基于共享注意力的多模态人物图像生成算法

**技术领域**

本发明涉及计算机视觉领域，具体指一种基于共享注意力的多模态人物图像生成算法

**背景技术**

视觉内容生成是一个迅速发展的研究领域，得益于生成模型的不断进步，诸如生成对抗网络(GANs)、自回归模型(AR)和扩散模型，这些模型为各种创新应用开辟了新的可能性。作为一个新兴领域，人物图像生成引起了众多研究者的关注，并为数字人生成、电影制作、时尚设计等下游任务提供了巨大支持。在这个领域中，一个关键性的问题是如何基于不同的输入实现可控的人物图像生成。例如，姿势引导的人物图像生成旨在将人物图像从初始姿势转化为指定的目标姿势，而基于图像的虚拟试穿则致力于将目标服装图像映射到参考人物图像上。

目前，大部分可控的人物图像生成方法仅基于单一的视觉模态（如人物姿势或服装）进行条件生成。然而，不同输入模态传达的条件信息均具备独特的优势。以文本描述为例，它可以以更自然且灵活的方式来指定所需的人物图像属性，比如服装的颜色和图案。同样，人物分割图可以对人物图像的形状和结构提供更精准和直观的控制。显然，一种能够融合来自多种模态的输入信息的可控人物图像生成模型将能够实现更具创意的应用，比如时尚设计和内容创作领域。这样的模型可以更好地满足用户的个性化需求，为创意产业带来新的可能性。

多模态控制的可控人物图像生成任务具有许多优势，但也面临着几个挑战：

1）通用合成框架设计问题。

首要挑战是如何设计一个通用的人物图像生成框架，能够有效地整合来自不同模态的多个控制信号。这需要考虑如何融合不同模态的信息，以及如何确保生成的图像在多模态条件下仍然具有合理性和一致性。

2）多模态控制和高质量生成的平衡问题。

在满足多模态控制的要求的同时，保持生成人物图像的高质量是一个挑战。不同的模态输入可能会导致信息的冲突或不一致性，因此需要找到一种平衡的方法，以确保生成图像既满足多模态控制要求，又具有逼真的质量。

3）多种因素的综合考量问题。

生成逼真的全身人物图像需要综合考虑多种因素，如姿势、体形、发型、纹理图案和时尚风格等。这些因素之间可能存在复杂的相互关系，因此如何在生成过程中平衡这些因素，以确保最终生成的图像在各个方面都具有逼真性和一致性，这是一个挑战。

**发明内容**

为了应对前述挑战，本发明提出一种基于共享注意力的多模态人物图像生成算法。该方法为两阶段模型，包括基于小波变换的离散变分自编码器模型和基于共享注意力的多模态条件扩散模型（S-MCDN），旨在更好地实现多模态输入的可控人物图像生成。其中S-MCDN又由多模态特征融合模块（MFFB）和半解耦去噪解码器（S3D）组成。

本发明的技术方案为：

S1、对现有的多模态人物数据集进行预处理，构建人物图像、人物分割图、文本信息、语义信息数据组；

S2、在现有数据组下，构建包含基于小波变换的离散变分自编码器模型和基于共享注意力的多模态条件扩散模型，并初始化网络权重。通过输入的数据组，得到生成的人物图像。

S2-1、通过所述基于小波变换的离散变分自编码器模型对输入的人物图像进行重构，同时学习人物图像的码本特征。

S2-2、通过所述基于共享注意力的多模态条件扩散模型，学习S2-1中取得的码本索引的先验分布。将文本信息，人物分割图和人物素描图作为多模态输入，其中人物分割图和人物素描图视觉模态通过MFFB模块先行融合，融合后的特征同文本特征相互独立地输入到S3D方法中指导先验分布的预测，从而实现基于多模态输入的可控人物图像生成。

S3、利用反向传播算法学习两阶段多模态人物图像生成模型参数，直至收敛。

作为优选，所述步骤S1中，对现有的多模态人物数据集的预处理方法是，使用人体解析器来计算一个包含24个标签的人体语义信息，每个标签代表人体的特定部分，比如脸、头发、胳膊、腿和衣服区域。关于文本信息，本发明基于DeepFashion-MultiModal数据集中的独热编码形状属性和纹理属性为每个人物图像创建了一个文本模板。例如：一个穿着纯色长袖上衣和牛仔长裤的男人；

作为优选，所述步骤S2-1中，基于小波变换的离散变分自编码器模型由小波增强编码器、解码器和码本组成。码本包含一组码本条目，其中是码本的大小，是编码的维度。

小波增强编码器接受一个全身人物图像作为输入，并输出混合特征。首先，将输入图像输入到空间编码器中提取空间特征。为了恢复丢失的高频信息，对应用离散小波变换，得到四个离散小波子带。然后，将四个离散小波子带串联起来，并将其输入到小波编码器中提取小波域特征。最后，应用交叉注意力块将空间特征和小波域特征融合，得到混合特征，其中本文将小波域特征作为“查询”（），将空间特征作为“键”（）和“值”（）。将输入图像压缩为混合特征后，接下来的空间量化器将每个空间网格特征映射到最近的码本条目：

其中，是量化操作，是量化特征。表示中特定空间位置上的向量。解码器经过训练，可以从量化特征中重构出人物图像，即。

作为优选，所述步骤S2-2中，多模态条件扩散模型由多模态特征融合模块（MFFB）和半解耦去噪解码器（S3D）组成。MFFB模块将用来处理前期多种视觉模态信息的融合，尤其针对人物分割图与人物素描图两种模态创新地提出了共享多头注意力机制( SMHA)，通过多头结构和自适应层归一化（AdaLN），捕捉模态数据之间的复杂关联信息,将两种视觉模态的信息进行融合。将经由MFFB模块处理得到的视觉信息和文本信息利用半解耦多头注意力机制输入到S3D中通过与图像特征的先验噪声和时间步融合，逐步预测图像每一时间步的离散先验。

作为优选，所述步骤S3中，两阶段多模态图像生成模型的目标函数包括重构损失、码本损失、约束损失、最小化对抗损失及交叉熵损失，其中分别代表各个损失函数的权重：

利用收集到的数据组训练基于小波变换的离散变分自编码器模型和多模态条件扩散模型。

本发明具有以下的特点和有益效果：

采用上述技术方案，将提出的基于共享注意力的多模态人物图像生成算法用于可控的人物图像生成，将人物分割图和文本描述作为多模态控制，能够生成高保真且多样性的全身人物图像，并且能够有效地满足输入的多模态控制条件。所提出的方法可以在细粒度外观（如面部特征、手指、复杂纹理和时尚配饰）中生成具有高度多样性和保真度的人物图像。

该发明意味着可以通过不同的输入条件精准控制生成的人物图像，满足多种应用场景的需求，例如数字人生成、时尚设计以及内容创作等。这对于推动计算机视觉领域中的人物图像生成技术具有重要意义，为时尚设计领域带来了许多创新的应用可能性。同时，也为在人物图像领域应用多模态生成模型提供了新的思路和方法。

**附图说明**

图1为本发明实施例的具体流程示意图；

图2为本发明实施例中模型；

图3为本发明数据组示意图。

**具体实施方式**

下面结合附图和实施例对本发明作进一步说明。

本发明提出一种多模态引导的人物图像生成方法。该方法为两阶段模型，包括基于小波变换的离散变分自编码器模型和多模态条件扩散模型，旨在更好地实现多模态输入的可控人物图像生成。

一种多模态引导的人物图像生成方法，如图1所示，其步骤如下：

步骤（1）、借助现有的多模态人物数据集并进行预处理，构建人物图像、文本信息、语义信息数据组。例如：附图3中一个模特的数据组就由具体的人物照片图像，经过人体解析器分割后的抽象化语义信息图像和对人物特征详细的文本信息描述构成；

本示例中，借助现有的多模态人物数据集，是指在DeepFashion-MultiModal数据集上评估了的方法，它由11484张分辨率为1024512的高质量人物图像组成，其中每个图像都与人物分割图标签和服装属性（即形状和纹理）相关联；

步骤（2）、在现有数据组下，构建包含基于小波变换的离散变分自编码器模型和基于共享注意力的多模态条件扩散模型，并初始化网络权重。

步骤（3）、在重构损失、码本损失、约束损失、最小化对抗损失及交叉熵损失下，利用收集到的数据组训练基于小波变换的离散变分自编码器模型和多模态条件扩散模型。

步骤（4）、通过反向传播算法对步骤（3）中的模型中的网络参数进行训练，直至整个模型收敛。多模态条件扩散模型利用离散扩散过程最大化条件概率，包括在每一步添加噪声的前向扩散过程和从噪声状态恢复样本的反向去噪过程，从而生成具有多样性的逼真人物图像。

具体的，步骤（1）所述的构建人物图像、文本信息、语义信息数据组，是指对于对应的人物图像数据，使用先进的人体解析器来计算一个包含24个标签的人体语义信息，每个标签代表人体的特定部分，比如脸、头发、胳膊、腿和衣服区域。关于文本信息，本发明基于DeepFashion-MultiModal数据集中的独热编码形状属性和纹理属性为每个人物图像创建了一个文本模板。例如：一个穿着纯色长袖上衣和牛仔长裤的男人；

步骤（2）所述的构建两阶段模型，包括基于小波变换的离散变分自编码器模型和基于共享注意力的多模态条件扩散模型，实现多模态引导的人物图像生成，具体如下：

如图2顶部所示，第一阶段：基于小波变换的离散变分自编码器模型

通过基于小波变换的离散变分自编码器模型，学习人物图像的码本。通过将小波特征融合到编码器中，对传统的向量量化变分自编码器进行了增强，从而提升了图像重构和合成的质量。

具体而言，基于小波变换的离散变分自编码器模型由小波增强编码器、解码器和码本组成。码本包含一组码本条目，其中是码本的大小，是编码的维度。小波增强编码器接受一个人物图像作为输入，并输出混合特征。首先，将人物图像输入到空间编码器中提取空间特征。为了恢复丢失的高频信息，对应用离散小波变换，得到四个离散小波子带。然后，将这些离散小波子带串联起来，并将其输入到小波编码器中提取小波域特征。最后，应用交叉注意力块将空间特征和小波域特征融合，得到混合特征，其中本文将小波域特征作为“查询”（），将空间特征作为“键”（）和“值”（）。将输入图像压缩为潜在向量后，接下来的空间量化器将每个空间网格特征映射到最近的码本条目：

其中，是量化操作，是量化特征。表示中特定空间位置上的向量。解码器经过训练，可以从量化编码中重构出人物图像，即。

该阶段模型可以通过最小化以下函数进行端到端训练：

其中，第一项是重构损失，用于估计输入人物图像和重构图像之间的误差。第二项是码本损失，用于优化码本中的条目。第三项是约束损失，鼓励小波增强编码器的输出保持接近所选的码本向量。表示停止梯度操作，是第三个约束损失的超参数。

如图2所示，第二阶段：

通过所述多模态条件扩散模型，离散扩散过程中在多模态控制下估计图像离散空间上的潜在先验分布，从而实现了基于多模态输入的可控人物图像生成。该模型接收文本描述，人物分割图和人物素描图作为输入。

具体来说，多模态条件扩散模型由多模态特征融合模块（MFFB）、共享多头注意力机制（SMHA）和半解耦去噪解码器（S3D）组成。首先，利用多模态特征融合模块中的共享多头注意力机制，对人物分割图和人物素描图进行特征提取，得到查询（）、键（）和值（）。通过自适应层归一化操作来调整查询和键的特征，以实现不同模态之间的有效融合。再利用多头注意力机制并行处理不同部分的信息，每个头关注不同的特征组合。最后对所有头的输出进行拼接和求和，得到人物分割图和人物素描图融合后的视觉信息。

在处理完人物分割图和人物素描图后，将文本描述信息和MFFB模块处理后的视觉信息输入到S3D模块。在S3D模块中，使用半解耦多头注意力机制（SDMHA），分别处理视觉信息流和文本信息流。通过Transformer网络，将处理后的信息与噪声图像先验和时间步融合，逐步预测图像的离散先验分布。

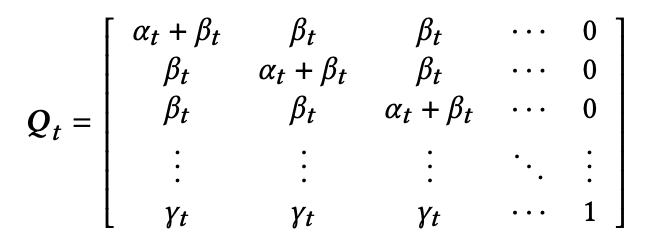
步骤（4）所述多模态条件扩散模型利用离散扩散过程最大化条件概率是指，给定一个四元组，使用预训练的基于小波变换的离散变分自编码器模型获得离散图像令牌，素描编码器、分割编码器和文本编码器分别获得离散形式的素描标记、分割标记和文本标记。接着，该多模态条件扩散模型利用离散扩散过程最大化条件概率，其中包括在每一步添加噪声的前向扩散过程和从噪声状态恢复样本的反向去噪过程，具体如下：

1）前向扩散过程

正向扩散过程通过固定的马尔可夫链逐步破坏初始图像。在预定数量的时间步长之后，正向过程生成一个与相同维度的日益嘈杂的潜变量序列，其中表示纯噪声令牌。其中，正向扩散过程可以表示为：

其中是一个独热列向量，索引为，条目为1。是从到的概率转移矩阵。

多模态条件扩散模型采用掩码替换扩散策略，其中概率转移矩阵定义为：



在这里，，和是令牌被替换为[MASK]令牌的概率，为与矩阵维数相关的常数。每个原始图像令牌有被[MASK]令牌替换的概率，有被均匀扩散的概率，使概率保持不变。同时，[MASK]令牌始终保持自己的状态。

2）逆向去噪过程

一旦通过前向扩散过程获得纯噪声令牌，多模态条件扩散模型将逐步对潜变量进行去噪，并依次恢复真实数据。相反的过程由后验分布给出：

这里，是累积转移矩阵。概率可以用闭合形式计算：

其中，，和可以预先计算和存储。

为了估计上述后验分布，使用半解耦去噪解码器 S3D，并引入了一

个额外的人物素描图模态。该过程可表示为：

最后，使用重参数化技巧训练该模型，以最小化变分下界，该下界可以表示为：

在这里，表示模型分布与真实分布之间的KL散度，是一个辅助去噪损失，鼓励多模态条件扩散模型预测无噪声令牌 ,是权重参数。表示时间步的先验分布。对于所提出的掩模替换扩散方法，先验分布定义为：

以上结合附图对本发明的实施方式作了详细说明，但本发明不限于所描述的实施方式。对于本领域的技术人员而言，在不脱离本发明原理和精神的情况下，对这些实施方式包括部件进行多种变化、修改、替换和变型，仍落入本发明的保护范围内。