附件5C：译文

指导教师评定成绩

(五级制)：

指导教师签字：

姿态引导的行人图像合成

# 摘要

本文提出了一种新颖的姿态导引人生成网络（PG2），该网络允许基于该人的图像和新姿势来以任意姿势合成人物图像。 我们的生成框架PG2明确地利用了姿态信息，并由两个关键阶段组成：姿态整合和图像细化。 在第一阶段，条件图像和目标姿态被馈送到类似U-Net的网络中以生成具有目标姿势的人的初始但略粗糙的图像。 然后，第二阶段通过以对抗方式训练类似U-Net的结构来提升和细化第一阶段生成的模糊的图像。 在分辨率为128\*64的行人重识别数据集和256\*256的时尚衣物数据集上进行的大量实验证明，我们的模型能够生成具有令人信服的细节的高质量人物图像。

# 1 引言

生成逼真的图像对于诸如面部编辑，电影制作和基于合成图像的图像检索等许多应用而言具有很大的价值。 因此，已经提出了广泛的方法，包括变分自编码器（VAE）[14]，生成对抗网络（GAN）[6]和自回归模型（例如，PixelRNN [30]）。 最近，GAN模型由于其通过对抗训练产生清晰图像的原理能力而特别受欢迎。 例如在[21,5,1]中，GAN被用来生成人脸和自然场景图像，并提出了几种方法来稳定训练过程并提高生成质量。

从应用的角度来看，用户通常具有特定的意图，例如更改背景，对象的类别，颜色或视角。我们方法的关键思想是通过适当表示该意图来明确引导生成过程，以实现对生成过程的直接控制。更具体地说，我们建议通过在参考图像和指定姿势上对其进行调节来生成图像。以参考图像作为条件，该模型预先具有关于期望对象的外观的足够信息。预期姿势给出的指导既明确又灵活。所以原则上这种方法可以将任何对象操纵为任意姿势。在这项工作中，我们专注于将一个人从一个给定的姿势转移到一个预期的姿势。这项任务有许多有趣的应用。例如，在电影制作中，我们可以直接操纵角色的人体到所需的姿势，或者，对于人体姿势估计，我们可以生成具有罕见但重要姿势的训练数据。

将一个人从一个姿势转移到另一个姿势是一项艰巨的任务。图1中可以看到一些例子。完整的端到端框架很难做到这一点，因为它必须同时生成正确的姿势和详细的外观。因此，我们采用分治策略，将问题分为两个阶段，分别关注学习整体人体结构和外观细节，分别类似于[35,9,3,19]。在阶段I，我们探索不同的方式来模拟姿势信息。采用U-Net的变体将目标姿势与人物图像进行整合。它输出粗糙的生成结果，捕捉目标图像中人体的全局结构。为了抑制条件图像与目标图像之间背景变化的影响，提出了一种基于蒙版的L1损失。但是，由于使用L1，它会产生模糊的结果。在阶段II，使用Deep Convolutional GAN（DCGAN）模型的变体来进一步优化初始生成结果。该模型通过对抗训练学习填充更多外观细节，并生成更清晰的图像。与通常使用直接学习从头开始生成图像的GAN不同，本文中我们训练GAN以生成初始生成结果与目标人员图像之间的差异图。训练收敛速度更快，因为这是一项更容易的任务。此外，我们添加蒙版的L1损失来规范生成器的训练，使其不会生成具有许多伪影的图像。对两个数据集（一个低分辨率行人重新识别数据集和一个高分辨率时尚照片数据集）的实验证明了该方法的有效性。

我们的贡献有三点。 i）我们提出了一种新的任务，即对参考图像和预期姿势进行图像生成调节，其目的是将图像中的人物操控为任意姿势。 ii）探索将姿势信息与人物图像集成的几种方法。 提出一种新颖的蒙版损失，以鼓励该模型专注于转移人体外观而不是背景信息。 iii）为解决姿势转移的挑战性问题，我们将问题分为两个阶段，第一阶段以人体全局结构为重点，第二阶段为基于第一阶段结果填充外观细节。

# 2 相关工作

最近有很多关于深度学习技术的生成图像建模的工作。 这些作品分为两类。 第一行作品遵循无监督的设置。 在这种情况下一种流行的方法是由Kingma和Welling [14]和Rezende等人提出的变分自动编码器。 [25]，它应用了重新参数化技巧来最大化数据可能性的下限。 方法的另一个分支是autogressive模型[28,30,29]，它以逐个像素的方式计算像素的条件分布的乘积，作为图像中像素的联合分布。 最常用的方法是生成对抗网络（GAN）[6]，它们同时学习生成器生成样本和鉴别器以区分生成的样本与真实样本。 许多作品表明，由于使用对抗性损失而不是L1损失，GAN可以生成清晰的图像。在这项工作中，我们还在我们的框架中使用对抗性损失，以便在图像中生成高频细节。

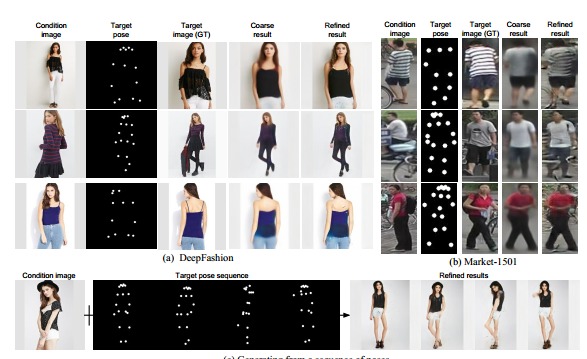
第二类工作生成以类别或属性标签，文本或图像为条件的图像。 Yan et al。 [32]提出了一种条件变分自动编码器（CVAE）来实现属性条件图像生成。 Mirza和Osindero [18]提出在边信息条件下对GAN的发生器和判别器进行条件处理，以执行类别条件图像生成。拉斯纳等人[15]通过对细粒身体部位进行调节，在衣服中产生全身的人。里德等人提出通过向生成器和鉴别器添加文本信息来生成以文本描述为条件的鸟的图像[24]，并进一步探索使用附加位置，关键点或分割信息来生成图像[22,23]。只有这些视觉线索作为条件并与我们对预期姿势的明确条件形成对比，对图像生成过程施加的控制仍然是抽象的。一些作品不仅在标签和文本上还在图像上进一步调节图像生成。研究人员[34,33,11,8]解决了以参考图像和特定脸部视点为条件的面部图像生成任务。陈等人[4]将看不见的推理解决为张量完成问题，并利用潜在因素在不可见的观点中推测姿态。赵等人[36]探讨了仅从单一视图输入生成多视图布料图像，这与我们的任务最相似。然而，各种各样的姿势与任何给定的观点相一致，使得调节不如我们的工作有表现力。在这项工作中，我们以更加明确和灵活的方式利用姿势信息，即使用关键点格式的姿势来模拟多样的人体外观。应该注意的是，我们不使用昂贵的姿态注释，而是使用最先进的姿态估计方法来获取所需的人体关键点。

图1：在DeepFashion数据集[16]（a）（c）和Market-1501数据集[37]（b）上生成的样本。

# 3 提出的算法

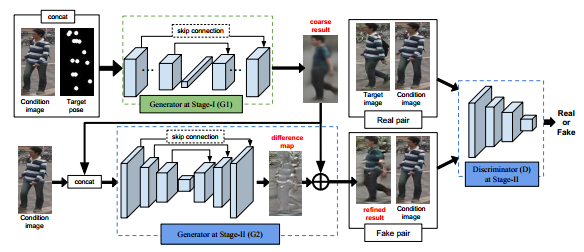
我们的任务是将一个人的外观从一个给定的姿势同时转移到一个理想的姿势，并保留身份的重要外观细节。 由于以端到端模式实施这一项目具有挑战性，因此我们提出两阶段方法来解决这一任务，每个阶段都侧重于一个方面。 对于第一阶段，我们提出并分析了几个模型变体，第二阶段我们使用有条件的DCGAN变体来填充更多的外观细节。 图2显示了我们提出的的姿势引导个人生成网络（PG2）的总体框架。

图2：我们的姿势指导人生成网络（PG2）的总体框架。 它包含两个阶段。 第一阶段侧重于姿势整合，并产生捕捉人体全局结构的初始结果。 第二阶段侧重于通过对抗训练来优化初始结果并生成更清晰的图像。

3.1 阶段1：姿态整合

在阶段I，我们将调节人员图像*IA*与目标姿势*PB*相结合，以生成捕捉目标图像*IB*中的人体的全局结构的粗略结果*I*ˆ*B*。

姿势嵌入。 为了避免昂贵的姿势注释，我们应用最先进的姿态估计器[2]来获得近似的人体姿势。 姿态估计器生成18个关键点的坐标。 将这些直接用作我们模型的输入将需要模型学习将每个关键点映射到人体的一个位置。 因此，我们将姿势*PB*编码为18个热力图。 每个热力图在相应关键点周围4个像素的半径中填充1，其他位置填充0（参见图3，目标姿态）。 我们连接IA和PB作为我们模型的输入。 这样，我们可以直接使用卷积层来整合这两种信息。

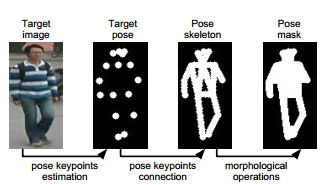
生成器G1。作为阶段I的发生器，我们采用类似于U-Net的体系结构[20]，即具有跳跃连接的卷积自动编码器，如图2所示。具体来说，我们首先使用几个叠层卷积层将IA和PB从小的局部区域集成到较大的区域，从而可以将外观信息集成并转移到相邻的身体部位。然后，使用全连接层，使得身体部位之间的信息也可以交换信息。之后，解码器由一组堆叠的卷积层组成，这些卷积层与编码器对称以生成图像。第一阶段的结果表示为*I*ˆ*B1*。在U-Net中，编码器和解码器之间的跳跃连接有助于直接从输入到输出传播图像信息。另外，我们发现使用残差块作为基本组件可以提高生成性能。具体而言，我们建议简化原始残差块[7]，并且在残差块内只有两个连续的卷积层-RELU激活层结构。

图3：计算姿态蒙版的过程

姿态蒙版损失。 为了将*I*ˆ*B1*与目标图像IB进行比较，我们采用L1距离作为阶段I的生成损失。 然而，由于我们只有一个条件图像和一个目标姿势作为输入，因此如果目标图像与条件图像背景不同，那么模型很难生成背景的样子。 因此，为了减轻背景变化的影响，我们添加了另一个项，其将增加姿势蒙版MB给L1损失，使得人体被赋予比背景更多的权重。 公式1给出了姿态掩模损失的表达式，表示像素方式的乘法：

(1)

姿势蒙版MB被设置为1用于前景，0被用于背景，并且通过连接人体部位并应用一组形态学操作来计算，使得它能够近似覆盖目标图像中的整个人体，参见示例图3.

G1的输出是模糊的，因为L1损失促使结果成为所有可能情况的平均值[10]。 但是，G1确实捕获了由目标姿态指定的全局结构信息（如图2所示）以及衣服颜色等其他低频信息。 身体外观的细节，即高频信息，将通过对抗训练在第二阶段完善。

3.2 阶段2：图像细化

由于第一阶段的模型已经合成了一幅粗糙但靠近姿势和基本色彩的目标图像的图像，在第二阶段，我们希望模型通过纠正错误或缺失来生成更多细节 在最初的结果。 我们使用条件DCGAN [21]的变体作为我们的基础模型，并在第一阶段的结果中对其进行调整。

生成器G2。 考虑到初始结果和目标图像在结构上已经相似，我们建议在第二阶段生成器G2的目的是生成一个外观差异图，使初始结果更接近目标图像。 使用类似于第一阶段的U-Net计算差异图，但是初始结果IˆB1和条件图像IA作为输入。 不同之处在于完全连接层从U-Net上移除。 这有助于从输入中保留更多细节，因为完全连接的层压缩输入中包含的大量信息。 差异图的使用加快了模型训练的收敛速度，因为模型重点在于学习缺少的外观细节，而不是从头开始合成目标图像。 特别是，训练已经从合理的结果开始。 G2的整体架构如图2所示。

鉴别器D.在传统的GAN中，鉴别器区分真实的实地图像和假生成的图像（由随机噪声生成）。 然而，在我们的条件网络中，G2将条件图像IA而不是随机噪声作为输入。 因此，真实图像不仅是自然的，而且也满足特定的要求。 否则，G2将被误导以直接输出IA本身是自然的，而不是精炼第一阶段IˆB1的粗略结果。 为了解决这个问题，我们将G2输出与条件图像配对以使鉴别器D识别对的方位，即（IˆB2，IA）对（IB，IA）。 这在图2中示出。 成对输入鼓励D学习IˆB2和IB之间的区别，而不仅仅是区分合成图像和自然图像。

与传统GAN的另一个不同之处在于噪声不再是必需的，因为发生器是以图像IA为条件的，这类似于[17]。 因此，鉴别器D和发生器G2分别具有以下损失函数，

(2)

(3)

其中bce表示二元交叉熵损失。 以前的工作[10,17]表明，混合对抗损失与损失最小Lp距离可以规范图像生成过程。 在这里，我们使用与第一阶段相同的蒙版L1损失，使得它比背景更注重目标人体的外观，

（4）

其中λ是L1损失的权重。 它控制着生成图像在低频时看起来像目标图像的接近程度。 当λ小时，对抗性损失主导着训练，并且更可能产生伪影; 当λ很大时，具有基本L1损失的发电机在训练中占主导地位，使得整个模型产生模糊的结果2。

在我们的DCGAN的训练过程中，我们交替地优化鉴别器D和发生器G2。 如图2的左侧部分所示，生成器G2将第一阶段结果和条件图像作为输入，目的是细化图像以混淆鉴别器。 鉴别器学习将一对状态图像和生成的图像分类为假，同时将包括目标图像的对分类为真实。

3.3 网络架构

我们总结了所提出的模型PG2的网络结构。在阶段I，G1的编码器由N个残余块和一个完全连接层组成，其中N取决于输入的大小。每个残差块由两个卷积层组成，其中步长等于1，接着是除最后一个块之外的一个子采样卷积层，其中步长等于 2。在阶段II，G2的编码器具有包括N-2个卷积块的完全卷积体系结构。每个块由步长等于1的两个卷积层和一个步长等于2的子采样卷积层组成。 G1和G2中的解码器与对应的编码器对称。此外，解码器和编码器之间存在跳跃连接，如图2所示。在G1和G2中，不应用批处理标准化。所有卷积层由3\*3的滤波器组成，并且每个块的滤波器数量线性增加。我们将整流线性单元（ReLU）应用于除完全连接层和输出卷积层以外的各层。对于鉴别器，由于不同的图像分辨率，我们采用与DCGAN [21]相同的网络架构，但输入卷积层的大小不同。

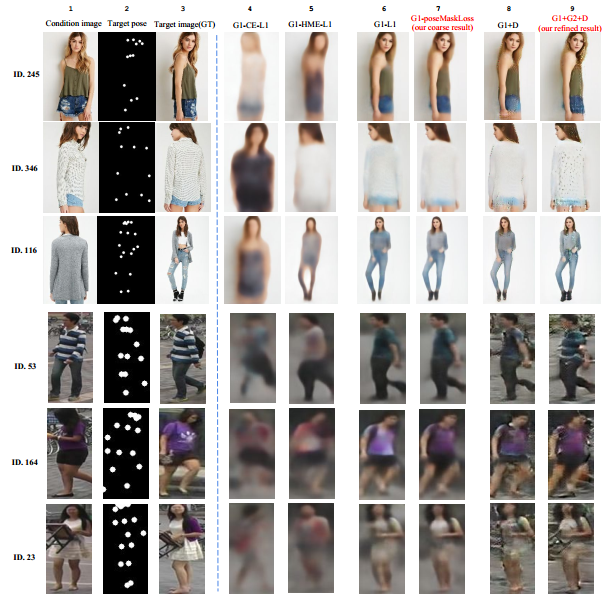
# 4 实验

我们在两个人数据集（Market-1501 [37]和DeepFashion [16]）上评估建议的PG2网络，其中包含具有不同姿势的人员图像。 我们提出PG2的三个主要方面的定量和定性结果：不同的姿态嵌入; 造成掩模损失与标准L1损失; 和两阶段模式与一阶段模式。 我们还将其与最相关的工作进行比较[36]。

4.1 实验结果

DeepFashion数据集[16]由52,712个店内服装图像和20万个交叉/姿势/对比图组成。所有图像的高分辨率为256\*256。在训练组中，我们有146,680对，每对包含同一人的两幅图像，但姿态不同。我们从测试集中随机选择12,800对进行测试。

我们还在一个更具挑战性的重新识别数据集Market-1501 [37]上进行实验，其中包含来自六台分离监视摄像机捕获的1501人的32,668幅图像。此数据集中的人员在姿势，照明，视角和背景方面各不相同，这使得人员生成任务更具挑战性。所有图像的尺寸为128×64，并按照[37]分为12,936 / 19,732训练/测试组。在训练组中，我们有439,420对，每对由两个同一个人但不同姿势的图像组成。我们从测试集中随机选择12,800对进行测试。

图4：DeepFashion（上3行，为了显示而剪切图像）和Market-1501数据集（下3行）的测试结果。 我们对G1进行了两个方面的测试：（1）G1-L1中三种姿态嵌入方法，即坐标嵌入（CE），热力图嵌入（HME）和姿态热图拼接;（2）两种损失，即提出的poseMaskLoss 和标准的L1损失。 第7,8,9列显示我们的阶段I（G1），一阶段对抗模型（G1 + D）和我们的两阶段对抗模型（G1 + G2 + D）之间的差异。 请注意，所有三个使用poseMaskLoss。 在分割数据集时，ID是随机分配的。

实现细节在两个数据集上，我们使用Adam [13]优化器，β1= 0.5，β2= 0.999。初始学习率设置为2e-5。在DeepFashion中，我们设置卷积块的数量N = 6。在阶段I和阶段II，分别用30k和20k迭代的大小为8的小批量训练模型。在Market-1501上，我们设置了N = 5的卷积块数。模型在阶段I和阶段II分别用22k和14k迭代的小型批次进行训练。对于数据增强，我们对这两个数据集做左右翻转。

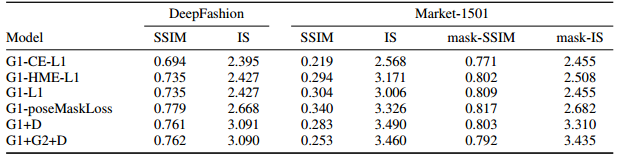
4.2定性结果

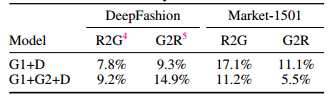
如上所述，我们验证了PG2网络的三个方面。在阶段I中比较不同的姿势嵌入和损失，然后我们证明我们的两阶段模型优于单阶段模型的优势。

不同的姿势嵌入。为了评估我们提出的姿态嵌入方法，我们实现了两种替代方法。首先是坐标嵌入（CE），我们通过两个完全连接的层传递关键点坐标，并将嵌入的特征向量与图像嵌入向量连接在瓶颈完全连接层上。第二种称为热力图嵌入（HME），我们将18个关键点的热图提供给一个独立的编码器，并提取完全连接的图层特征以在瓶颈完全连接图层上与图像嵌入矢量连接。

图4的第4,5和6列表示当在阶段I中使用时不同姿势嵌入方法的定性结果，其具有CE（G1-CE-L1），HME（G1-HME-L1）和我们的G1（ G1-L1）。所有三个使用标准的L1损失。我们可以看到G1-L1能够合成捕捉人体全局结构（如姿势和颜色）的合理外观图像。但是，另外两种嵌入方法G1-CE-L1和G1-HME-L1很模糊，颜色错误。而且，G1-CE-L1的结果都是错误的姿势。这可以解释为将关键点坐标映射到适当的图像位置使训练更具挑战性的额外困难。我们提出的使用18通道姿势热力图的姿势嵌入能够有效地引导生成过程，从而导致正确生成的姿势。有趣的是，G1-L1甚至可以生成合理的脸部细节，如眼睛和嘴巴，如DeepFashion样本所示。

姿态蒙版损失与L1损失。比较Market-1501数据集中L1训练与L1损失（G1-L1）和G1训练结果（使用poseMaskLoss（G1-poseMaskLoss））的结果，我们发现姿态蒙版损失确实带来了性能的改善（图4中的第6列和第7列）。通过将图像生成集中在人体上，合成图像变得更锐利，颜色看起来更漂亮。我们可以看到，对于人物ID164，由G1-L1产生的人的上身颜色比由G1-poseMaskLoss产生的人的上身更嘈杂。对于人员ID 23和346，带有姿势遮罩损失的方法为肩部和头部生成更清晰的边界。这些比较验证了我们的姿态掩膜丢失有效地减轻了嘈杂背景的影响，并引导发生器专注于人体的姿势转移。这两次损失为DeepFashion样本产生了类似的结果，因为背景更为简单。

表1：定量评估。 对于所有的数值，越高越好。

表2：AMT的用户研究结果

两阶段与一阶段。另外，我们证明了我们的两阶段模型优于单阶段模型的优点。为此，我们使用G1作为生成器，但以对抗方式训练它，直接生成新图像，给定条件图像和目标姿势作为输入。这一阶段模型表示为G1 + D，我们的完整模型表示为G1 + G2 + D。从图4中，我们可以看到，我们的完整模型能够生成相片逼真的结果，其中包含比单阶段模型更多的细节。

例如，对于DeepFashion样本，脸部和衣服上的更多细节会转移到生成的图像上。对于ID245，G1 + D的短裤比G1 + G2 + D具有更浅的颜色和更模糊的边界。对于ID 346，两阶段模型能够为衣服生成正确的颜色和纹理，而一阶段模型只能生成正确的颜色。在Market-1501样品上，由于更具挑战性的设置，两种方法生成的图像质量都会下降。但是，两阶段模型仍然能够产生比单阶段方法更好的结果。我们可以看到，对于ID53，T恤上的条纹由我们的完整模型保留，而一阶段模型只能生成蓝色斑点作为衣服。此外，我们也可以清楚地看到女人手中的凳子（ID 23）。

4.3定量结果

我们也给出了两个数据集的定量结果。结构相似性（SSIM）[31]和初始评分（IS）[26]被用来衡量合成的质量。请注意，在Market-1501数据集中，条件图像和目标图像可能具有不同的背景。由于输入中没有关于目标图像中背景的信息，因此我们的方法无法想象新背景是什么样子。为了减少评估中背景的影响，我们提出了一个

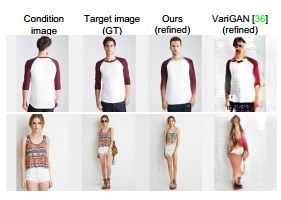
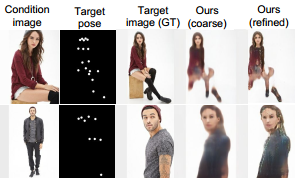
图5：与[36]比较的例子

图6：我们在DeepFashion上的失败案例

称为蒙版SSIM的SSIM变体。在计算SSIM之前，合成和目标图像都会添加姿态遮罩。这样我们只关注测量一个人外观的综合质量。同样，我们使用mask-IS来消除背景的影响。然而，应该注意，图像质量并不总是对应于这种图像相似性度量。例如，在图4中，我们的完整模型比G1-poseMaskLoss生成更清晰，更逼真的结果，但后者具有更高的SSIM。这在超分辨率论文中也被观察到[12,27]。

表1中的数值分数也清楚地显示了这些优点。例如。所提出的姿态嵌入（G1-L1）在所有测量和两个数据集中始终优于G1-CE-L1。 G1-HME-L1可能由于两个嵌入的相似性而获得相似的定量数字。将L1的损失更改为提议的poseMaskLoss（G1-poseMaskLoss）可以在所有测量和两个数据集中进一步提高。在第一阶段（G1 + D）或完整模型（G1 + G2 + D）后，在训练期间添加鉴别器会导致可比数字，尽管我们观察到上述结果的明显差异。这可以通过以下事实来解释：模糊的图像通常会获得较好的SSIM，尽管不太令人信服且不太具有照片真实感[12,27]。

4.4用户研究

我们针对这两个数据集执行Amazon Mechanical Turk（AMT）的用户研究。对于每一个，我们以30个用户的随机顺序显示55个真实图像和55个生成的图像。在[10,15]之后，每个图像显示1秒。前10个图像用于实践，因此在计算分数时被忽略。从表2中的结果可以看出，（1）在DeepFashion中，我们生成的G1 + D和G1 + G2 + D图像分别对9.3％和14.9％显示了G1 + G2 + D优于G1 + D的优势; （2）在Market-1501上，G2R的平均得分较低，因为背景比DeepFashion杂乱得多; （3）在Market-1501上，G1 + G2 + D得分低于G1 + D，因为G1 + G2 + D从条件图像中传递更多的背景，可以在图4中看出，但同时它带来了额外的文物背景，导致用户评价'假'; （4）关于R2G，我们注意到Market-1501获得明显的高分（> 10％），因为人类用户在面对低质量监控图像时有时会感到困惑。

4.5进一步分析

由于我们的姿势条件是新颖的，所以没有直接的比较工作。我们只与最相关的one6 [36]进行比较，它在DeepFashion数据集上进行了多视图人物图像合成。注意到，[36]使用了目标视图的条件图像和附加单词矢量，例如， “边”作为网络输入。比较例如图5所示。很明显，我们的改进结果比[36]更好。以第二排为例，我们可以生成以上身为条件的高质量全身图像，而[36]的全身综合只具有粗糙的身体形状。

另外，我们在图6中用我们的模型给出了两个失败的DeepFashion示例。在第一排中，只有上半身是一致生成的。这些“腿”是由这种复杂姿势的罕见训练数据造成的。最后一行显示男性/女性训练数据不平衡造成的性别不准确。此外，条件人穿着与他的内衣短袖相似颜色的长袖外套，使得生成的布料看起来像是两者的混合物。

5 结论

在这项工作中，我们提出了姿态引导人生成网络（PG2），通过对参考图像和目标姿势进行调节来解决合成人物图像的新任务。采用分治策略将生成过程分为两个阶段。第一阶段旨在捕捉一个人的全球结构并产生一个初步结果。进一步提出姿态掩模损失以减轻背景对人体图像合成的影响。 Stage-II通过对抗训练填充更多外观细节，以生成更清晰的图像。对两个人数据集的广泛实验结果表明，我们的方法能够生成照片逼真和姿势正确的图像。在未来的工作中，我们计划在姿态和属性上产生更多可控和多样化的人物图像调节。

参考文献

[1] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. arXiv, 1701.07875, 2017.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. arXiv, 1611.08050, 2016.

[3] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iter- ative error feedback. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4733–4742, 2016.

[4] Chao-Yeh Chen and Kristen Grauman. Inferring unseen views of people. In CVPR, pages 2003–2010, 2014.

[5] Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In NIPS, pages 2172–2180, 2016.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[8] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. arXiv, 1704.04086, 2017.

[9] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. arXiv preprint arXiv:1612.04357, 2016.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017.

[11] Xu Jia, Amir Ghodrati, Marco Pedersoli, and Tinne Tuytelaars. Towards automatic image editing: Learning to see another you. In BMVC, 2016.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, 2016.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv, 1412.6980, 2014. [14]Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. arXiv, 1312.6114, 2013.

[15] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. arXiv preprint arXiv:1705.04098, 2017.

[16] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In CVPR, pages 1096–1104, 2016.

[17] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In ICLR, 2016.

[18] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv, 1411.1784, 2014.

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In

European Conference on Computer Vision, pages 483–499. Springer, 2016.

[20] Tran Minh Quan, David G. C. Hildebrand, and Won-Ki Jeong. Fusionnet: A deep fully residual convolu- tional neural network for image segmentation in connectomics. arXiv, 1612.05360, 2016.

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv, 1511.06434, 2015.

[22] Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In NIPS, 2016.

[23] Scott Reed, Aäron van den Oord, Nal Kalchbrenner, Victor Bapst, Matt Botvinick, and Nando de Freitas. Generating interpretable images with controllable structure. Technical report, 2016.

[24] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In ICML, 2016.

[25] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approxi- mate inference in deep generative models. In ICML, 2014.

[26] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In NIPS, pages 2226–2234, 2016.

[27] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network.

[28] Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregres- sive distribution estimation. arXiv, 1605.02226, 2016.

[29] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In NIPS, pages 4790–4798, 2016.

[30] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In

ICML, pages 1747–1756, 2016.

[31] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing, 13(4):600–612, 2004.

[32] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In ECCV, pages 776–791, 2016.

[33] Jimei Yang, Scott Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In NIPS, 2015.

[34] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Du-Sik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In CVPR, 2015.

[35] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. arXiv:, 1612.03242, 2016.

[36] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, and Jiashi Feng. Multi-view image generation from a single-view. arXiv, 1704.04886, 2017.

[37] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In ICCV, pages 1116–1124, 2015.

译文原文出处：

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