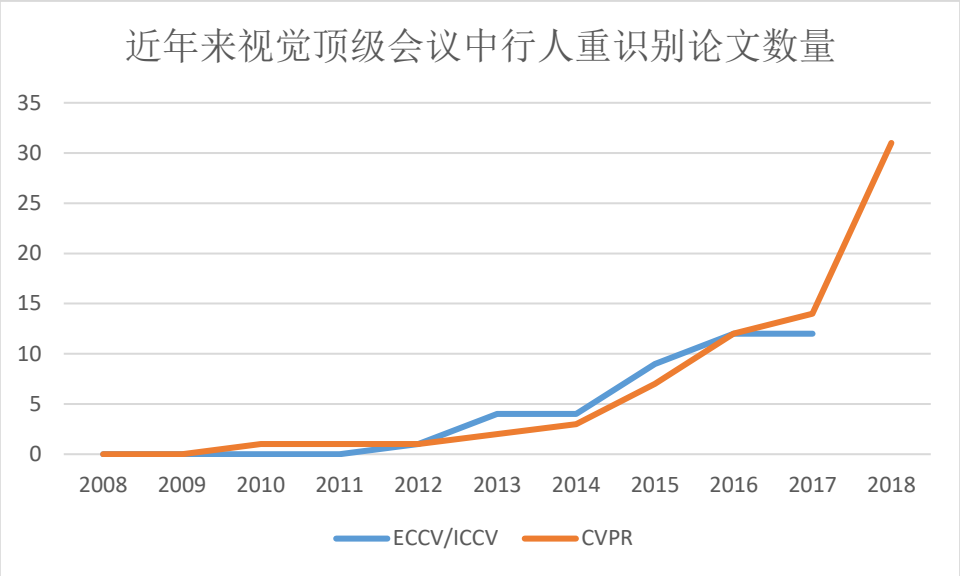


图 1.3 近年来计算机视觉顶级会议中行人再识别相关论文占比

1.2 难点分析

今年来，随着越来越多的注意力被注入到行人重识别这一领域，计算机视觉领域的众多学者也提出了大量算法与解决方案，不断尝试在公开行人重识别数据集上提高模型的性能和准确率，伴着深度卷积神经网络的发展，目前已经在公开行人重识别数据集上取得了十分理想的成绩，但这并不意味着该任务的完美解决。本小节对行人重识别任务的主要挑战和难点进行了总结，如下：

① **行人图像分辨率低** 行人重识别任务的数据主要来源于监控摄像头系统。在实际场景中，为了使得单个摄像头尽可能覆盖更大的监控区域，所采集到的行人图像分辨率通常只有 128x64 左右。这样极低的清晰度和分辨率使得行人重识别系统更多的依赖色彩等低等级的图像信息，更加细致的如面部信息等高等级信息几乎不具备。

② **类内差异性** 因为摄像头所处位置及周围环境变化导致，不同摄像头间，甚至同一摄像头下的同一行人也会出现外贸差异，影响因素主要有以下几点：

1） **视角和姿态变化** 监控摄像头网络中的相机位置、高度和角度固定，故同一行人在不同相机下呈现不同的姿态和视角，如图 1.4(a)，由于人体本身的特性，摄像头捕捉到的行人的正面、侧面和背面本身就具备着相当的差异性，如双肩包、人脸等，且行人多变的姿态更加增加了行人重识别的难度

图 1.4 行人重识别任务中的干扰因素，(a)(c)展现了视角、姿态和光照变化给行人图像造成的类内差异，每一对红色框表示该对属于同一个行人。(b)中的三张图分别来自三个行人，分别是行人间遮挡、物体遮挡给行人重识别任务带来的干扰。(d)中红色框和绿色框分别属于不同的行人，但由于分辨率的局限性，不能很好地区分开来。

2） **遮挡** 如图 1.4 (b)，监控摄像头下采集的行人图像存在着严重的遮挡问题，这些遮挡通常由其他行人、物体等造成，该问题在人流密集处更为严重，遮挡后的行人图像一方面使得行人的检测定位任务变得更加困难，同时也造成了行人部分关键信息的缺失，若直接利用该提取后的信息，则会对模型的训练造成很大的干扰。

3） **光照变化和背景干扰** 由于摄像头所处的位置不同，如室内和室外差异，以及一天内光照情况的周期性变化，使得即使同一个人同一个摄像头下的图像由于光照情况的差异有着相当的表征差异。同时，由于背景的复杂和多样性，使得区分背景和前景成为行人重识别任务中又一个挑战，局限于当前的图片分辨率和计算机视觉算法，还不能很好地消除这一个干扰。

③ **类间相似性** 由于之前提到的清晰度带来的局限性，使得区分行人的主要特征是着装和携带物特征，如图 1.4(d)，当出现行人间着装相当类似，或者由于光照等环境因素使得着装有着相当的相似度的情况时，行人重识别的性能会受到极大的影响。

④ **数据集规模有限** 如表 1.1 示，就当前而言，公开行人重识别数据集中的图片数量相比深度神经网络所需要的巨大数据量相比仍显不足，如用于图像分类的

imageNet 数据集，通过 1500 万图片，2.2.万类实现了在分类领域超过人类的视觉能力，但当前的勉强达到万张图片，行人数量仅为 1000-2000，这大大限制了行人重识别进一步突破的可能性，但大规模的标注数据集本来就十分困难，由于行人重识别任务需要在众多图片中找到属于同一行人的图片挑选出来，使得行人重识别的标注任务更加困难，尤其是要扩充至更大规模数据集。

表 1.1 行人再识别常用数据库统计资料

数据库	时间	身份	图片	相机	标注方式	裁剪大小
VIPeR[1]	2007	632	1264	2	手工	128x48
iLIDS[2]	2009	119	476	2	手工	任意
GRID[3]	2009	250	1275	8	手工	任意
CAVIAR[4]	2011	72	610	2	手工	任意
PRID2011[5]	2011	200	1134	2	手工	128x64
CUHK01[6]	2012	971	3884	2	手工	160x60
CUHK02[7]	2013	1816	7264	10(5 对)	手工	160x60
CUHK03[8]	2014	1467	13164	2	手工/DPM	任意
Market-1501[9]	2015	1501	32668	6	手工/DPM	128x64

综上所述，由于监控摄像头条件下的低分辨率，以及各干扰情况的客观存在，使得行人重识别成为计算机视觉领域非常具有挑战性的问题，也吸引了众多学者的关注和工作。近年来，借助深度神经网络的惊人表现，众多模型已经能够达到相当优异的性能，在工业界实际中的实践表明，现有的数据集规模成为行人重识别任务算法进一步突破的一大障碍。因此，如何在不通过人工标注的前提下，通过合理的数据增广方案进一步扩大现有数据集规模成为当前一个富有挑战性的热点研究领域。

1.3 国内外研究现状

近年来，深度学习作为计算机视觉领域最成功、最受认可和欢迎的方法之一，已经在行人重识别任务中展现了惊人的潜力和前所未有的性能。经典的 AlexNet, GoogleNet、ResNet 等众多卷积神经网络[1-5]的提出将计算机视觉领域的各项任务的准确率和性能提到了传统机器学习方法从未达到过的高度，例如在目前最大的图像分类任务数据集 ImageNet[6]上，基于深度学习的方法[7-12]已经在 2017 年首次超过了人类的准确率。但一个强大的深度学习模型的训练，最大的前提就是需要一个足够大的数据集，才能避免模型的过拟合，从而具备强大的鲁棒性和泛化能力，实现在该领域真正的突破，如深度卷积神经网络在分类任务上已经有了超越人类

的性能，主要原因是分类任务上已经构建了 ImageNet 这一拥有 2.2 万类的 1500 万图片的数据集。

表 1.2 行人重识别任务数据集 Market-1501 代表性算法性能

算法	rank-1	mAP
BOW[48]	44.4	20.8
LDNS[49]	61.0	35.7
SVDNET[50]	82.3	62.1
TriNet[51]	84.9	69.1
CamStyle[26]	89.5	71.6
DuATM[52]	91.4	76.6

表 1.3 行人重识别任务数据集 CUHK 代表性算法性能

算法	rank-1	mAP
BOW[48]	25.1	12.2
LOMO[53]	30.8	17.0
SVDNET[50]	76.7	56.8
TriNet[51]	72.4	53.5
CamStyle[26]	78.3	57.6
DuATM[52]	81.8	64.6
DeepTransfer[54]	84.6	87.4

深度学习在计算机视觉领域的惊人表示也使得行人重识别任务得到了进一步的发展和突破[13-21]，如表 1.2,1.3，在当前最广泛使用的行人重识别数据集 Market-1501 和 CUHK 上，已经有个别算法在 top-1 准确率上超过了 90%，但这并不意味着行人重识别已经能够被很好地解决。

众多算法的提出一方面展示了深度卷积神经网络惊人的学习能力，并验证了众多思路和算法的能效，却也暴露了行人重识别领域目前最大的瓶颈：数据集规模远远满足不了一个强大模型的训练需求。

Tian 等人的实验[22]指出，即便将前后景分离后的背景作为数据集进行模型训练，top-1 的准确率也能够达到 35%，即训练过程中背景因素对行人重识别的性能有着相当大的贡献，也就是说，现有数据集的规模过小，以至于训练的模型泛化能力太差，对干扰因素有着很差的鲁棒性，不能够很好地学习对人体信息进行学习。

针对这一问题，解决方案上主要有两种问题，一种是以 Sun 为代表[23-25]的对人体更加精细化地学习，这样一来能够排除遮挡、背景等干扰因素，从而获得模型性能的提升，但该思路有以下缺陷：1，这样更加精细的标注需要更加大规模的人力，这无疑是更加不现实的；2，还是不能从根本上训练一个对各种干扰因素，如光照、姿态等，从而训练具有鲁棒性和强大泛化能力的模型。

另一种思路则是进一步扩充数据集[26-28]，如通过强大的计算机视觉算法生成新的更大规模的、具备更大多样性的数据集，如图片合成技术，编码器-解码器架构，声称对抗网络等，相当多的工作验证了这一方向的有效性，其中的一个主要方向是多姿态的行人图像合成。

多姿态行人图像合成致力于通过训练算法模型，使得该模型可以基于已有行人的图像，批量合成该行人在已有图像中不具备的姿态，从而在视角、姿态两个维度赋予数据集更大的多样性。该研究一方面能够极大的节省标注数据所需要的海量的人力、物力，同时也能够基于姿态、视角合成更多的数据，从而极大扩充数据集的规模。在之前提到的干扰因素中，由于光照、遮挡问题的独特性，在新的数据集设计上很少考虑到这两点，对于光照因素，在极暗或者逆光情况下，即便行人肉眼也很难进行识别和辨认，遮挡问题同理，故在设计行人重识别算法时，普遍默认忽略这两个因素，更多的算法集中于解决多视角、多姿态下行人的重识别任务，故多姿态的行人图像合成被普遍认为是行人重识别数据集进行进一步扩充，突破行人重识别任务当前瓶颈的主要策略。

在刚刚公布的计算机视觉顶级会议 CVPR2018 中，在被接收的 33 篇行人重识别论文中，仅仅多姿态行人合成的算法研究就有 5 篇，在这样一个相当小的领域里有着如此这样高的关注度，也从侧面说明了这一方向的较高的关注度和极大的潜力。

1.4 本文主要工作

本文针对当前行人重识别任务的主要瓶颈，即行人重识别数据集规模过小这一局限性，尝试基于生成对抗网络，通过姿态引导的多姿态行人图片生成，生成更多符合真实样本特征空间分布的新样本，加深对深度神经网络对人体结构的理解和学习，为行人重识别数据集的数据增广提供了一种可能性，为训练和实现更具鲁棒性和具有更强泛化能力的深度神经网络的提供必要的数据条件。创新点如下：

① 提出一种新的图像合成算法 创造性的将生成对抗网络运用到多姿态行人图像合成中来，基于生成对抗网络强大的模拟与学习特性，利用生成器与判别器间的博弈，能够极为精确地学习到已有数据的特征分布，并模拟生

成出新的类似于已有的真实训练样本。通过在训练模型的输入中加入目标的pose信息，该信息能够在生成模型的学习和生成中起到关键的引导作用，从而使得设计的模型可以通过在输入中给予任意姿态信息，能够生成任意在输入中指定的动作。

② 提出了新的生成对抗网络模型 通过在经典的编码器-解码器的瓶颈处添加给予 softmax 的分类器，加强卷积神经网络提到的的特征信息中的行人个体特征，从而保证在生成对抗网络的特征空间中较为稳定的保证个体信息的特征。

③ 设计了新的目标函数 通过多层特征信息及多个分辨率下的特征信息距离计算，以及 patch 损失函数的加入，保证生成数据在底层信息、高层信息、低分辨率、高分辨率下都能尽可能保留输入的行人图像的与姿态无关、与个体信息相关的信息。

1.5 章节安排

本论文章节安排如下：

第一章首先介绍了行人重识别问题的定义、研究背景与意义。在分析了行人再识别问题当前的主要瓶颈之后，就国内外在该瓶颈上所做的工作进行了简要介绍，并总结了本文的主要工作。

第二章对国内外相关研究工作进行了详细介绍。

第三章对基于生成对抗网络的多姿态行人图像合成模型进行了介绍，介绍网络的各个子部分以及各部分对应的目标函数。

第四章给出生成对抗网络模型的具体设计与细节，并就对网络训练后生成的结果进行分析，与已有方法、优化前的模型结果进行比对和解读。

第五章对全文工作进行了总结，并对未来的研究方向进行了进一步的讨论。

2 相关工作

本章主要从生成对抗网络模型、基于条件信息的生成对抗网络、图像-图像翻译和基于信息分离的生成对抗网络四个方面对图像合成算法领域当前国内外的研究进展和突破进行详细的介绍。

2.1 生成对抗网络模型

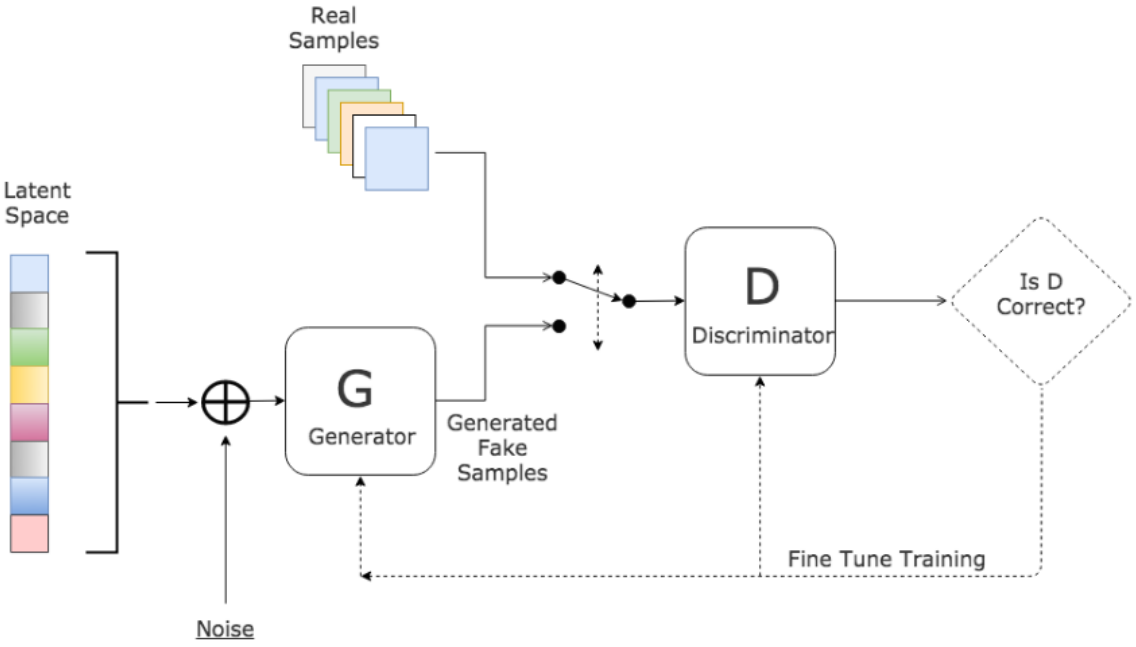


图 2.1 生成对抗网络的架构

生成对抗网络[29]一直是近年来生成模型领域相当理想的图像合成模型，并已经被应用到大量的数据集和不同类型的任务中。生成对抗网络是通过模拟目标分布，来生成新的数据样本，其由两部分构成：生成器（generator）和判别器(Discriminator)。生成器的目标是生成新的尽可能符合目标特征分布的新样本，判别器的目标则是尽可能降低生成样本与目标样本间的特征分布的距离，两个模型同时训练并进行优化，生成器不断尝试提高生成样本的质量、使之与原样本分布尽可能一致，而判别器则不断尝试提升判别器进行判别的准确性，通过这样的对抗训练，实现对目标样本空间的特征空间进行模仿和学习。

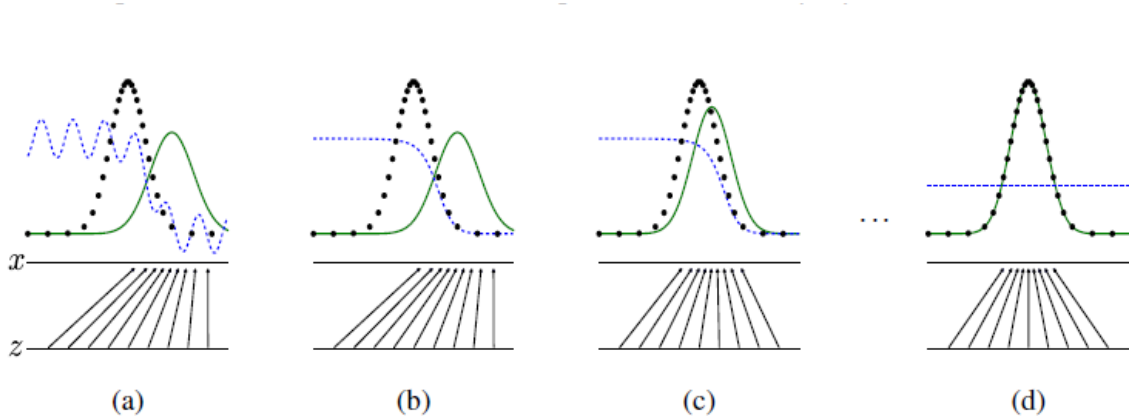


图 2.2 生成对抗网络的学习过程

如图 2.2，生成对抗网络在训练的同时更新判别器（蓝色虚线）与生成器的生成数据分布（黑色虚线），通过对抗训练，使得判别器能够区分由生成器生成的数据（黑色虚线）和原有的数据分布（绿色实线），z 则是初始的噪声分布，z 则是生成器的映射过程，通过不断的生成器与判别器的迭代训练，生成器生成的数据不断与目标数据分布接近，判别器也在不断地提升自身判别能力，最终目标是生成的数据与目标数据不能被判别器区分开来。

生成器网络将潜在变量 z 映射到数据空间以合成假样本 G（z），判别器将真实样本 x 与假样本 G（z）进行判别，生成是真实样本的概率 D（x）或 D（G（z））。生成器和判别器都以相互对抗的方式进行训练：生成器试图欺骗判别器，而判别器试图区分真实数据和由生成器合成的样本。 在数学上，该训练过程可以通过值函数 V（D，G）上的极大极小函数来建模：

$$\min_G \max_D V(D,G) = \mathbb{E}_{X\sim p_{data}}[logD(X)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(Z))] \qquad (2.1)$$

具体而言，对于生成对抗网络，目前已经有许多方法来提高其不稳定训练过程的稳定性，也有很多工作给生成对抗网络提供条件信息，使得生成的图像样本不仅逼真，而且还与条件施加的约束相匹配。一些人在离散的标签上调节生成对抗网络[35,36]，而许多其他作品通过在图像上进行调节生成对抗网络来合成图像，用于诸如域转移（domain adaptation）[26,27,28]的任务，在施加约束下的图像编辑[59,61]，超分辨率图像[47]，风格转移[46,12]合成图像。在这里，条件信息也可以是文本格式，可以依据图像描述来生成匹配该描述的图像。里德等人[37]首先提出了一个基于条件信息的生成对抗网络框架的端到端深度神经架构，它根据自然语言描述成功地生成了逼真的图像（64×64）。他们进一步开发了另一个模型[67]，可以使用文本描述，对象位置和其他注释等条件合成 128×128 图像。 Zhang 等人[45]提出了 stackGAN，

它将文本到图像合成过程分解为两个阶段，并成功地生成了逼真的256×256图像。

(a) VAE (b)DCGAN

2.3 基于自编码器、DCGAN等传统生成模型合成的图像

近年来，生成对抗网络[29]、对抗自动编码器[30]、变分自动编码器[31]和ARMs[32]等能够生成逼真、锐利且清晰图像的模型一直引领着图像合成的研究。传统图像生成算法主要用生成对抗网络[29]、自动编码器[31]来由图像噪声生成的特征分布映射到真实数据的分布上。卷积自动编码器以及AAE[30]已经验证了将自编码器用作生成器的可行性，但是在这种情况下，对于较为复杂的生成分布，如他人体图像，不能很好地学习该特征映射。因此，当要合成的图像时行人图像这样具有复杂特征分布的图像时，传统的图像合成方法不已经不再适用了。比如，郑等人[33]直接利用DCGAN[34]结构来合成行人图像，但如图2.3(b) 所示，该模型生成的图像相当不逼真。

2.2 基于条件信息的生成对抗网络

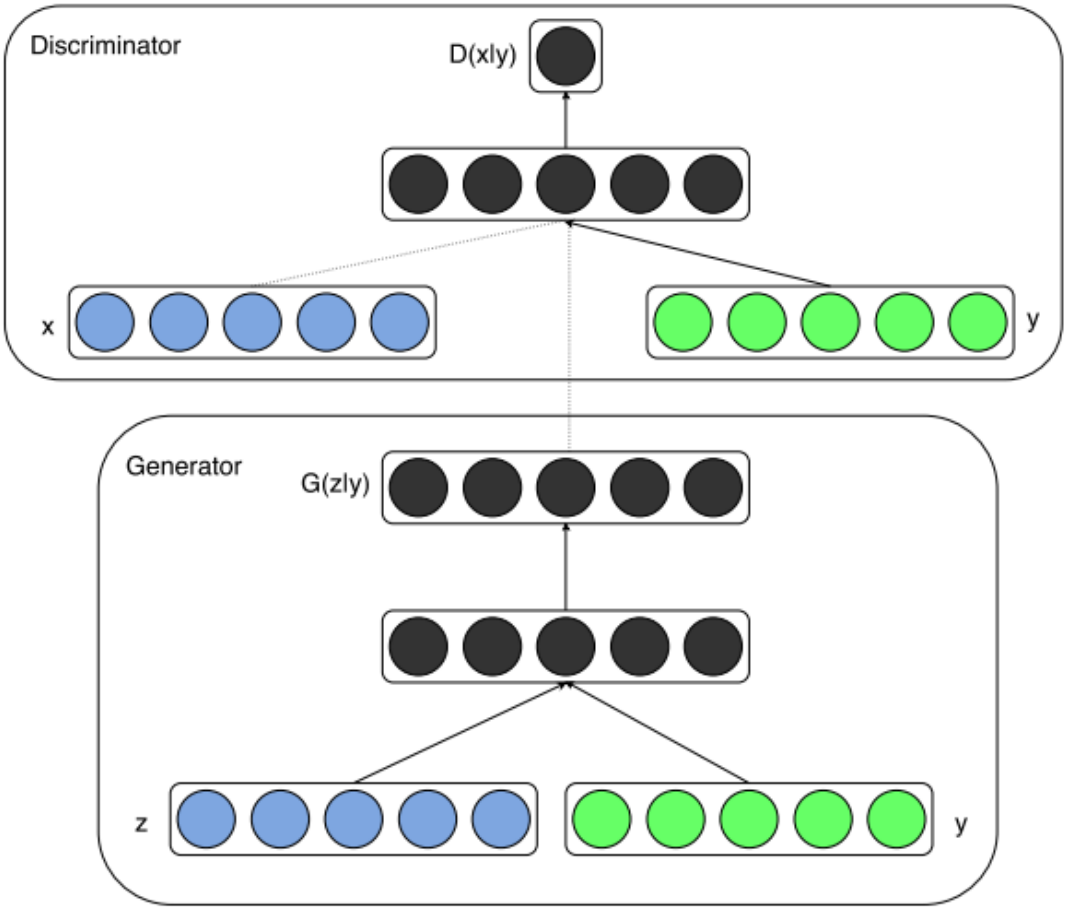
如果生成器和判别器都以一些额外的信息y为条件，则生成对抗网络可以扩展到条件模型。 y 可以是任何类型的辅助信息，例如类别标签或来自其他模式的数据。 我们可以通过将y输入到鉴别器和发生器作为附加输入层来执行调节，如图2.4所示。

在生成器中，先验输入噪声 $p_z(z)$和y以联合隐藏表示的形式进行组合，并且对抗训练框架在如何隐藏特征空间的组织和表示有着相当大的灵活性，故在辅助信息起到限制作用的同时，同时保证生成样本的多样性。目标函数表示为：

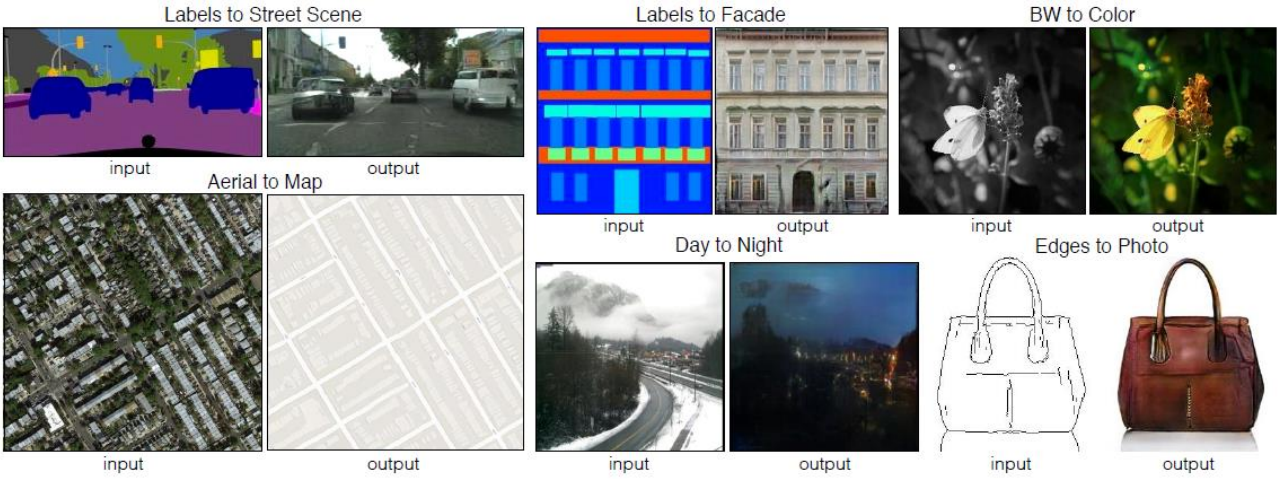
$$\min_G \max_D V(D,G) = \mathbb{E}_{X\sim p_{data}}[logD(x|y)] + \mathbb{E}_{z\sim p_Z}[\log(1-D(G(z|y))] \qquad (2.2)$$

基于条件来训练生成对抗网络并限制生成过程的方法，到目前为止，已经探索了几个条件，如离散标签[35,36]和文本[36]。图像也被用作一种条件，例如在图像到图像翻译[37]，未来帧预测[38]，图像修复[39]和人脸对齐[40]的问题中。最近Zhu等人[41]使用文字描述和图像作为生成新服装的条件。他们都提出了多视图下行人图像生成问题的生成对抗网络模型。然而，这两种方法使用全监督，即两个不同姿势的同

一人的两副图像穿着相同。以完全不受监督的方式处理问题变得非常困难，需要更复杂的网络设计，尤其是在估计生成图像的损失时。



2.4 条件生成对抗网络



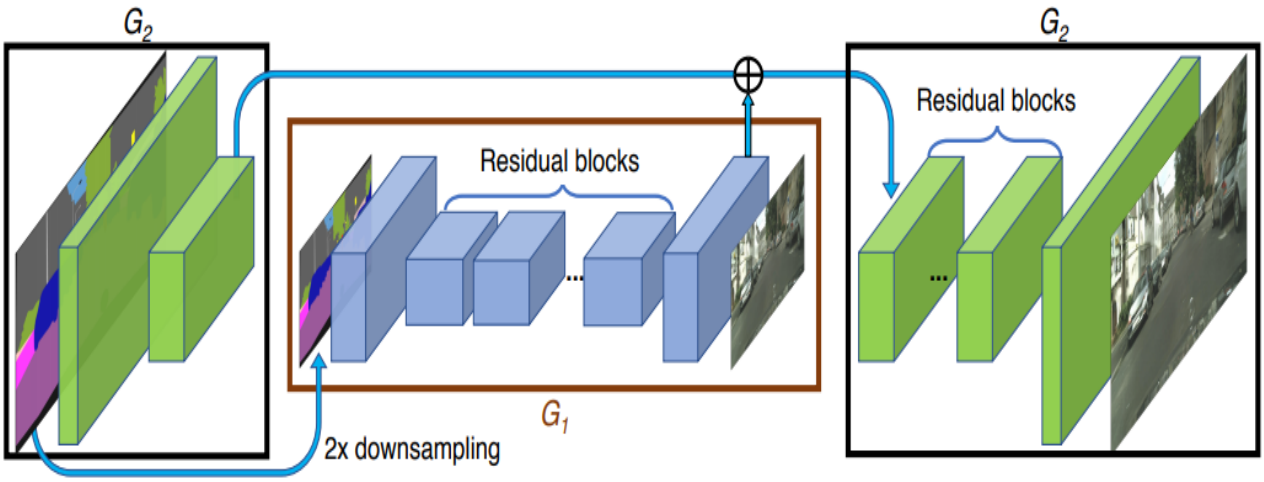
2.5 条件生成对抗网络几个应用示例

2.3 图像-图像翻译

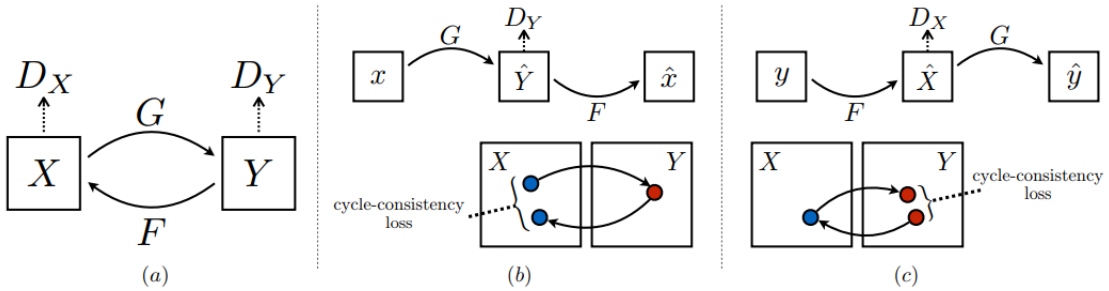
在基于条件信息的生成对抗网络基础上，SiIsola 等人[38]提出了图像到图像翻译（image-to-image translation）的第一个统一化的框架，已经引起了极大的关注度并具备极强的鲁棒性，如图 2.5 所示，已经能够执行色彩渲染、风格转移、图像填充等诸多任务。

就在今年，基于 Silsola 提出的经典的图像到图像的翻译框架，Wang 等人提出了生成更高分辨率图像[47]的算法模型，最终生成图像的分辨率达到了 2048x1024 . 如图 2.6 所示，该模型首先通过G_1在低分辨率下进行训练，然后将G_2与已有的G_1组合，将组合后的网络在高分辨率下共同训练，值得注意的是在图中右边，即G_2反卷积部分中，其输入时G_2卷积部分与G_1 逐元素的和，其训练策略另一个值得注意的点则是通过多尺度的判别器，三个判别器分别在原分辨率、二分之一分辨率，四分之一分辨率下对生成图片与目标的原图片进行判别，从而保证在低层与高层信息都保证尽可能的一致性。

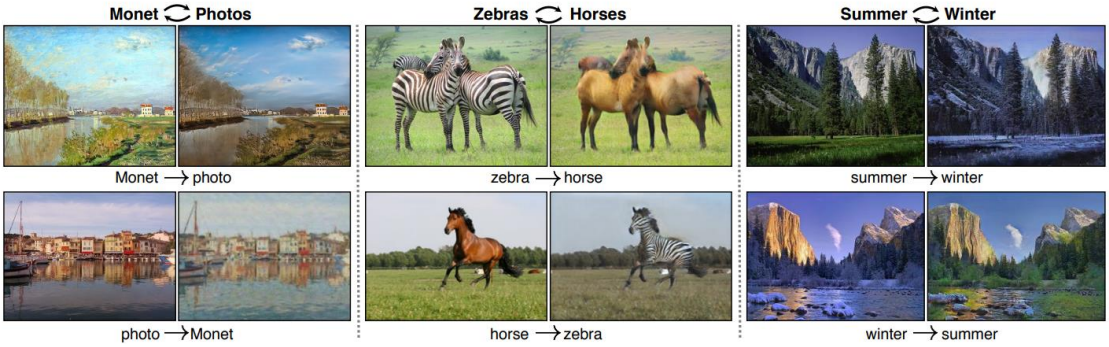
最近的研究也试图在没有监督的情况下学习图像对图像的翻译，这个思路本质上是有缺陷的，需要额外的限制才能达到一定的效果。 一些工作尝试在图像翻译过程中强制保留源特征空间中的某些属性，如像素值，像素梯度，语义特征，类标签或成对样本距离。



2.6 高分辨率的图像到图像翻译模型



2.6 基于循环一致性损失的图像到图像翻译模型



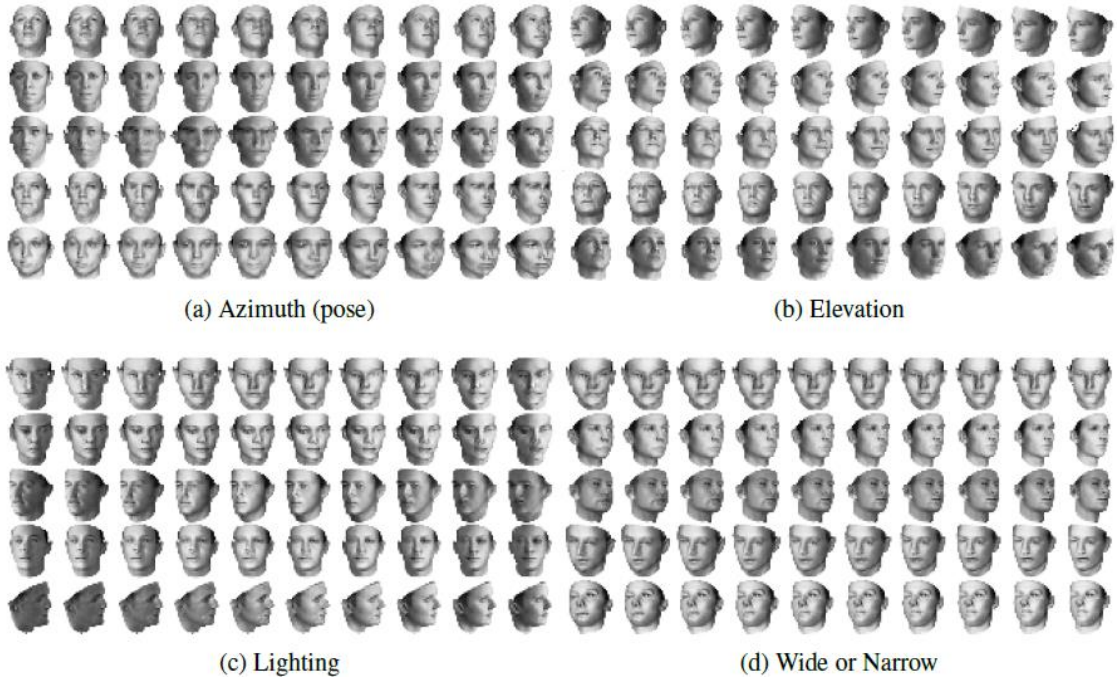
2.7 基于循环一致性损失的图像到图像翻译

另一个受欢迎的约束是循环一致性损失（Cycle-Consistent Loss）[61]。 如图 2.6 示，该模型包含两个映射函数 G：X→Y 和 F：Y→X，以及相关的对抗判别器D_X和D_Y。D_Y鼓励 G 将 X 转换为与 Y 无法区分的输出，D_X和 F 则相反。为了进一步规范映射，我们引入了两个周期一致性损失（Cycle-Consistent Loss），（a）正向循环一致性损失：x→G（x）→F（G（x））≈x，（b）反向循环一致性损失：y→F（y）→G（F（y））≈y。更直观来讲，以图 2.7 中最中间的斑马-马转换为例，一个映射函数将斑马映射为马，另一个映射函数将生成的马的图片作为模型输入，将马映射回原来的斑马，若训练后的模型真正地学习并能成功模仿目标特征空间，则能够通过这两次映射将图片-图片的翻译映射回原点，即斑马。

另外，刘等人[55]提出了 UNIT 框架，它假定一个共享的潜在空间，使得两个域中的相应图像映射到相同的潜在代码。（multimodal-img2img）

大多数现有的图像到图像翻译方法的一个重要限制是翻译输出中缺乏多样性。为了解决这个问题，一些作品建议在给定相同的输入的情况下同时生成多个输出，并鼓励它们不同[56]。 尽管如此，这些方法只能生成离散数量的输出。 朱等人[57]提出了一个可以为连续和多模态分布建模的 BicycleGAN。 然而，所有上述方法都需要以输入-输出对的形式进行监督。

2.4 基于信息分离的生成对抗网络



2.8 基于 InfoGAN 的特征解耦学习,通过信息的分离学习，能够在一个维度或属性上对生成图像进行相应的改变，如图示，分别改变了生成人脸图片的姿态、高度、光照、宽窄

基于信息分离的生成对抗网络框架尝试将图像中的信息分离为两类：共有信息（mutual information）和隐藏表示（hidden information）,如图 2.8 中，人脸的结构、肤色等信息就是共有信息，而光照、姿态等就是隐藏表示，而基于信息分离的生成对抗网络框架则是尝试在最大化保留共有信息的前提下，学习可理解的某个维度或者属性上的表示。

很少有工作试图通过学习输入图像的分离表示，即对图像特征解耦，在指定特征空间上进行学习。陈等人[62]提出 InfoGAN 是 GAN 的一个扩展，以无监督的方式使用共有信息来学习解耦表示，例如从 MNIST 数据集中的数字形状中抽取书写风格这一特征，从 3D 渲染图像的中抽取光照、姿势特征等。 Cheung 等人[59]在半监督自动编码器体系结构中增加了一个交叉协方差，以便将特征分布解耦，如数字的手写风格和面孔中的主体身份。 Tran 等人[25]提出 DRGAN 从一个或多个人脸图像中学习共有信息，并生成在某一个维度进行区分表示的图像，以合成在保留个体原有信息的前提下，在目标姿态下的人脸。

2.5 基于生成对抗网络的多姿态行人图像合成

由于人体具有复杂的非刚性结构，具有很大的自由度[61]，因此一些作品使用结构条件来生成人物图像。里德等人。Reed 等人提出了使用姿态关键点和文本描述作为条件的生成对抗网络网络[62]，而在 Reed 等的模型[64]中，除了调整部分关键点，分割掩码（latent code）和文本以外，还使用了 PixelCNN 的扩展 MPI 人体姿态数据集等。拉斯纳等人[65]通过对细粒身体和衣服部分进行调理而产生衣服中的人的全身图像，例如，姿势，形状或颜色。赵等人[44]将生成对抗网络的优势与变分推理（variational inference）相结合，生成着装行人的在多视图下的图像。马等人[43]提出对图像进行解耦处理，并提出将关键点以灵活的方式传递人体姿势，且面部个关键信息可以进行相应的转移。然而，他们的方法需要成对的人员图像对的训练集，这会花费昂贵的人类注释。最近，朱等人[58]提出了使用循环一致性来实现域之间不成对的图像到图像翻译的 CycleGAN。他们在外观变化方面取得了令人瞩目的成就，但在几何变化方面几乎没有成功。

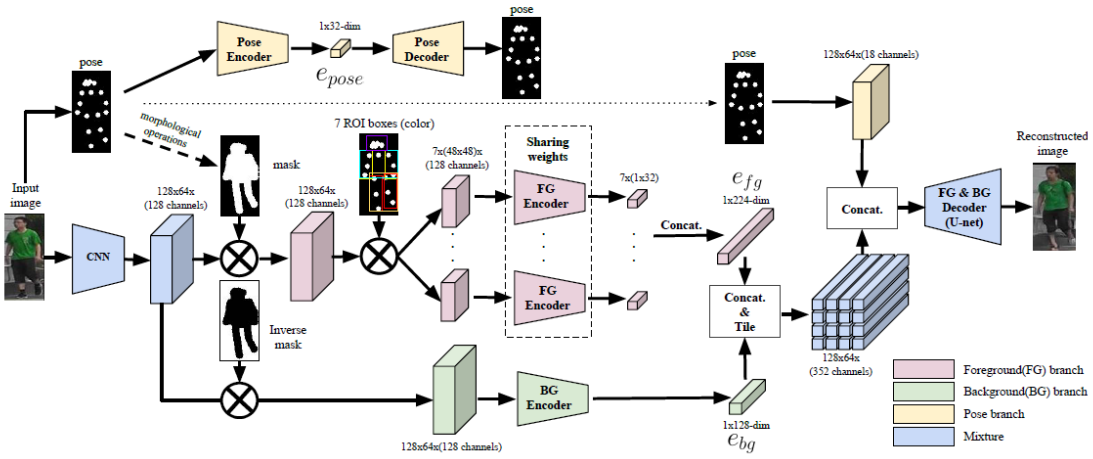


图 2.9 基于人体部位解耦的生成对抗网络，用以合成多姿态人体图像的框架

在多姿态行人图像合成上，Ma 等人提出了基于行人部位特征解耦的生成对抗网络框架[60]，如图 2.9，网络模型一共分为三条支路，第一支路是姿态信息的学习，第二条是背景信息的学习，第三条则是人身体信息的学习，该支路又根据人体的各个部位进行了进一步的分割，并分别针对七个感兴趣区域（ROI, region of interest）单独进行学习，在三条的解耦信息分别编码后，基于编码器-解码器结构，这九个编码器编码得到的特征向量被组合起来，生成一个总的特征向量，并将该特征向量进行解码，最终生成与目标姿态相符的图片。

3 基于生成对抗网络的行人图像合成

3.1 引言

值得注意的是，大多数基于条件信息的生成对抗网络（condition GAN）更多的将精力集中于潜码的表达或者图像质量上，而忽略了行人个体信息的保留以及个体之间特征的差异性。W 等人也提出基于生成对抗网络的特性，同时也是缺点：生成器远远比判别器难训练到收敛，因此判别器很容易早于生成器收敛，从而出现过于强大的判别器，导致生成器与判别器之间的对抗训练与学习无从进行，并抑制生成器对特征空间的学习和模仿。考虑到以上两点，本文基于经典的编码器-解码器结构，提出了包含多分类器、小分辨率图片卷积器在内的新的生成器模块设计，保证了个体信息尽可能的保留、生成图片低层、高层信息真实图片尽可能的一致；基于 Patch 损失函数，缓和生成器与生成器之间的对抗训练，避免了两者之以提前收敛，从而使得训练尽可能的稳定。

3.2 网络模型设计

如图 3.1，是本文提出的模型的整体架构，生成器网络由四部分组成：编码器、解码器、分类卷积网络、小分辨率图像生成卷积网络；共设计了两个判别器网络，分别对应生成器中间层特征卷积而来的小分辨率图像和生成器网络最终生成的图像。

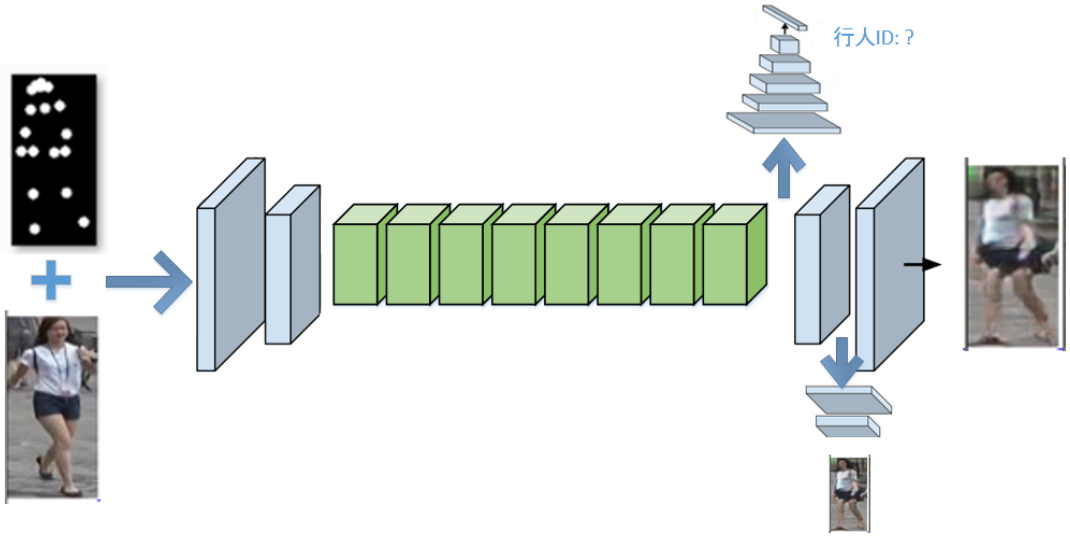
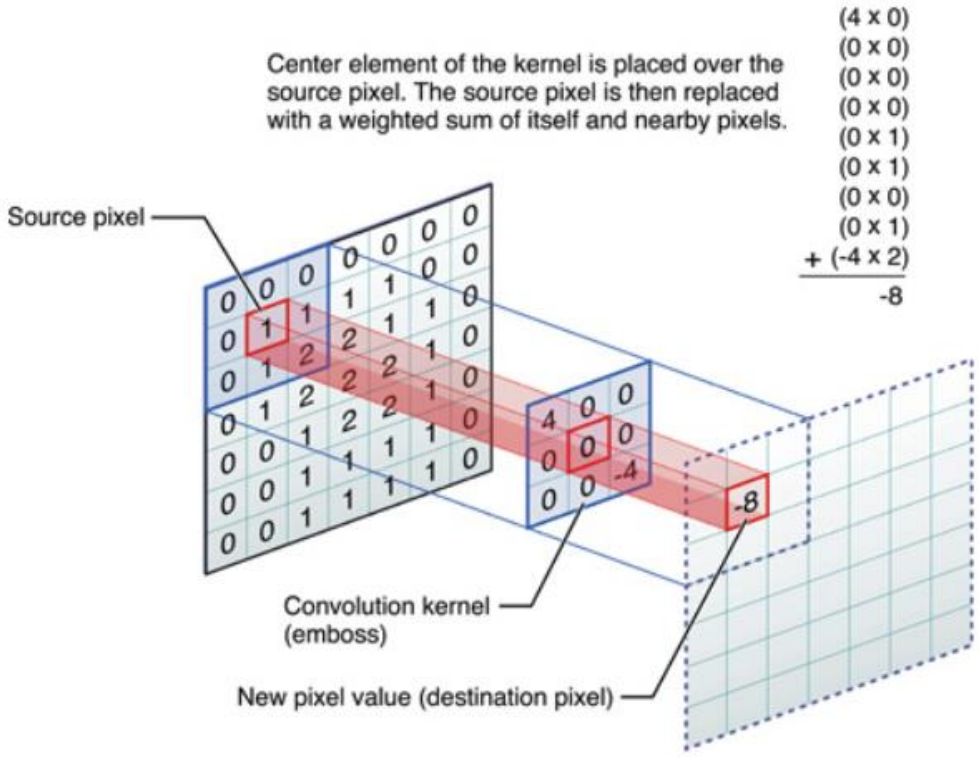


图 3.1 网络生成器设计

3.2.1 生成器设计



3.2 卷积神经网络中卷积过程示意图

如图 3.2，从后方的九个源像素源的通过卷积核生成一个新像素的过程即卷积，通过卷积核，相比于原始的神经网络的全连接，能够在学习各层次的图像信息的同时极大地减少网络的大小与参数量，并极大地降低计算复杂度。卷积神经网络的诞生是推动整个计算机视觉近年来在包括行人重识别在内的各个领域任务突破的关键因素。卷积的逆过程即反卷积，通过卷积核，将一个像素上采样至卷积核大小的像素范围，通过将同等步长、卷积核大小的卷积、反卷积层对阵连接，实现了经典的编码器-解码器（encoder-decoder architecture）结构。但卷积和反卷积的以下特性使得编码器-解码器结构不能完全满足作为生成对抗网路中生成器的条件：

① 卷积过程捕捉全局的信息而非局部信息，诸如形状、色彩和等信息，丢失了很多低层次的信息，又因为随着层数的逐渐加深，提取的特征也逐渐向高维度、高层次的信息靠近和集中，故丢失了很多小区域的细节，如衣服的条纹信息，较小的物体，锐利的边缘信息等。

② 反卷积过程原理上是上采样过程（up-sample），基于输入的特征进行提升图片的分辨率，但由于低层次信息已经丢失，反卷积后易生成模糊、仅具高层信息的图片

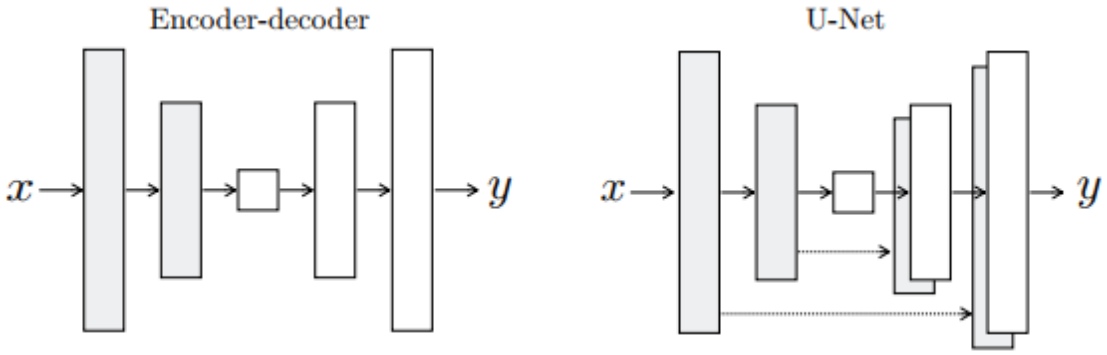


图 3.3　编码器-解码器模型（左）与 U-net（右）与生成对抗网络结合的效果对比

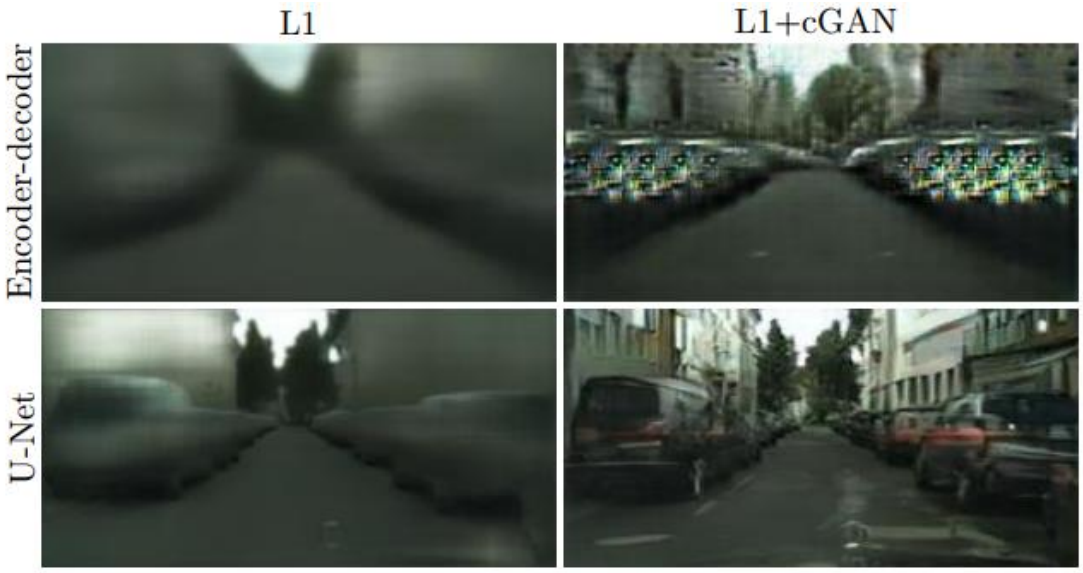


图 3.4 编码器-解码器模型（左）与 U-net（右）与生成对抗网络结合的效果对，可见跳跃连接确实能够在生成对抗网络中较好的保留细节信息

在网络的设计上，基本遵循了经典的编码器-解码器结构，但不同于广泛使用的 U-net 构型，模型使用了基于 residual block 的编码器-解码器结构，如图 3.3（a）所示，编码器通过左侧 n 层卷积（阴影部分），生成在瓶颈处（中间正方形）的特征向量，再通过与先前卷积过程中步长、卷积核大小等相同的 n 层反卷积，生成新的图像，但存在的一个问题由于卷积的特性，会丢失很多低层次的信息，从而使得生成的图片失真、模糊；U-net 则能够较好的解决这个问题，如图 3.1（b）所示，U-net 使用跳跃连接（skip-connection），通过将对称结构中卷积层的输出直接与对应的反卷积层的输入结合起来，从而使得低层次信息能够很好地保存。

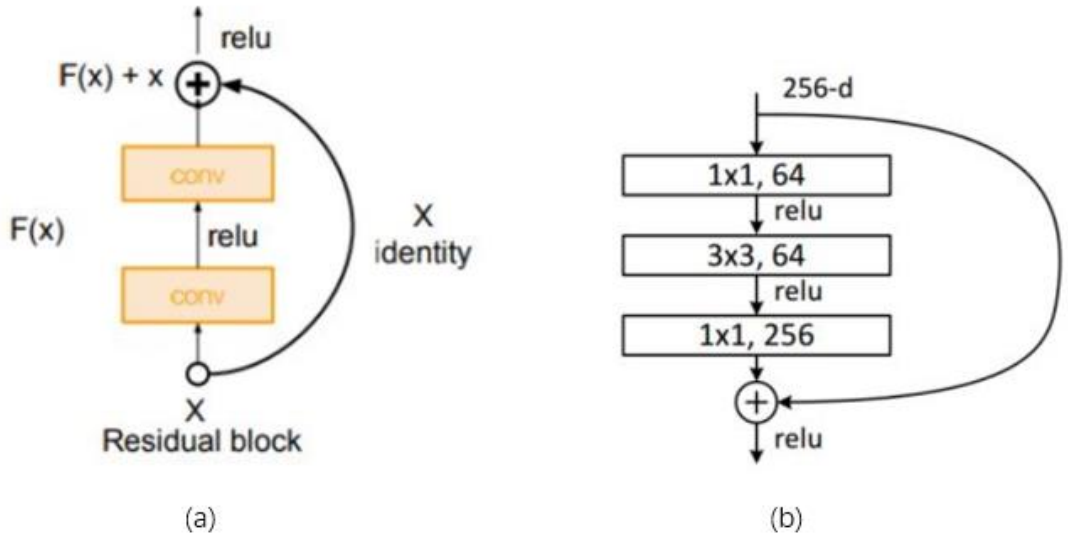


图 3.6 残差模块示意图（a）,本文中用到的残差模块（b）

不同于之前提到的两种构型，本文章基于残差模块（residual block），通过另一种形式的跳跃连接，既能够较好的保留低层次信息，又能免去跳跃连接带来的计算和模型上的负担。

如图 3.6(a)，残差模块（residual block）由两层卷积核一条跳跃连接组成，x 作为该模块的输入，经过两层卷积生成 F(x), 该模块的输出则是将 x 与 F(x)加起来，作为下一个模块的输出。在本文的网络中，采用了图 3.4(b)中的设计。

本文在解码器的中间层抽取了特征，并通过一个卷积层输出了原图四分之一的图片，作为生成器的限制之一，具体用途在 3.2.3 ，3.2.4 还会详细提到。

3.2.2 分类器设计

不同于传统的图像合成任务，合成的行人图像的有效性很大程度上基于个体信息的保留，换句话说，基于已有行人图像生成该个体的新图像，新图像必须符合原有个体的信息，即保留具备个体不变性（identity-invariant）的信息，若生成的新图像不能同原有个体有较好的匹配，即便能够将目标姿态信息习得，生成图片质量再高，也是无效图像。

截止目前为止，基于生成对抗网络的行人多姿态生成还未将个体信息的保留与监督纳入考虑，本文创造性地在生成器的瓶颈处，添加了一个基于 softmax 多分类的卷积神经网络，直接将瓶颈处，即前面编码器卷积而得来的特征向量进行针对个体的多分类，相当于在编码器处增加了一个针对个体信息的监督。

通过 softmax 层实现了基于行人 id 的多分类任务，基于在瓶颈处的分类器，作为一个限制，起到了如下的作用：

① 限制了瓶颈处的特征，即之前的卷积层提取的特征，都是与个体信息有着强烈相关性的，防止了个体重要信息的丢失。

② 在前面的卷积层提取行人信息的同时，保证了个体间的（intra-identity）差异，避免丢失独属于行人个体的、与其余行人不同的信息，从而避免了在已有工作中出现的，不同行人的生成图片间一定的相似性。

3.2.3 判别器设计

众所周知，在不同的尺度（即分辨率）下，信息的富集程度不同，一个好的生成模型，应当在不同的尺度下都与原图有着很强的相似性和尽可能小的差异，基于这个原理，本文的模型中设计了两个分类器，分别针对整个生成器的输出和 3.2.1 中提到的，基于中间层特征通过卷积生成的原图四分之一大小的图片，由于两个判别器的输入的尺度不同，故针对小尺度的判别器的层数比针对原图的判别器少两层。

通过多尺度的判别式，该模型实现了：

① 基于小尺度更多集中于全局信息的优点，通过小尺度的判别器实现了对生成图片与原图片全局信息、高层信息的一致性约束；通过大尺度图片和其对应的判别器实现了对生成图片与原图片局部信息、低层信息一致性的约束，两者相互辅助，实现对生成图片在局部信息与全局信息的约束。

② 通过对中间层的特征信息进行约束，进一步约束了生成器的解空间，相当于在图片生成过程中加了一层约束，作为增强信息促进反卷积层在上采样过程中学习更加有利、有用的特征，从而加快生成器的收敛，使得训练过程更加平稳。

3.3 潜码设计

作为基于条件的生成对抗网络以及 InfoGAN 的重要条件，潜码（latent code）的设计和选择直接影响着生成图像的质量和最终模型训练的方向，基于我们模型设计的目标：生成多姿态、视角的行人图像，因此姿态信息是我们需要加入到网络输入中的潜码。

3.3.1 潜码的提取

潜码的提取，即姿态信息的提取上，已经有着众多的思路和算法提出，在其中目前业界最领先的是 OpenPose 模型[63]，如图 3.7 所示。

在该模型中，能够提取人身体上 18 个关键点的信息：鼻子、脖子、膝盖、脚

踝、臀部、手腕、手肘、眉毛、肩膀。

图 3.7 生成样本示例

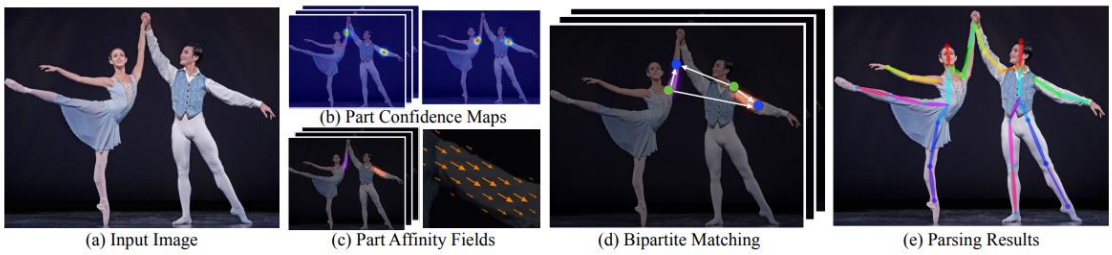


图 3.8 生成过程示意，(a)：输入图片，(b)：部位置信图，(c)相邻亲和场，(d)：二部匹配，(e)分割结果

如图 3.8，该算法提取姿态信息的流程是：

① 以图片作为输入，通过反向传播网络，同时预测身体部位位置的一组二维置信度图，以及用于编码身体部分间关联程度的相邻亲和力（part affinity）的一组二维向量

② 基于相邻亲和场（part affinity field）进行身体部位的聚合，如图中将手肘与肩部联合来形成胳膊。

③ 基于贪心推理（greedy inference），进行人与人之间姿态信息的分割。

3.3.2 潜码的输入

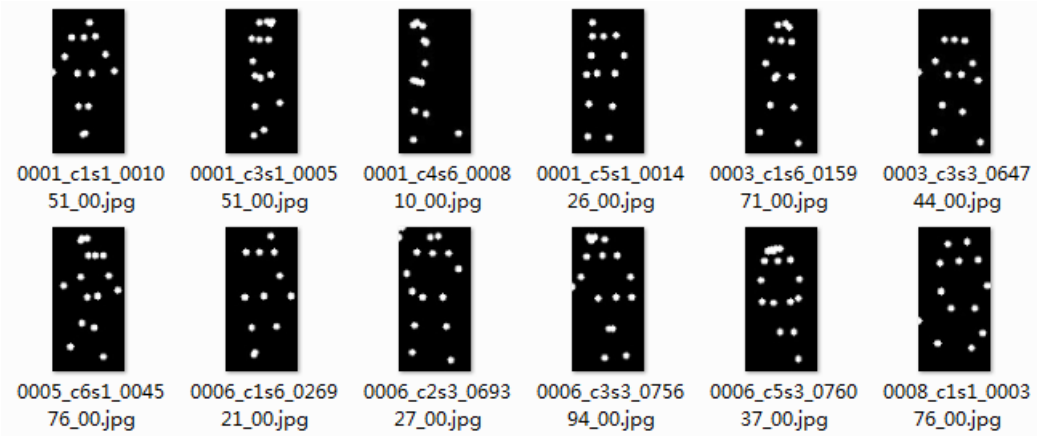


图 3.9 姿态输入信息样本示例

与之前众多基于条件信息的生成对抗网络（condition GAN）一样，通过 OpenPose 库提取到各个关键点的坐标信息后，如图 3.9，我们生成了一张黑色背景，以半径为 5 的白色圆形表达关键点的图片，相比于传统的神经网络的三通道对应 RGB 图像的输入不同，我们将含有姿态信息的图像作为第四通道与原有 RGB 图像叠加起来，作为输入的第四个通道。

3.4 目标函数设计

在本部分，I_A指某张姿态为 A 的行人图像，I_B是姿态为 B 的目标图像，将目标图像的姿态P_B与I_A 相加并作为生成器 G 的输入，输出为具有目标姿态P_B的合成图像$\hat{I}_B$，小分辨率、原分辨率判别器分别以D_1, D_2指代。

3.4.1 生成对抗网络损失函数

传统生成对抗网络的损失函数表达形式为：

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim data}[logD(x)] + \mathbb{E}_{z\sim noise}[\log(1 - D(G(z))] \tag{3.1}$$

目标是通过学习使得生成器的分布$p_G(x)$ 与真实的数据分布$p_{data}(x)$尽可能一致，GAN 通过生成器网络 G，以噪声数据$z\sim p_{noise}(z)$为输入，生成生成器样本分布p_G，生成器通过与判别器网络 D 对抗训练，对于给定生成器，最优化的判别器应当是 $D(x) = P_{data}(x)/(P_{data}(x) + P_G(x))$ 。

但本文设计的模型不仅于此，借鉴自 InfoGAN，，在整个生成对抗网络的生成

对抗损失函数中将潜码（latent code）纳入了考虑，在最大化学习并保留共有信息（mutual information）的同时，学习多样化的姿态信息。

在这里我们用 c 表示潜码，在加入潜码的情况下，对于生成器而言，形式则变成了 G(z,c)，但在传统的生成器网络中，作为输入的一部分，潜码会对生成图片起到相当干扰作用，破坏原有的结构，故应当寻找一种表示，使得$P_G(x|c) = P_G(x)$, 从另一个角度来说，是寻求一种共同信息（mutual information），使得潜码包含在原有输入中。

在信息论中，X 与 Y 共同信息表达为 I(X;Y)，意为从 Y 中能够学到的关于 X 的信息的多少。共同信息可以表达为两个熵值的差：

$$\mathrm{I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)} \tag{3.2}$$

直觉上来解释，I(X;Y) 是当观察到 Y 时，对 X 的不确定性的减少程度。若 X,Y 完全独立，则 I(X;Y)=0，相反，若为 1，则两者互相间有着直接的决定性的联系。

以上的解释可以使得我们得出这样的代价函数：对于给定的$\mathrm{x}\sim P_G(x)$，则要使$P_G(c|x)$有着尽可能低的熵值。换句话说，潜码 c 中的信息在生成器的生成过程中不应被丢失。因此，本文中的生成对抗网络的目标函数如下设计：

$$\min_G \max_D V_1(D,G) = \mathrm{V(D,G)} - \lambda \mathrm{I(C;G(Z,C))} \tag{3.3}$$

然而现实中，共同信息项$\mathrm{I(C;G(Z,C))}$很难直接最大化，因为其的优化需要后验概率 P(c|x)。但是，可以定义一个辅助分布 Q(c|x)来逼近 P(c|x)，从而获取一个下界（lower bound）：

$$\begin{aligned}
&\mathrm{I\big(C;G(Z,C)\big) = H(c) - H(c|G(z,c))} \\
&= \mathbb{E}_{x\sim G(z,c)}\Big[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|X)]\Big] + H(c) \\
&= \mathbb{E}_{x\sim G(z,c)}\Big[D_{KL}(P(\,) \parallel Q(|x)) + \mathbb{E}_{c'\sim P(c|x)}[\log P(c'|X)]\Big] + H(c) \\
&\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|X)]] \quad + H(c)
\end{aligned} \tag{3.4}$$

因此，生成对抗网络部分的损失函数表示为：

$$\mathcal{L}_{GAN} = \mathbb{E}_{\mathrm{I_A}\sim p_{data}(\mathrm{I_A})}[\log D_1(\mathrm{I_A})] + \mathbb{E}_{\mathrm{I_A}\sim p_{data}(\mathrm{I_A})}[\log(1 - D_1(\mathrm{G(I_A|P_B)}))] \tag{3.5}$$

$$\mathcal{L}^S_{GAN} = \mathbb{E}_{\mathrm{I_A}\sim p_{data}(\mathrm{I_A})}[\log D_2(\mathrm{I_A})] + \mathbb{E}_{\mathrm{I_A}\sim p_{data}(\mathrm{I_A})}[\log(1 - D_2(G_{mid}(\mathrm{I_A|P_B})))] \tag{3.6}$$

3.4.2 L1 距离损失函数

基于 L1 距离的损失函数用于衡量并惩罚生成图片和目标图片的差异，两个分辨率下的损失函数表达形式分别为：

$$\mathcal{L}_{L1} = \|G(I_A, P_B) - I_B\|_1 \tag{3.7}$$

$$\mathcal{L}_{L1}^{S} = \|C_S(G_{mid}(I_A, P_B)) - I_B^S\|_1 \tag{3.8}$$

其中，$G_{mid}(I_A, P_B)$是生成器中间层的输出，C_S是将生成器中间层生成为小分辨率图像的小型卷积神经网络。$\mathcal{L}_{L1}$为原分辨率图像的 L1 损失函数，$\mathcal{L}_{L1}^{S}$为小分辨率图像的 L1 损失函数。

不同于传统机器学习方法中使用的 L2 作为距离度量，L1 距离在生成对抗网络中使用的更为广泛，原因是其能够更好地作为图片质量度量的情况下，鼓励生成器生成边缘更加锐利的图片，从而尽可能的保留生成图片的边缘信息。

直观上说，就是直接将生成图片与原图进行像素值直接进行差值运算，从而引导训练尽可能与原图接近。

与传统的利用 L1 距离的判别器不同，本文利用了两个判别器D_1和D_2，分别对应中间层特征信息通过卷积生成的小分辨率图片，和原图大小的生成器的最终生成，在这两个部分的 L1 信息有着不同的意义和作用：

① 相比于原分辨率而言，低分辨率更多的压缩了底层信息、细节信息，而保留了高层信息、结构信息，故小分辨率的判别器强化了对高层信息的学习，而大分辨率对细节信息、底层信息的学习进行了强化；

② 从编码器-解码器结构角度而言，随着卷积层的加深，每一层卷积层的输出越来越向高层信息靠近，而解码器部分的反卷积则可以认为是卷积运算的逆过程，故浅层位置的反卷积是在基于高层信息进行解码、上采样，而反卷积层数越深，则越偏向底层信息，而两个基于 L1 距离的损失函数刚好与反卷积层不同位置的对低层、高层信息的学习对应起来。

3.4.3 图片块损失函数

传统判别器的损失函数一般基于传统机器学习方法中的二分类问题，即分类结果是离散的 0-1 分布，但由于生成器的生成图片的质量十分有限，而判别器由于卷积神经网络强大的特征提取能力，很容易通过个别细节判定生成的图片为假，而离散的 0-1 分布在反向传播算法中不能很好地鼓励生成图片的质量，故本文提出的模型的判别器创造性地使用了基于 patch 的损失函数

所谓 patch 的产生，是基于卷积神经网络的特性决定的，对于相邻两层，通过大小为 3*3 的卷积核生成的后的一个像素则对应上一层的一个 3*3 的 patch，若再通过一层步长为 2 的 3*3 卷积进行卷积运算，则新生成特征向量中的一个特征值则对应输入图片中的一个 5*5 的 patch。

在本文实现的判别器模型中，Patch 是基于判别器最后一层的输出进行判定，由于卷积神经网络的特征，最后一层的每一个特征值，基于感受野的原理，对应着原图中的一个 patch，以原图大小（128*64 像素）作为输入的判别器中，最后一层的特征向量中每一个特征值对应着 7*7 的 patch。

对于每一个 patch，通过对原图和生成图片对应位置的特征值判定，生成一个结果为 0-1 分布的结果，然后根据 patch 的分类结果生成一个连续的值，从而在反向传播算法中能够得将基于目标函数的反馈反向传播，从而很好地鼓励生成器生成的更高图片质量的图片。

基于 patch 的损失函数表达形式为：

$$\mathcal{L}_{patch} = \frac{1}{N}\sum_{i=0}^{n}\left(\frac{\mathcal{G}_{\hat{\mathrm{I}}_B,i} - \mathcal{G}_{I_B,i}}{H'W'}\right)^2 \tag{3,9}$$

$$\mathcal{L}_{patch}^{s} = \frac{1}{N}\sum_{i=0}^{n}\left(\frac{\mathcal{G}_{\hat{\mathrm{I}}_B^s,i} - \mathcal{G}_{\mathrm{I}_B^s,i}}{H''W''}\right)^2 \tag{3.10}$$

$\mathcal{G}_{\hat{\mathrm{I}}_B,i}$表示图片$\hat{\mathrm{I}}_B$中的第 i 个 patch，并将原图与生成图像的对应位置的 patch 进行对比，H,W 表示的是当前图像的基于 patch 的高度和宽度，$\mathcal{L}_{patch}$和$\mathcal{L}_{patch}^{s}$分别表示原分辨率下、小分辨率下的基于 patch 的损失函数。

3.4.4 交叉熵损失函数

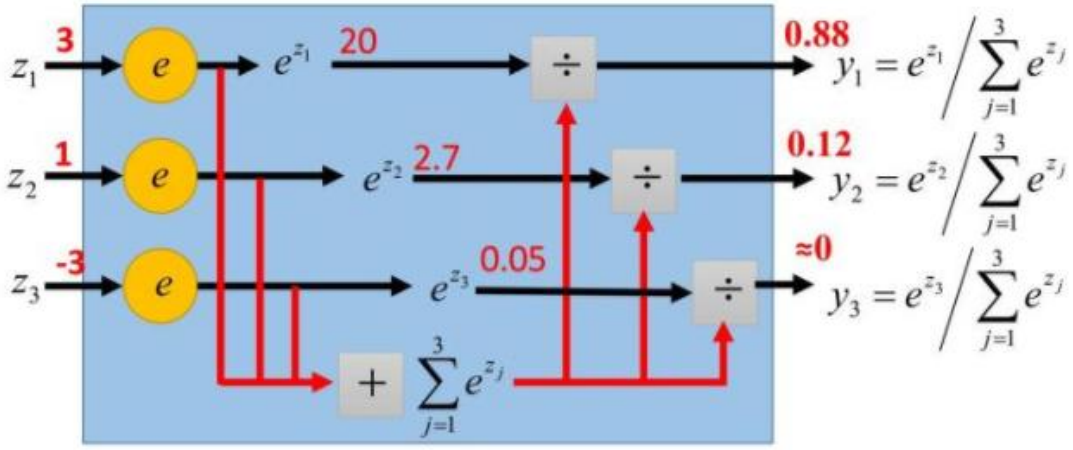


图 3.10 softmax 分类的示意图

在生成器网络的瓶颈处，本文设计了一个分类器网络，基于每个个体的 ID 进行多分类，在这里，分类器的设计模型基于 softmax 分类层进行训练。

$$\mathrm{z} = \Psi(m_{bottleneck}) \tag{3.11}$$

此处，z 表示为在瓶颈处的卷积神经网络Ψ的输出，$m_{bottleneck}$则是在生成器瓶颈处的特征向量。

传统的 one-hot 编码，给预测的 label 赋值为 1，其余赋值为 0，这样虽然非常符合直觉，有着诸多缺点：

① 离散的输出不符合神经网络的反向传播算法，不能很好地将损失函数的差值进行反向传播；

② One-hot编码一定程度上不符合典型的概率分布，即每次只预测一个结果，而忽略其余 ID 的可能性

基于以上 one-hot 的缺陷，本文的多分类器使用了基于 softmax 的多分类层。

Softmax 是将多个神经元的输出，映射到（0,1）区间内，从而进行多分类。设 softmax 层前的输出为一个向量 V，v_j代表 v 中第 i 个元素，则该元素的 softmax 值为：

$$y_j = \frac{e^{v_j}}{\sum_{i=0}^{n} e^{v_i}} \tag{3,12}$$

因此，基于 softmax 的分类器的损失函数为：

$$\mathcal{L}_{softmax} = \sum_{i=0}^{n} -z_i \log(y_i) \tag{3.13}$$

3.4.5 总目标函数

基于以上，本模型的总的损失函数表达为：

$$\mathcal{L} = \mathcal{L}_{GAN} + \mathcal{L}_{GAN}^{s} + \mathcal{L}_{L1} + \lambda \mathcal{L}_{L1}^{s} + \mathcal{L}_{patch} + \mathcal{L}_{patch}^{s} + \mathcal{L}_{softmax} \tag{3.14}$$

3.5 本章小节

在本章中，我们提出了一种基于生成对抗网络的行人图像合成网络模型。通过在中间层、瓶颈处插入卷积神经网络的方式在图像生成过程中添加针对个体的、高低层信息的、类间差异的有效地监督及对应的损失函数，有效地缩小了生成器的解空间，使得生成对抗网络训练更加平稳，生成高质量的多姿态的行人图片。

4 实验与分析

4.1 网络架构

如表 4.1 为网络的编码器结构，可以看到，先进行边缘填充，由 个卷积层，和 个残差模块组成，其中每个卷积层后跟随者一层批正则化层和 ReLu 激活层。具体参数如表格所示。

每个残差模块两个卷积层，两个批正则化层，一个激活层，一个边缘填充层组成，与前面卷积层不同的是，残差模块中的卷积层的卷积核大小为 1x1。

表 4.1 编码器网络结构

层名称	具体细节与参数	输入通道数	输出通道数
边缘填充	大小：3	4	4
卷积层	卷积核大小=(7, 7), 步长=(1, 1), 无偏置， 输入：输出通道数=4:64	4	64
批正则化层	分母系数=1e-05, 动量=0.1, 无仿射变换参数	64	64
ReLU 激活层		64	64
卷积层	卷积核大小=(3, 3), 步长=(2, 2), 边缘填充=(1, 1), 无偏置	64	128
批正则化层	分母系数=1e-05, 动量=0.1, 无仿射变换参数	128	128
ReLU 激活层		128	128
卷积层	卷积核大小=(3, 3), 步长=(2, 2), 边缘填充=(1, 1), 无偏置	128	256
批正则化层	分母系数=1e-05, 动量=0.1, 无仿射变换参数	256	256
ReLU 激活层		256	256
残差模块 1	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256
残差模块 2	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256

残差模块 3	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256
残差模块 4	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256
残差模块 5	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256
残差模块 6	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256
残差模块 7	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256
残差模块 8	卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数) ReLU 激活层 边缘填充层((1, 1, 1, 1)) 卷积层(256, 256, 卷积核大小=(3, 3), 步长=(1, 1), 无偏置) 批正则化层(256, 分母系数=1e-05, 动量=0.1, 无仿射变换参数)	256	256

如表 4.2，是生成器中解码器的网络架构与具体参数，反卷积层与编码器中残差模块外的的卷积层一一对应，包括每一层的具体参数，步长、卷积核等。

表 4.2 解码器网络架构：

层名称	具体细节与参数	输入通道数	输出通道数
反卷积层	卷积核大小=(3, 3), 步长=(2, 2), 边缘填充=(1, 1),边缘填充=(1, 1), 无偏置	256	128
批正则化层	分母系数=1e-05, 动量=0.1, 无仿射变换参数	128	128
ReLU 激活层		128	128
反卷积层	卷积核大小=(3, 3), 步长=(2, 2), 边缘填充=(1, 1), output_边缘填充=(1, 1), 无偏置	128	64
批正则化层	分母系数=1e-05, 动量=0.1, 无仿射变换参数	64	64
ReLU 激活层		64	64
边缘填充层	(3, 3, 3, 3)	64	64
卷积层	卷积核大小=(7, 7), 步长=(1, 1)	64	3

如表 4.3，为用于生成小分辨率图片的卷积神经网络，在具体实现中，这个卷积神经网络的输入是在解码器中第一个反卷积层的输出的特征向量，经过对比验证，使用 tanh 激活函数的效果最好。

表 4.3 小分辨率图片生成卷积网络

层名称	具体细节与参数	输入通道数	输出通道数
卷积层	卷积核大小=(1, 1), 步长=(1, 1)	128	3
Tanh 激活层		3	3

如表 4.4，为用于监督行人个体特征的基于卷积神经网络的多分类器，先以卷积层调整特征向量的尺寸，然后以全连接层生成 1 x 739 的特征向量，739 维的原因是 Market-1501 数据集的训练集中的行人个体数量为 739.

表 4.4 基于 softmax 的多分类器卷积网络

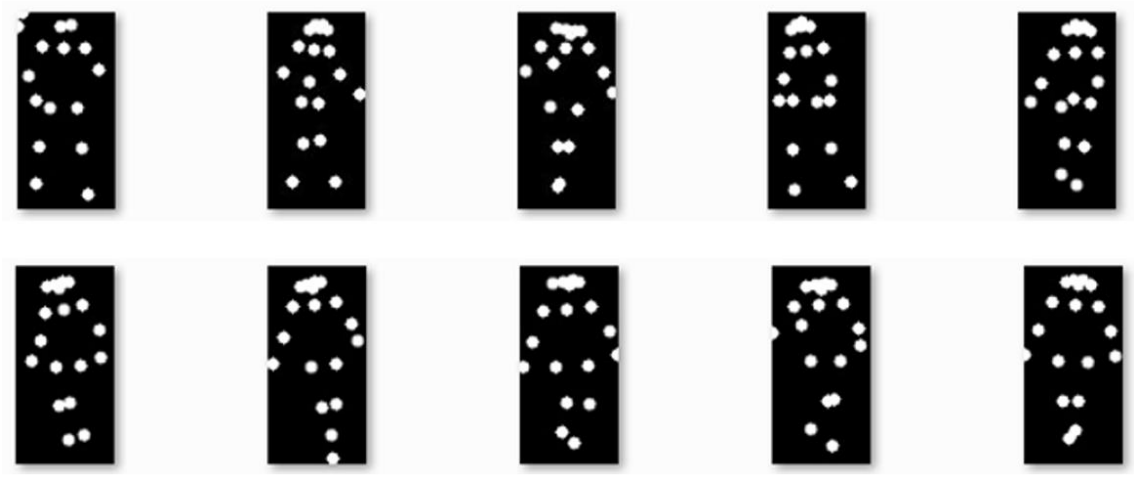
层名称	具体细节与参数	输入通道数	输出通道数
卷积层	卷积核大小=(3, 3), stride=(2, 2), 边缘填充=(1, 1),无偏置	256	512
卷积层	卷积核大小=(3, 3), stride=(2, 2), 边缘填充=(1, 1),无偏置	512	1024
自适应全局池化层	输出大小：1x1	1024	1024
全连接层		1024	739

图 4.2 Market-1501 数据集中的样本示例

4.2 数据集

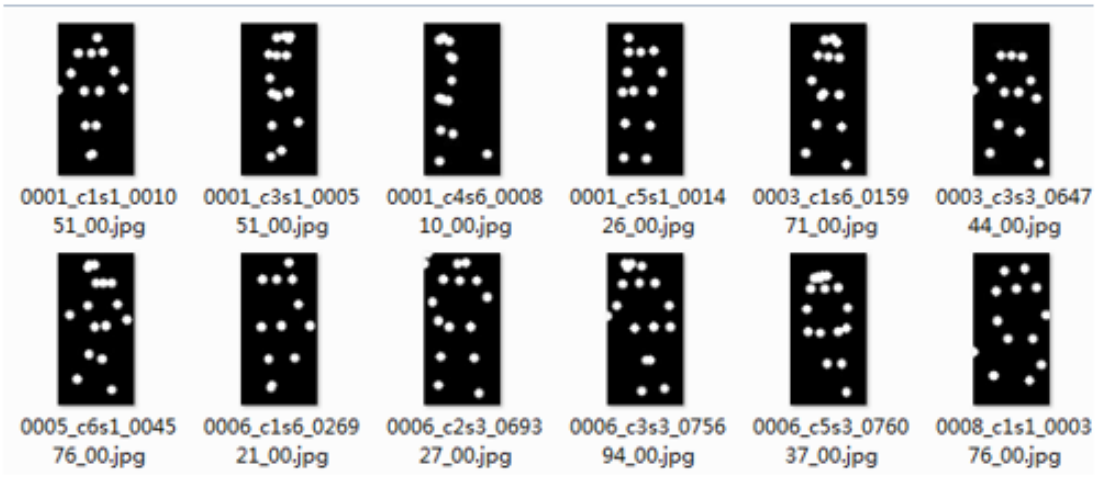
模型的训练和测试都是在当前最大的行人重识别任务数据集 Market-1501 上，如图 4.2，其中包含来自六台分离的监视摄像机捕获的 1501 人的 32688 张行人图像，此数据集中的行人在姿态、照明、视角和背景方面各不相同，从而使得生成新的行人图像极具挑战性，所有图像的尺寸为 128*64， 并分成分别为 12936 张、19732 张的训练集与测试集。

模型的训练并非完全基于以上所有的图片，由于 OpenPose 库并非对每一张图片的每一个关键点都能做到完美地提取，故我们筛选出了能够检测数 14 个关键点以上的共 4641 张图片作为训练数据集，并在同一个训练行人的 ID 下，对不同的姿态的属于同一行人的图片进行组合，形成了共 58706 对的训练数据集。



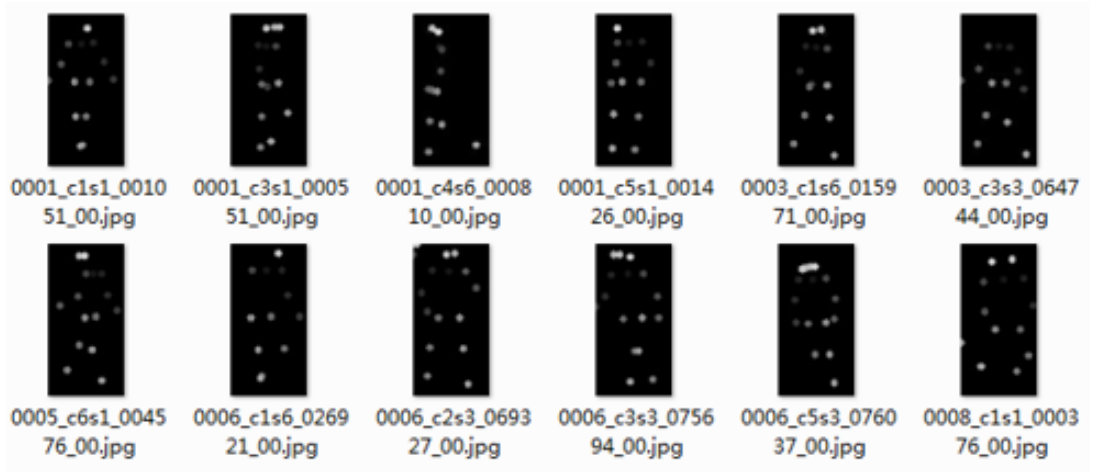
4.2 用于测试阶段的十个模板姿态

在测试过程中，我们从测试数据集中随机选取 10 个能够完好检测到所有身体关键点的姿态作为模板姿态，然后从测试集随机选取 200 张图片，每一张图片分别和模板姿态中的每一个作组合并输入生成器，即对应每个测试集中的图片，生成 10 个不同姿态的图片。如图 4.2，为用于测试阶段的基于点信息的姿态模板。



(a)

(b)



(c)

图 4.3 不同姿态信息表示示例,(a):等值原点表示，(b):腿部连接表示，(c):不等值原点表示

在潜码的输入形式上，我们进行了多种尝试，如图 4.3，为起中的示例。

如（a），是算法模型中输入的潜码的最终表示形式，人身体的 18 个关节点包含：

鼻子、颈、左肩膀、左手肘、左手腕、左肩膀、右手肘、右手腕、左臀部、左膝盖、左脚踝、右臀部、左膝盖、右脚踝、左眼睛、右眼睛、左耳朵、右耳朵、背景

以纯黑色为背景，以半径为 5 的白色圆圈标注以上 18 个关节点。

（b）尝试将腿部连接，以增强腿部信息的学习，但实验证明，连接线起到了较强的干扰作用，随符合直觉，但不能很好地适用于卷积神经网络。

（c）尝试用不同的灰度值赋值关键点，给予不同图像间相同关节点赋值相同，并且不同关节点间灰度值不同，实验证明，效果仍不理想，原因是关节点的颜色越深，即关节点的灰度越接近黑色，由于难以与背景区分，导致此类关节点很难学习到。

4.3 实验设置

硬件环境为:

CPU: Intel Core i7-5820K CPU@3.30GHz x12

内存：128G

GPU：4 x NVIDIA GeForce TITAN Xp

硬盘：12TB

操作系统：64 位 Ubuntu 14.04

开发语言为 python，框架为 PyTorch1.0，CUDA 版本为 8.0，CuDNN 版本为 5.0。

4.4 实验结果分析

接下来，基于训练后的生成对抗网络模型，在测试阶段，随机抽取测试数据集中的图片和模板姿态作为输入，生成了以下结果，在每组展示样本中，最左侧为输入的原图，右侧十张为生成的对应于图 4.2 的十个姿态模板的生成图像。

在 Market-1501 数据集中，图像分辨率为 128*64，故生成图像分辨率也为 128*64。

图 4.2 生成图像示例，最左侧为输入图片，右侧为合成的十个不同姿态的行人图像

如图 为实验结果，以上为模型的最终结果，在尽可能保留行人原图片细节的前提下，对行人姿态进行了任意的变换，并且即使在腿部交叉的情况下依然生成了相当自然的图片，边缘锐利、清晰。

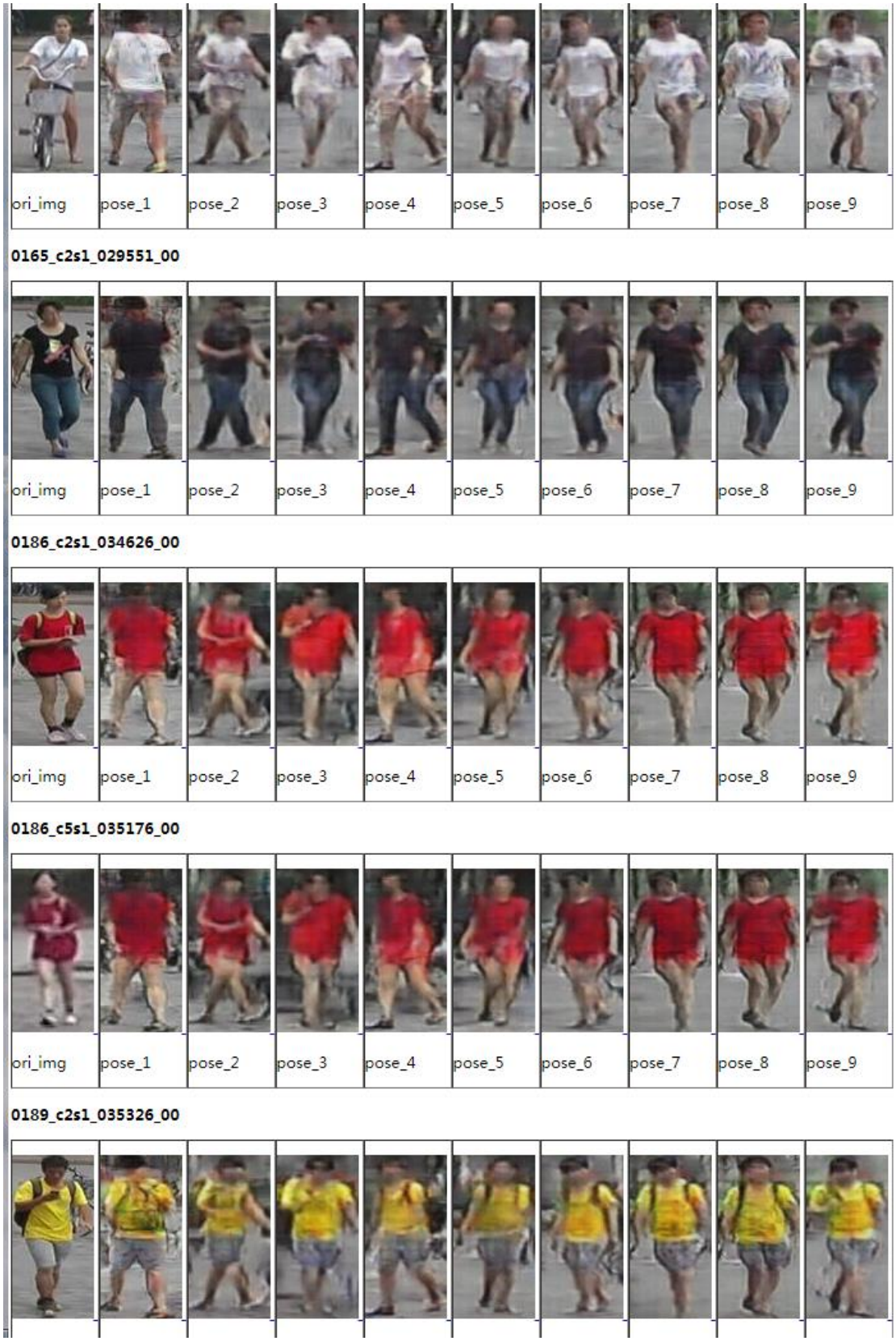


图 4.3 无小分辨率图像监督情况下生成图像示例

此为无小分辨率图片情况下的结果，出现了与目标姿态一定程度的偏离和变形。

图 4.4 无多分类器监督情况下生成图像示例

此为无分类器情况下结果，如图中最下面一列，行人个体信息不能够很好地得到保留，出现了女性转变成男性的倾向。

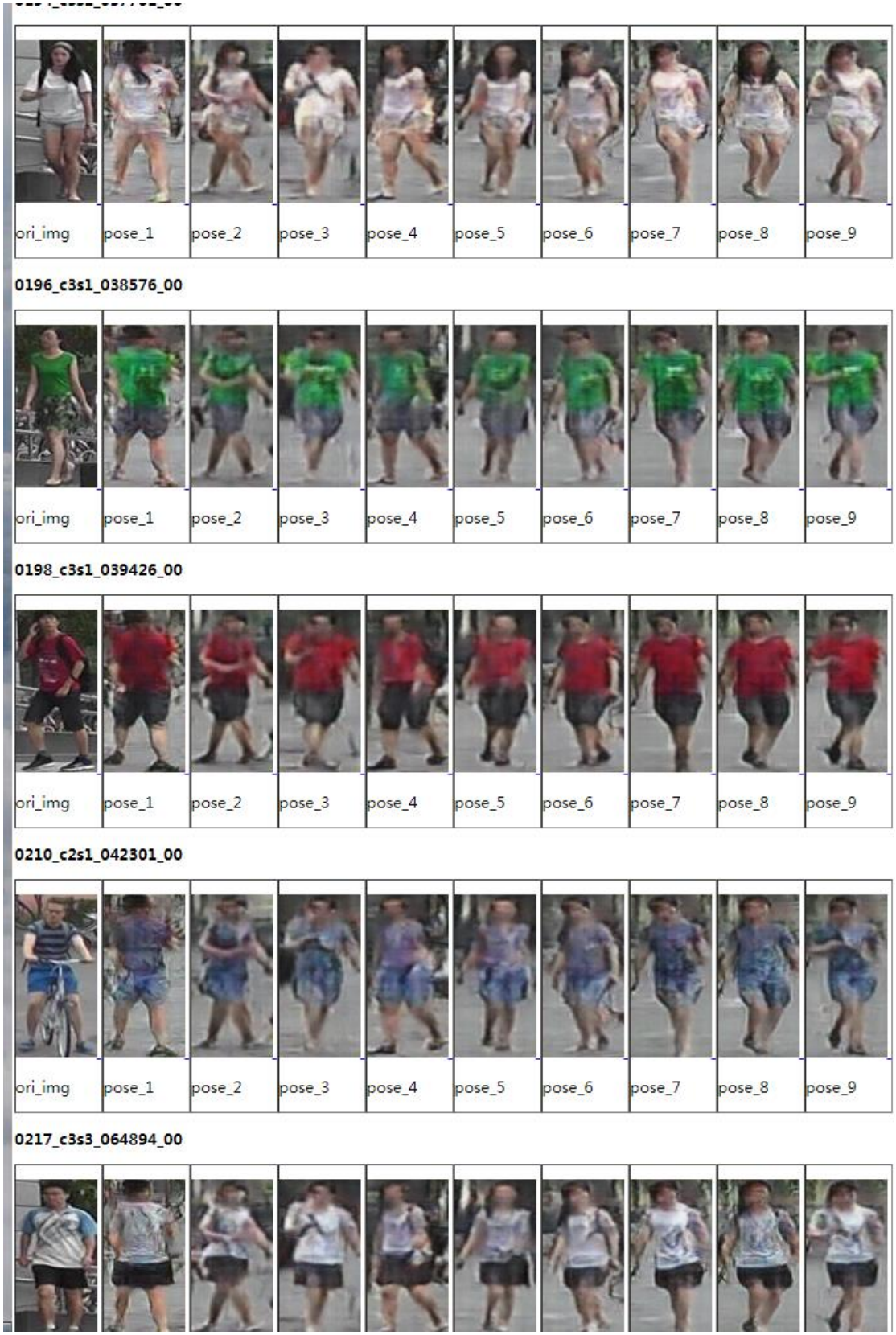


图 4.5 当前领域内最先进方法生成图像示例

此为当前最先进的行人图像合成模型复现后的生成结果，可以看出，细节信息、个体信息都不能得到很好地保留，且边缘模糊。

图 4.6 输入行人状态为骑行时的生成结果

由图示，当输入的行人状态是骑行时，结合以目标的姿态信息，虽然会一定程度上影响生成图片的质量，但依然能够有效、成功地生成基于新的姿态的图片，并且在生成图像中已经消除掉了自行车，可见训练的模型已经具有着相当强的鲁棒性和泛化能力。

4.5 本章小结

在本章中，在实现了模型的基础上，我们在目前最大的行人重识别数据集 Market-1501 上，通过更改模型及输入，进行了大量的实验，并对每一部分网络的改动做出了解释和针对性地对比，验证了众多思路的实现效果，验证了我们的思路的可行性，并证实了我们所设计的模型在个体信息保留、低高层信息一致上皆优于目前的现有所有方案。

5 结论与展望

5.1 结论

本文的主要贡献是：

① 提出了一种新的生成对抗网络模型设计用以进行行人多姿态图像的合成，一方面，借助优化后的生成对抗网络生成了高质量的、逼真的合成图像，可以用于行人重识别任务数据集的数据增光；另一方面，通过生成对抗网络与潜码的结合，进一步发掘了人工神经网络对人体部位、姿态信息的理解和学习，为进一步的应用提供了前提和可能性。

② 为了加强生成后行人图像的个体信息，并保留行人间提取后的特征的差异信息，本文在生成器的瓶颈处创造性地设计了基于 softmax 的多分类器，使得经过编码器卷积运算后提取的特征的有效性。

③ 本文采用在解码器中间层插入卷积神经网络以生成小分辨率图片，小分辨率、原分辨率的合成图片分别对应着高层、低层信息，与原图进行 L1 距离的计算，一方面可以视作在网络不同层次保证图片生成质量的约束；另一方面，通过多分辨率的生成及其对应的判别器，进一步缩小了生成器的解空间，从而提高生成图片的质量，与生成的特征分布与原图数据分布的一致性。

④ 本文摒弃传统的在判别器上的二分类方法，创造性的使用了基于 patch 的损失函数，从根本上解决了当前基于像素的判别器的致命性的短板——过于强的判别器无法促使生成器向生成更高质量图片的方向靠近；从而减缓了判别器的收敛速度，避免了生成器与判别器学习进度的不对等，使得训练更加的平稳。

⑤ 通过一系列监督网络和辅助网络的设计和增益，实现了目前效果最好的基于姿态的行人图像合成模型。

5.2 展望与未来的工作

但在算法的设计和具体实现过程中，我们也发现了诸多问题：

① 生成对抗网络的一大特性——训练的不稳定性，深深的影响着模型的训练和最终的收敛；

② 生成对抗网络过于大的解空间和过于确定的目标函数，使得生成更高的分辨率的图像有着相当高的难度；

③ 针对生成对抗网络，在潜码的发掘上仍不够充分，目前在行人生成图像问题上，仅仅有姿态信息作为潜码而进行利用；

④ 目前生成的行人图像无论是在观感上还是具体的比对上都与原有的真是图像有着相当差距，因此还不能满足数据增强的目的。

未来可能的研究方向：

① 进一步调整生成对抗网络模型的设计或者训练策略，使训练过程更加平缓和稳定；

② 通过生成对抗网络训练来生成更高分辨率的图像；

③ 发掘更多的图像生成领域，基于生成对抗网络模型的潜码以及条件信息；

④ 生成更高质量的行人图像，使之能够真正的用于行人重识别数据集的数据增强。

致谢

首先，感谢葛永新老师在我的毕业设计完成过程中，给予的珍贵的指导和耐心的教导，以及葛永新老师在我的学业生涯中给予的至关重要的指导和帮助。

感谢悉尼科技大学的杨易教授，在我的毕业设计过程中给予的珍贵的指导和帮助，感谢华为分布式计算与并行处理实验室与胡芝兰博士在计算资源和技术上的慷慨且至关重要的支持，感谢郑哲东师兄在我毕业设计完成过程中给予的珍贵的指导。

于我而言，四年本科给我带来的学识倒是其次，最最重要的是让我看到了更大的世界，认识了更多的人，逐渐认识到自己的局限性，从而越来越明确自己是一个怎样的人，以及要成为怎样的一个人，也许这就是所谓的“与这个世界妥协”。

尽管我存在着一些难以消解的缺点及理想主义倾向，但这也正是我本人最为珍贵与闪亮的特征。

感谢三年的每日长跑，不仅仅是强健体魄，更使得我每晚都能有三四十分钟的沉思时间。

感谢叔本华的几本著作，促使我完成在思想上的一次重要的蜕变，塑造了我思想上的脉络。

感谢陈静刘礼等老师对我的支持和指导，感谢魏松霖饶焱竣等朋友在生活上的陪伴和理解。

感谢我的父亲、母亲对我一如既往的资助、支持和鼓励。

参 考 文 献

[1] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551.

[2] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.

[3] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[5] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[6] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 248-255.

[7] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[8] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

[9] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Cvpr, 2015.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[11] Normalization B. Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[13] D. Yi, Z. Lei, S. Liao, S. Z. Li. Deep metric learning for person re-identification[C]. IEEE Conference on International Conference on Image Processing, 2014: 34–39.

[14] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 152–159.

[15] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3908–3916.

[16] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S. Z. Li. Embedding Deep Metric for Person Re-identification: A Study Against Large Variations[C]. European Conference on Computer Vision. Springer, 2016: 732-748.

[17] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification[C]. European Conference on Computer Vision. Springer, 2016: 135-153.

[18] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification[C]. European Conference on Computer Vision. Springer, 2016: 791-808.

[19] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based CNN with improved triplet loss function[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1335–1344.

[20] T. Xiao, H. Li, W. Ouyang, X. Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1249-1258.

[21] M. Geng, Y. Wang, T. Xiang, Y. Tian. Deep Transfer Learning for Person Re-identification. arXiv preprint arXiv: 1611.05244v2, 2016.

[22] Tian M, Yi S, Li S et al Eliminating Background-bias for Robust Person Re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018

[23] Sun Y, Zheng L, Yang Y, et al. Beyond Part Models: Person Retrieval with Refined Part Pooling[J]. arXiv preprint arXiv:1711.09349, 2017.

[24] Xu J, Zhao R, Zhu F, et al. Attention-Aware Compositional Network for Person Re-identification[C]. . IEEE Conference on Computer Vision and Pattern Recognition, 2018

[25] Zhou Q, Fan H, Zheng S, et al. Graph Correspondence Transfer for Person Re-identification[J]. arXiv preprint arXiv:1804.00242, 2018.

[26] Zhong Z, Zheng L, Zheng Z, et al. Camera Style Adaptation for Person Re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018

[27] Deng W, Zheng L, Kang G, et al. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018

[28] Wei L, Zhang S, Gao W, et al. Person Transfer GAN to Bridge Domain Gap for Person Re-

Identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018

[29] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[30] Makhzani A, Shlens J, Jaitly N, et al. Adversarial autoencoders[J]. arXiv preprint arXiv:1511.05644, 2015.

[31] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[32] Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks[J]. arXiv preprint arXiv:1601.06759, 2016.

[33] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]. IEEE international conference on computer vision 2017

[34] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[C]. International Conference on Learning Representations 2016

[35] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[36] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans[J]. arXiv preprint arXiv:1610.09585, 2016.

[37] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[C]. International Conference on computer vision 2017

[38] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017

[39] Mathieu M, Couprie C, LeCun Y. Deep multi-scale video prediction beyond mean square error[J]. International Conference on Learning Representations 2016

[40] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2536-2544.

[41] Huang R, Zhang S, Li T, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis[J]. arXiv preprint arXiv:1704.04086, 2017.

[42] Zhu S, Fidler S, Urtasun R, et al. Be your own prada: Fashion synthesis with structural coherence[C]. International Conference on computer vision 2017

[43] Ma L, Jia X, Sun Q, et al. Pose guided person image generation[C]//Advances in Neural Information Processing Systems. 2017: 405-415.

[44] Zhao B, Wu X, Cheng Z Q, et al. Multi-view image generation from a single-view[J]. arXiv

preprint arXiv:1704.04886, 2017.

[45] Zhang H, Xu T, Li H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//IEEE Int. Conf. Comput. Vision (ICCV). 2017: 5907-5915.

[46] Chidambaram M, Qi Y. Style transfer generative adversarial networks: Learning to play chess differently[C]. International Conference on Learning Representations 2017

[47] Wang T C, Liu M Y, Zhu J Y, et al. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018

[48] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124.

[49] Zhang L, Xiang T, Gong S. Learning a discriminative null space for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1239-1248.

[50] Sun Y, Zheng L, Deng W, et al. Svdnet for pedestrian retrieval[J]. International Conference on computer vision 2017

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

[52] Si J, Zhang H, Li C G, et al. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification[J]. IEEE Conference on Computer Vision and Pattern Recognition, 2018

[53] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2197-2206.

[54] Geng M, Wang Y, Xiang T, et al. Deep transfer learning for person re-identification[J]. arXiv preprint arXiv:1611.05244, 2016.

[55] Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks[C]//Advances in Neural Information Processing Systems. 2017: 700-708.

[56] Chen Q, Koltun V. Photographic image synthesis with cascaded refinement networks[C]//The IEEE International Conference on Computer Vision (ICCV). 2017, 1.

[57] Zhu J Y, Zhang R, Pathak D, et al. Toward multimodal image-to-image translation[C]//Advances in Neural Information Processing Systems. 2017: 465-476.

[58] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. International Conference on computer vision 2017

[59] Cheung B, Livezey J A, Bansal A K, et al. Discovering hidden factors of variation in deep networks[J]. arXiv preprint arXiv:1412.6583, 2014.

[60] Ma L, Sun Q, Georgoulis S, et al. Disentangled Person Image Generation[C]. International Conference on computer vision 2017

[61] Moeslund T B, Hilton A, Krüger V. A survey of advances in vision-based human motion capture and analysis[J]. Computer vision and image understanding, 2006, 104(2-3): 90-126.

[62] Reed S E, Akata Z, Mohan S, et al. Learning what and where to draw[C]//Advances in Neural Information Processing Systems. 2016: 217-225

[63] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//CVPR. 2017, 1(2): 7..

[64] Reed S, van den Oord A, Kalchbrenner N, et al. Generating interpretable images with controllable structure[J]. 2016.

[65] Lassner C, Pons-Moll G, Gehler P V. A generative model of people in clothing[J]. arXiv preprint arXiv:1705.04098, 2017.