(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112990283 A
(43)申请公布日 2021.06.18

(21)申请号 202110237774.6

(22)申请日 2021.03.03

(71)申请人 网易(杭州)网络有限公司
地址 310052 浙江省杭州市滨江区长河街道网商路599号4幢7层

(72)发明人 袁燚 许曼玲 范长杰 胡志鹏

(74)专利代理机构 北京超凡宏宇专利代理事务所(特殊普通合伙) 11463
代理人 钟扬飞

(51)Int.Cl.
G06K 9/62(2006.01)
G06K 9/00(2006.01)

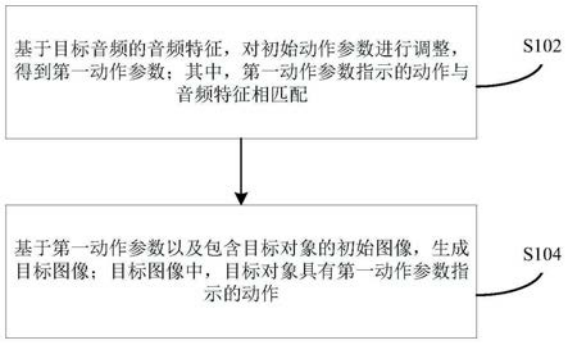
权利要求书2页 说明书11页 附图6页

(54)发明名称

图像生成方法、装置和电子设备

(57)摘要

本发明提供了一种图像生成方法、装置和电子设备;其中,该方法包括:基于目标音频的音频特征,对初始动作参数进行调整,得到第一动作参数;其中,第一动作参数指示的动作与音频特征相匹配;基于第一动作参数以及包含目标对象的初始图像,生成目标图像;目标图像中,目标对象具有第一动作参数指示的动作。该方式中,通过音频的音频特征调整动作参数,可以使得到的第一动作参数所指示的动作与该音频特征相匹配,进而使生成的图像中的目标对象具有第一动作参数指示的动作,因而该方式可以通过音频控制图像中对象的动作,使图像中对象的动作随着音频内容的变化而变化,在播放音频的过程中,图像内容变化多样,提高了用户的视觉体验。



CN 112990283 A

1.一种图像生成方法,其特征在于,所述方法包括:

基于目标音频的音频特征,对初始动作参数进行调整,得到第一动作参数;其中,所述第一动作参数指示的动作与所述音频特征相匹配;

基于所述第一动作参数以及包含目标对象的初始图像,生成目标图像;所述目标图像中,所述目标对象具有所述第一动作参数指示的动作。

2.根据权利要求1所述的方法,其特征在于,所述目标对象包括人脸;所述第一动作参数指示的动作包括所述人脸的表情动作。

3.根据权利要求1所述的方法,其特征在于,所述目标音频的音频特征用于调整所述初始动作参数指示的动作的动作幅度;所述第一动作参数指示的动作的动作幅度与所述音频特征相匹配。

4.根据权利要求1所述的方法,其特征在于,基于目标音频的音频特征,对初始动作参数进行调整,得到第一动作参数的步骤,包括:

根据所述目标音频的音频特征,确定参数调整权重;

基于所述参数调整权重,对所述初始动作参数进行放缩处理,得到第一动作参数。

5.根据权利要求4所述的方法,其特征在于,根据所述目标音频的音频特征,确定参数调整权重的步骤,包括:

在所述音频特征的时间维度上,对所述时间维度上的各个时间点对应的特征向量求取平均值,得到初始参数;

将所述初始参数映射至预设的数值范围中,得到所述参数调整权重。

6.根据权利要求4所述的方法,其特征在于,根据所述目标音频的音频特征,确定参数调整权重的步骤之前,所述方法还包括:

对所述音频特征中,任意两个相邻的初始时间点之间插入指定数量的中间时间点,以及每个所述中间时间点对应的特征向量,得到最终的所述音频特征;其中,所述中间时间点对应的特征向量,基于与所述中间时间点相邻的两个初始时间点对应的特征向量确定。

7.根据权利要求1所述的方法,其特征在于,所述目标音频的音频特征,通过下述方式得到:

提取所述目标音频的梅尔频率倒谱系数MFCC参数;所述MFCC参数包括预设时间间隔的多个时间点,以及每个时间点对应一个MFCC数值;

将所述MFCC参数输入至预先训练完成的特征提取网络中,输出所述目标音频的音频特征。

8.根据权利要求7所述的方法,其特征在于,所述特征提取网络包括多个依次串联的特征提取模块;所述特征提取模块包括卷积层、批量归一化层和激活函数层。

9.根据权利要求7所述的方法,其特征在于,将所述MFCC参数输入至预先训练完成的特征提取网络中,输出所述目标音频的音频特征的步骤之前,所述方法还包括:

基于预设的填充值,对所述MFCC参数的频率维度上的数值进行数值填充,以使所述频率维度上的数值数量与所述MFCC参数的时间维度上的数值数量相匹配;

将数值填充后的所述MFCC参数进行复制,得到指定通道数量的所述MFCC参数。

10.根据权利要求7所述的方法,其特征在于,所述特征提取网络通过下述方式训练得到:

将样本音频的MFCC参数输入至编码网络中，输出所述样本音频的特征向量；将所述样品音频的特征向量输入至解码网络中，得到所述样本音频的输出参数；

基于预设的损失函数，计算所述输出参数与所述样本音频的MFCC参数之间的损失值，基于所述损失值训练所述编码网络和所述解码网络，将训练完成后的所述编码网络确定为所述特征提取网络。

11.根据权利要求10所述的方法，其特征在于，所述解码网络包括多个依次串联的解码模块；所述解码模块包括转置卷积层、批量归一化层和激活函数层。

12.根据权利要求1所述的方法，其特征在于，基于所述第一动作参数以及包含目标对象的初始图像，生成目标图像的步骤，包括：

提取所述包含目标对象的初始图像的图像特征；其中，所述图像特征包括全局特征和细节特征；

将所述第一动作参数与所述细节特征进行融合处理，得到融合特征；

基于所述融合特征和所述全局特征，生成所述目标图像。

13.根据权利要求12所述的方法，其特征在于，将所述第一动作参数与所述细节特征进行融合处理，得到融合特征的步骤，包括：

从所述图像特征中获取第一指定通道中的特征数据；其中，所述第一指定通道中的特征数据包含所述细节特征；

从所述第一动作参数中获取第二指定通道中的参数数据；

将所述第一指定通道中的特征数据和所述第二指定通道中的参数数据进行逐点相加处理，得到所述融合特征。

14.根据权利要求12所述的方法，其特征在于，基于所述融合特征和所述全局特征，生成所述目标图像的步骤，包括：

将所述融合特征和所述全局特征作为隐向量，输入至预先训练完成的图像生成网络中，输出所述目标图像；其中，所述隐向量用于：控制所述图形生成网络输出与所述隐向量指示的特征相匹配的图像。

15.一种图像生成装置，其特征在于，所述装置包括：

参数调整模块，用于基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；其中，所述第一动作参数指示的动作与所述音频特征相匹配；

图像生成模块，用于基于所述第一动作参数以及包含目标对象的初始图像，生成目标图像；所述目标图像中，所述目标对象具有所述第一动作参数指示的动作。

16.一种电子设备，其特征在于，包括处理器和存储器，所述存储器存储有能够被所述处理器执行的机器可执行指令，所述处理器执行所述机器可执行指令以实现权利要求1-14任一项所述的图像生成方法。

17.一种机器可读存储介质，其特征在于，所述机器可读存储介质存储有机器可执行指令，所述机器可执行指令在被处理器调用和执行时，所述机器可执行指令促使所述处理器实现权利要求1-14任一项所述的图像生成方法。

图像生成方法、装置和电子设备

技术领域

[0001] 本发明涉及图像处理技术领域，尤其是涉及一种图像生成方法、装置和电子设备。

背景技术

[0002] 终端设备播放音频时，在显示屏幕上显示特定的图像，并使图像内容随着音频律动的变化而变化，可以提高用户在倾听音频时的视觉体验感。相关技术中，随着音频律动变化的图像通常为跳跃的条状频谱图；将正在播放的音频进行傅里叶变换，得到音频的频域特征，基于该频域特征即可生成上述条状频谱图。但这种图像内容较为单一，对用户而言缺乏吸引力，用户视觉体验感较低。

发明内容

[0003] 有鉴于此，本发明的目的在于提供一种图像生成方法、装置和电子设备，以在播放音频的过程中，图像内容变化多样，提高用户的视觉体验。

[0004] 第一方面，本发明实施例提供了一种图像生成方法，方法包括：基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；其中，第一动作参数指示的动作与音频特征相匹配；基于第一动作参数以及包含目标对象的初始图像，生成目标图像；目标图像中，目标对象具有第一动作参数指示的动作。

[0005] 上述目标对象包括人脸；第一动作参数指示的动作包括人脸的表情动作。

[0006] 上述目标音频的音频特征用于调整初始动作参数指示的动作的动作幅度；第一动作参数指示的动作的动作幅度与音频特征相匹配。

[0007] 上述基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数的步骤，包括：根据目标音频的音频特征，确定参数调整权重；基于参数调整权重，对初始动作参数进行放缩处理，得到第一动作参数。

[0008] 上述根据目标音频的音频特征，确定参数调整权重的步骤，包括：在音频特征的时间维度上，对时间维度上的各个时间点对应的特征向量求取平均值，得到初始参数；将初始参数映射至预设的数值范围中，得到参数调整权重。

[0009] 上述根据目标音频的音频特征，确定参数调整权重的步骤之前，方法还包括：对音频特征中，任意两个相邻的初始时间点之间插入指定数量的中间时间点，以及每个中间时间点对应的特征向量，得到最终的音频特征；其中，中间时间点对应的特征向量，基于与中间时间点相邻的两个初始时间点对应的特征向量确定。

[0010] 上述目标音频的音频特征，通过下述方式得到：提取目标音频的梅尔频率倒谱系数MFCC参数；MFCC参数包括预设时间间隔的多个时间点，以及每个时间点对应一个MFCC数值；将MFCC参数输入至预先训练完成的特征提取网络中，输出目标音频的音频特征。

[0011] 上述特征提取网络包括多个依次串联的特征提取模块；特征提取模块包括卷积层、批量归一化层和激活函数层。

[0012] 上述将MFCC参数输入至预先训练完成的特征提取网络中，输出目标音频的音频特

征的步骤之前，方法还包括：基于预设的填充值，对MFCC参数的频率维度上的数值进行数值填充，以使频率维度上的数值数量与MFCC参数的时间维度上的数值数量相匹配；将数值填充后的MFCC参数进行复制，得到指定通道数量的MFCC参数。

[0013] 上述特征提取网络通过下述方式训练得到：将样本音频的MFCC参数输入至编码网络中，输出样本音频的特征向量；将样品音频的特征向量输入至解码网络中，得到样本音频的输出参数；基于预设的损失函数，计算输出参数与样本音频的MFCC参数之间的损失值，基于损失值训练编码网络和解码网络，将训练完成后的编码网络确定为特征提取网络。

[0014] 上述解码网络包括多个依次串联的解码模块；解码模块包括转置卷积层、批量归一化层和激活函数层。

[0015] 上述基于第一动作参数以及包含目标对象的初始图像，生成目标图像的步骤，包括：提取包含目标对象的初始图像的图像特征；其中，图像特征包括全局特征和细节特征；将第一动作参数与细节特征进行融合处理，得到融合特征；基于融合特征和全局特征，生成目标图像。

[0016] 上述将第一动作参数与细节特征进行融合处理，得到融合特征的步骤，包括：从图像特征中获取第一指定通道中的特征数据；其中，第一指定通道中的特征数据包含细节特征；从第一动作参数中获取第二指定通道中的参数数据；将第一指定通道中的特征数据和第二指定通道中的参数数据进行逐点相加处理，得到融合特征。

[0017] 上述基于融合特征和全局特征，生成目标图像的步骤，包括：将融合特征和全局特征作为隐向量，输入至预先训练完成的图像生成网络中，输出目标图像；其中，隐向量用于：控制图形生成网络输出与隐向量指示的特征相匹配的图像。

[0018] 第二方面，本发明实施例提供了一种图像生成装置，装置包括：参数调整模块，用于基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；其中，第一动作参数指示的动作与音频特征相匹配；图像生成模块，用于基于第一动作参数以及包含目标对象的初始图像，生成目标图像；目标图像中，目标对象具有第一动作参数指示的动作。

[0019] 第三方面，本发明实施例提供了一种电子设备，包括处理器和存储器，存储器存储有能够被处理器执行的机器可执行指令，处理器执行机器可执行指令以实现上述图像生成方法。

[0020] 第四方面，本发明实施例提供了一种机器可读存储介质，机器可读存储介质存储有机器可执行指令，机器可执行指令在被处理器调用和执行时，机器可执行指令促使处理器实现上述图像生成方法。

[0021] 本发明实施例带来了以下有益效果：

[0022] 上述图像生成方法、装置和电子设备，首先基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；该第一动作参数指示的动作与音频特征相匹配；然后基于第一动作参数以及包含目标对象的初始图像，生成目标图像；该目标图像中的目标对象具有第一动作参数指示的动作。该方式中，通过音频的音频特征调整动作参数，可以使得到的第一动作参数所指示的动作与该音频特征相匹配，进而使生成的图像中的目标对象具有第一动作参数指示的动作，因而该方式可以通过音频控制图像中对象的动作，使图像中对象的动作随着音频内容的变化而变化，在播放音频的过程中，图像内容变化多样，提高了用户的视觉体验。

[0023] 本发明的其他特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点在说明书、权利要求书以及附图中所特别指出的结构来实现和获得。

[0024] 为使本发明的上述目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合所附附图，作详细说明如下。

附图说明

[0025] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案，下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施方式，对于本领域技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0026] 图1为本发明实施例提供的一种图像生成方法的流程图；

[0027] 图2为本发明实施例提供的特征提取网络的训练流程示意图；

[0028] 图3为本发明实施例提供的基于音频特征调整图像特征的流程示意图；

[0029] 图4为本发明实施例提供的一种StyleGAN网络的示意图；

[0030] 图5为本发明实施例提供的在播放目标音频的过程中生成多张目标图像的示意图；

[0031] 图6为本发明实施例提供的人脸图像随着音频律动变化的示意图；

[0032] 图7为本发明实施例提供的一种图像生成装置的结构示意图；

[0033] 图8为本发明实施例提供的一种电子设备的结构示意图。

具体实施方式

[0034] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合附图对本发明的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0035] 考虑到相关技术中，终端设备在播放音乐时，屏幕上显示的图像内容较为单一，导致用户视觉体验感较低的问题，本发明实施例提供的图像生成方法、装置和电子设备，可以应用于网页页面、APP(Application，应用程序)等播放音乐时的图像显示场景中。

[0036] 首先，参见图1所示的一种图像生成方法的流程图，该方法包括如下步骤：

[0037] 步骤S102，基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；其中，第一动作参数指示的动作与音频特征相匹配；

[0038] 上述目标音频可以是音乐、语音或其他类型声音等音频；目标音频的音频特征，可以通过训练好的音频特征提取网络提取得到；目标音频的音频特征中，通常可以包含音频中声音的声调、节奏、语速、感情等特征。初始动作参数通常对应预设的初始动作，以人脸的表情动作为例，该初始动作可以为人脸不具有表情动作，或者具有默认的表情动作。该初始动作参数可以预先设置，也可以从图像中提取，例如，可以从下述目标图像中提取。动作参数发生变化时，动作参数对应的动作也发生变化，上述目标音频在播放过程中，随着时序的变化，播放内容在变化，进而导致不同时间点或时间点的音频特征也在变化；基于音频特征

对初始动作参数进行调整，得到的第一动作参数也在不断变化。具体的调整方式可以为对参数进行缩放、更换、取反、按照一定的规则映射等等。

[0039] 不同的动作参数指示的动作类型可以不同，或者，不同的动作参数指示的动作类型相同，但动作幅度不同。当然，当音频特征相同时，得到的第一动作参数通常也相同。因而通过上述步骤，音频特征与第一动作参数具有一定的对应关系，因而第一动作参数指示的动作与音频特征也具有一定的对应关系，二者相匹配。

[0040] 步骤S104，基于第一动作参数以及包含目标对象的初始图像，生成目标图像；目标图像中，目标对象具有第一动作参数指示的动作。

[0041] 目标对象可以为人脸、人体其他部位、人体、或者其他动物或静物，本实施例对目标对象的类型不作限定。初始图像中的目标对象具有默认动作，以人脸为例，该人脸的默认动作可以为“无表情”，或者“微笑”的表情。该初始图像包含目标对象的图像信息，还可以包含背景区域的图像信息。上述第一动作参数可以对目标对象的动作进行改变，使得目标对象具有第一动作参数所指示的动作，从而得到目标图像。

[0042] 在实际实现时，可以提取初始图像的图像特征，将第一动作参数作为目标对象的动作特征融入初始图像的图像特征中，使图像特征中具有第一动作参数相关的特征，在基于图像特征得到上述目标图像。

[0043] 上述图像生成方法，首先基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；该第一动作参数指示的动作与音频特征相匹配；然后基于第一动作参数以及包含目标对象的初始图像，生成目标图像；该目标图像中的目标对象具有第一动作参数指示的动作。该方式中，通过音频的音频特征调整动作参数，可以使得到的第一动作参数所指示的动作与该音频特征相匹配，进而使生成的图像中的目标对象具有第一动作参数指示的动作，因而该方式可以通过音频控制图像中对象的动作，使图像中对象的动作随着音频内容的变化而变化，在播放音频的过程中，图像内容变化多样，提高了用户的视觉体验。

[0044] 为了进一步丰富图像变化内容，本实施例中的上述目标对象可以包括人脸；包含目标对象的初始图像可以预先设置，也可以由用户提供；第一动作参数指示的动作包括人脸的表情动作；通过音频控制图像中人脸的表情动作，使人脸表情随着音频内容的变化而变化，可以提高图像显示的趣味性，进一步提高用户的视觉体验。

[0045] 一种具体的实现方式中，上述目标音频的音频特征用于调整初始动作参数指示的动作的动作幅度；具体的，当音频较为和缓轻柔时，调整的动作幅度可以较小，当音频紧张激烈时，调整的动作幅度较大。以人脸表情动作为例，音频由和缓轻柔逐渐变为紧张激烈时，表情动作可以逐渐变为“微笑”“含笑”“大笑”“狂笑”，即表情动作笑的幅度越来越大。第一动作参数指示的动作的动作幅度与音频特征相匹配，该方式可以使对象随着音频变化的动作更加连贯逼真，使用户在视觉和听觉上同时体验音频律动的变化，提高用户的体验。

[0046] 下面描述目标音频的音频特征的获取方法。首先，提取目标音频的梅尔频率倒谱系数MFCC(Mel-scale Frequency Cepstral Coefficients，梅尔频率倒谱系数)参数；该MFCC参数包括预设时间间隔的多个时间点，以及每个时间点对应一个MFCC数值；由于音频是具有时序性的，在提取MFCC参数时，可以每隔一定的时间间隔对目标音频裁剪一次，然后对裁剪到的这段音频计算MFCC，例如，每间隔100毫秒裁剪一次，对100毫秒内的音频计算MFCC。因而，对于具有一定时间长度的目标音频而言，通常会被裁剪为多段音频，因而目标

音频的MFCC参数包括多个MFCC数值，每个MFCC数值对应一个时间点，相邻的时间点的时间间隔与音频裁剪的时间间隔相同。另外，考虑到人耳所能捕获的声音特征主要集中在低频阶段，为了减低不必要的数据处理量，可以对音频进行滤波处理，滤除高频阶段的音频，仅对低频阶段的音频计算MFCC。在实际实现时，可以预先设置一个频率阈值，例如该阈值设置为13，将高于该频率阈值的频率滤除。然后将MFCC参数输入至预先训练完成的特征提取网络中，输出目标音频的音频特征。

[0047] 上述特征提取网络包括多个依次串联的特征提取模块；该特征提取模块包括卷积层、批量归一化层和激活函数层。在实际实现时，可以采用八个特征提取模块依次串联；每个特征提取模块中的结构相同，但各个模块在训练完成后的参数可能不同。上述卷积层具体可以为2D卷积层，上述批量归一化层也可以称为batchnormalization层，上述激活函数层具体可以为Relu函数。

[0048] 为了使输入数据的格式满足特征提取网络的要求，MFCC参数在输入至特征提取网络之前，需要进行预处理。具体的，首先基于预设的填充值，对MFCC参数的频率维度上的数值进行数值填充，以使频率维度上的数值数量与MFCC参数的时间维度上的数值数量相匹配；目标音频的MFCC参数属于二维数据，包括时间维度和频率维度，在时间维度上包括多个预设间隔排列的时间点，频率维度上包括每个时间点对应的MFCC数值。当目标音频较长时，在时间维度上的时间点数量可能会远大于频率维度上MFCC数值的数据量，为了得到宽度与高度相同的二维矩阵，需要对MFCC参数的频率维度上的数值进行数值填充，例如，在MFCC数值的后面填充上述填充值，该填充值可以为零或其他数值，从而使频率维度上的数值数量与时间维度上的数值数量相同。其他实施方式中，特征提取网络可能对输入数据的通道数也有要求，上述得到的MFCC数值为单通道数据，如果特征提取网络需要输入指定通道数量的数据，则需要将上述数值填充后的MFCC参数进行复制，得到指定通道数量的MFCC参数。通过数据堆叠的方式增加输入数据的维度或通道数。

[0049] 另外，MFCC参数中的MFCC数值可能分布在一个较大的数值范围内，此时还需要对MFCC参数中的数值进行归一化处理，将MFCC参数中的数值映射至一个指定的数据范围内，如[0,1]的范围之内，以便于后续的数据处理。预处理后的MFCC参数输入至上述特征提取网络后，即可输出目标音频的音频特征。

[0050] 上述特征提取网络需要预先训练，本实施例中采用无监督训练的方式训练该特征提取网络，从而减少对样本数据进行标签标记的难度和时间人力成本。具体的，特征提取网络通过下述方式训练得到：将样本音频的MFCC参数输入至编码网络中，输出样本音频的特征向量；将样品音频的特征向量输入至解码网络中，得到样本音频的输出参数；基于预设的损失函数，计算输出参数与样本音频的MFCC参数之间的损失值，基于损失值训练编码网络和解码网络，将训练完成后的编码网络确定为特征提取网络。

[0051] 为了便于理解，图2示出了特征提取网络的训练流程示意图，输入样本音频后，提取样本音频的MFCC参数，对样本音频的MFCC参数进行预处理，例如，数值填充、堆叠增维等；将预处理后的MFCC参数输入至编码网络，输出样本音频的编码向量，即上述样本音频的特征向量；该编码向量输入至解码网络中，得到样本音频的输出参数；基于预设的损失函数，计算输出参数与样本音频的MFCC参数之间的损失值，基于该损失值优化编码网络和解码网络的网络参数，判断是否达到预设的训练次数，如果没有训练次数，则继续执行前述提取样

本音频的MFCC参数的步骤，如果达到了训练次数，则训练结束。

[0052] 上述编码网络的网络结构可以参考前述特征提取网络的网络结构，解码网络的网络结构与编码网络相对应，具体的，解码网络包括多个依次串联的解码模块；解码模块包括转置卷积层、批量归一化层和激活函数层。在实际实现时，可以采用八个解码模块依次串联；每个解码模块中的结构相同，但各个模块在训练完成后的参数可能不同。上述卷积层具体可以为2D卷积层，上述批量归一化层也可以称为batchnormalization层，上述激活函数层具体可以为Relu函数。

[0053] 通过上述解码网络和编码网络进行无监督学习，从而获取一个适用于音频特征提取的编码网络，作为特征提取网络。特征提取网络训练完成后，将目标音频的MFCC参数作为网络的输入数据，可在指定的时间间隔下提取音频特征，输出音频特征，该音频特征的维度可以为一个具有时序性的向量Wc，该向量的维度可以为(1,512)。

[0054] 得到目标音频的音频特征后，则基于该音频特征对初始动作参数进行调整。具体的实现方式中，可以根据目标音频的音频特征，确定参数调整权重；基于该参数调整权重，对初始动作参数进行放缩处理，得到第一动作参数。该参数调整权重可以对初始动作参数中的参数值进行放大或缩小，从而控制动作的幅度；具体的，将该参数调整权重与初始动作参数相乘，即可得到第一动作参数。当参数调整权重不为一时，上述第一动作参数所指示的动作的幅度通常与初始动作参数不同。该方式可以基于音频特征确定调整动作参数的权重，进而调整动作参数，从而实现基于音频特征调整动作幅度。

[0055] 在基于音频特征确定参数调整权重时，一种具体的实现方式中，在音频特征的时间维度上，对时间维度上的各个时间点对应的特征向量求取平均值，得到初始参数；音频特征中的每个时间点对应一个特征向量，该特征向量具有一定的长度，如(1,512)，因而每个时间点对应的特征向量中包含多个特征值，对特征向量中包含的多个特征值求取平均值，即可得到上述初始参数。考虑到特征向量中特征值的分布范围较大，因而得到的平均值可能过大或者过小，如果直接采用该初始参数作为权重调整初始动作参数，可能导致最终目标对象的动作过于夸张，给用户造成惊悚或恐惧的不良体验。为了避免该问题，本实施例将初始参数映射至预设的数值范围中，得到参数调整权重。该数值空间可以预先设置，例如[0,1]，也可以为其他的数值范围。通过该方式，可以将参数调整权重保持在一个合理的范围内，使最终生成的目标对象的动作合理逼真，给用户提供良好的视觉体验。

[0056] 由上述实施例可知，目标音频的音频特征中包括多个预设时间间隔的时间点，以及每个时间点对应的特征向量。如果时间间隔较长，则相邻的两个时间点的特征向量差异较大，导致目标对象的动作变化较为跳跃、不平滑。为了避免该问题，在确定参数调整权重之前，对音频特征中，任意两个相邻的初始时间点之间插入指定数量的中间时间点，以及每个中间时间点对应的特征向量，得到最终的音频特征；其中，中间时间点对应的特征向量，基于与中间时间点相邻的两个初始时间点对应的特征向量确定。上述的初始时间点为提取特征向量的时间点，中间时间点为插入的时间点。该方式既考虑了音频特征的生成速度，又可以使生成的图像满足人眼对动作变化的平滑度要求。

[0057] 作为示例，如果每隔0.25秒提取一个特征向量，则相邻的初始时间点之间的时间间隔为0.25秒，此时，可以在每两个相邻的初始时间点中插入7个中间时间点，这7个中间时间点可以均匀分布在相邻的两个初始时间点之间。中间时间点的特征向量，可以对前后相

邻的两个初始时间点对应的特征向量进行加权平均得到。加权平均的权值，可以根据每个中间时间点的位置确定，例如，按照时序排列的两个初始时间点A和初始时间点B，对于靠近初始时间点A的中间时间点，计算其特征向量时，需要给与初始时间点A较大的权值，以使该中间时间点对应的特征向量与初始时间点A的特征向量相似度较高；同理，对于靠近初始时间点B的中间时间点，计算其特征向量时，需要给与初始时间点B较大的权值，以使该中间时间点对应的特征向量与初始时间点B的特征向量相似度较高；对于处于中间位置的中间时间点，计算其特征向量时，需要给与初始时间点A和初始时间点B同样的权值，通过该方式，可以实现初始时间点A至初始时间点B这段时间内，目标对象的动作的平滑变化。

[0058] 得到第一动作参数后，基于第一动作参数以及包含目标对象的初始图像，生成目标图像。具体的，首先提取包含目标对象的初始图像的图像特征；其中，该图像特征包括全局特征和细节特征；将第一动作参数与细节特征进行融合处理，得到融合特征；基于融合特征和全局特征，生成目标图像。

[0059] 可以预先训练一个图像特征提取网络，将上述初始图像输入至该网络中，即可输出初始图像的图像特征。该图像特征提取网络可以通过反向编码网络StyleGAN2 Encoder实现，初始图像输入至该反向编码网络后，经过多轮迭代后，例如1000次迭代，即可输出尺度为(18,512)的编码向量Wf，该编码向量即上述初始图像的图像特征。在上述反向编码网络中，初始化隐向量w为0值，并用该值生成一张随机的图像，利用VGG16网络对随机生成的图像和目标样本图像分别进行特征提取，比较两者的差值作为损失函数，迭代优化隐向量w，使隐向量w作为输入所产生的图像无限逼近目标图像，从而获得该目标图像对应的反向编码，即上述编码向量。

[0060] 在上述反向编码网络的基础上，将同一对象的标准图像和具有某一指定动作的图像分布输入至该反向编码网络中，得到标准图像对应的第一隐向量，以及具有某一指定动作的图像对应的第二隐向量，第一隐向量和第二隐向量的差值，即上述指定动作对应的控制向量，即该指定动作对应的初始动作参数。以人脸图像为例，将同一人脸的不同表情(如不笑和笑)进行反向编码，所得隐向量之差即为“笑”这一人脸特征的控制向量，该控制向量可以作为一种初始动作参数；通过该方式，可以得到多种表情对应的初始动作参数，也可以得到多种动作对应的初始动作参数。

[0061] 该图像特征中包括初始图像的全局特征和细节特征，以包含人脸的初始图像为例，全局特征可以为人脸的性别、年龄等特征；细节特征可以为人脸的五官姿势、发丝纹理等特征。以上述尺度为(18,512)的编码向量为例，该编码向量包括18个维度(维度也可以称为通道)，每个维度的特征长度为512，全局特征可以包含在第1至10个维度的特征中，细节特征可以包含在第11至18个维度的特征中。

[0062] 为了避免第一动作参数的融入导致图像整体的相似性减弱，图像中目标对象的变化太大，造成视觉上的韵律效果较差，本实施例中，将第一动作参数与细节特征进行融合，通过第一动作参数仅改变图像中目标对象的局部细节，例如，人脸的嘴部姿态、眼部姿态等局部的动作；同时图像其他区域以及图像的整体视觉效果不变，从而提高图像随音频变化的律动效果。

[0063] 将第一动作参数与细节特征进行融合处理时，一种具体的实现方式中，从图像特征中获取第一指定通道中的特征数据；其中，第一指定通道中的特征数据包含细节特征；从

第一动作参数中获取第二指定通道中的参数数据;将第一指定通道中的特征数据和第二指定通道中的参数数据进行逐点相加处理,得到融合特征。

[0064] 上述图像特征中通常包括多通道的特征数据,每个通道的特征数据所包含的特征类型不同,基于此,可以将包含细节特征的通道作为第一指定通道,并获取第一指定通道中的特征数据。假如图像特征包括18个通道,排序靠前的通道中的特征数据包含全局特征,也可以称为图像的大尺度特征,排序靠后的通道中的特征数据包含细节特征;此时,可以选择部分排序靠后的通道作为第一指定通道,例如可以选择最后8个通道作为第一指定通道。仅将第一动作参数融入包含细节特征的图像特征中,从而实现仅在图像的细节处进行变化,不影响图像的全局效果。以人脸图像为例,该方式可以对人脸的细微表情进行控制,但对人脸的其他特征保持不变。

[0065] 另外,第一动作参数通常也包括多通道的参数数据,为了便于特征融合,本实施例中,仅选取第一动作参数中一部分通道的参数数据融入至图像特征中。假如第一动作参数也包括18个通道,则可以选择最后8个通道作为第二指定通道。一种具体的实现方式中,上述第一指定通道中的特征数据和上述第二指定通道中的参数数据的通道数可以相同,每个通道的特征尺度也可以相同,在该情况下,将第一指定通道中的特征数据和第二指定通道中的参数数据进行逐点相加处理,得到融合特征。逐点相加可以理解为,对于第一指定通道中的某一位置上的特征值,获取第二指定通道中该位置上的参数值,将该特征值与参数值相加,即相加的两个数值的位置在第一指定通道和第二指定通道中相同或相对应。除了逐点相加的融合方式以外,还可以有逐点相乘或其他特征融合方式。

[0066] 为了便于理解,图3示出了基于音频特征调整图像特征的流程示意图;以人脸图像为例,初始动作参数也可以称为表情控制编码。输入目标音频的音频特征,即音频特征的编码向量组W_c,对W_c中时间点相邻的两个特征向量之间插值,得到插值后的特征向量组W_c';对W_c'中的每个时间点对应的特征向量求均值,得到一维数据形式的权重向量,该权重向量中,每个时间点对应一个权重值。通过该权重向量对表情控制编码We进行放缩;将放缩后的表情控制编码与人脸图像编码W_f叠加得到W_f';这里的人脸图像编码可以理解为上述图像特征的具体示例;最后输出具有时序性的音频控制人脸编码W_f'。该音频控制人脸编码包括前述实施例中的融合特征和全局特征。

[0067] 得到融合特征后,基于该融合特征和全局特征,生成目标图像。具体的,将融合特征和全局特征作为隐向量,输入至预先训练完成的图像生成网络中,输出目标图像;其中,该隐向量用于:控制图形生成网络输出与隐向量指示的特征相匹配的图像。该图像生成网络可以通过StyleGAN网络实现,图4所示为StyleGAN网络的示意图。该StyleGAN网络可以生成分辨率较高的图像,该网络将一个随机噪声z通过八层的全连接网络进行解耦,转化为隐向量w。w作为输入A部分控制生成图片的整体风格,噪音Noise作为输入B部分控制生成细节(如生成人脸的头发丝等),共同输入生成器synthesis network,在与鉴别器的博弈中完成对图像的生成。具体到本实施例中,上述融合特征和全局特征作为隐向量w,输入至图3中A部分,从而控制生成器生成目标图像。

[0068] 图5示出了在播放目标音频的过程中,生成多张目标图像,从而生成视频的过程。首先,加载StyleGAN网络模型;输入音频控制的人脸编码,通过StyleGAN网络生成目标图像,将图像按照时序排列生成视频;判断目标音频是否结束,如果否,继续执行输入音频控

制的人脸编码的步骤，如果结束，则结束生成视频的流程。图6示出了人脸图像随着音频律动变化的示意图；以嘴部动作为例，随着音频的播放，人脸做出不笑、微笑和大笑的表情，实现音频驱动图像中对象细微动作的变化。

[0069] 通过上述方式，可以实现音频律动的可视化，用户可以输入任意图像，如自己的照片等，实现音乐对图像中对象的细微动作控制，可以获得更高的体验感，优化了音乐播放产品的界面视觉体验。

[0070] 对应于上述方法实施例，参见图7所示的一种图像生成装置的结构示意图，该装置包括：

[0071] 参数调整模块70，用于基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；其中，第一动作参数指示的动作与音频特征相匹配；

[0072] 图像生成模块72，用于基于第一动作参数以及包含目标对象的初始图像，生成目标图像；目标图像中，目标对象具有第一动作参数指示的动作。

[0073] 上述图像生成装置，首先基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；该第一动作参数指示的动作与音频特征相匹配；然后基于第一动作参数以及包含目标对象的初始图像，生成目标图像；该目标图像中的目标对象具有第一动作参数指示的动作。该方式中，通过音频的音频特征调整动作参数，可以使得到的第一动作参数所指示的动作与该音频特征相匹配，进而使生成的图像中的目标对象具有第一动作参数指示的动作，因而该方式可以通过音频控制图像中对象的动作，使图像中对象的动作随着音频内容的变化而变化，在播放音频的过程中，图像内容变化多样，提高了用户的视觉体验。

[0074] 上述目标对象包括人脸；第一动作参数指示的动作包括人脸的表情动作。

[0075] 上述目标音频的音频特征用于调整初始动作参数指示的动作的动作幅度；第一动作参数指示的动作的动作幅度与音频特征相匹配。

[0076] 上述参数调整模块，还用于：根据目标音频的音频特征，确定参数调整权重；基于参数调整权重，对初始动作参数进行放缩处理，得到第一动作参数。

[0077] 上述参数调整模块，还用于：在音频特征的时间维度上，对时间维度上的各个时间点对应的特征向量求取平均值，得到初始参数；将初始参数映射至预设的数值范围中，得到参数调整权重。

[0078] 上述装置还包括：插值模块，用于：对音频特征中，任意两个相邻的初始时间点之间插入指定数量的中间时间点，以及每个中间时间点对应的特征向量，得到最终的音频特征；其中，中间时间点对应的特征向量，基于与中间时间点相邻的两个初始时间点对应的特征向量确定。

[0079] 上述装置还包括特征提取模块，用于通过下述方式得到目标音频的音频特征：提取目标音频的梅尔频率倒谱系数MFCC参数；MFCC参数包括预设时间间隔的多个时间点，以及每个时间点对应一个MFCC数值；将MFCC参数输入至预先训练完成的特征提取网络中，输出目标音频的音频特征。

[0080] 上述特征提取网络包括多个依次串联的特征提取模块；特征提取模块包括卷积层、批量归一化层和激活函数层。

[0081] 上述装置还包括：预处理模块，用于：基于预设的填充值，对MFCC参数的频率维度上的数值进行数值填充，以使频率维度上的数值数量与MFCC参数的时间维度上的数值数量

相匹配;将数值填充后的MFCC参数进行复制,得到指定通道数量的MFCC参数。

[0082] 上述装置还包括网络训练模块,用于下述方式训练得到特征提取网络:将样本音频的MFCC参数输入至编码网络中,输出样本音频的特征向量;将样品音频的特征向量输入至解码网络中,得到样本音频的输出参数;基于预设的损失函数,计算输出参数与样本音频的MFCC参数之间的损失值,基于损失值训练编码网络和解码网络,将训练完成后的编码网络确定为特征提取网络。

[0083] 上述解码网络包括多个依次串联的解码模块;解码模块包括转置卷积层、批量归一化层和激活函数层。

[0084] 上述图像生成模块还包括:提取包含目标对象的初始图像的图像特征;其中,图像特征包括全局特征和细节特征;将第一动作参数与细节特征进行融合处理,得到融合特征;基于融合特征和全局特征,生成目标图像。

[0085] 上述图像生成模块还包括:从图像特征中获取第一指定通道中的特征数据;其中,第一指定通道中的特征数据包含细节特征;从第一动作参数中获取第二指定通道中的参数数据;将第一指定通道中的特征数据和第二指定通道中的参数数据进行逐点相加处理,得到融合特征。

[0086] 上述图像生成模块还包括:将融合特征和全局特征作为隐向量,输入至预先训练完成的图像生成网络中,输出目标图像;其中,隐向量用于:控制图形生成网络输出与隐向量指示的特征相匹配的图像。

[0087] 本实施例还提供一种电子设备,包括处理器和存储器,存储器存储有能够被处理器执行的机器可执行指令,处理器执行机器可执行指令以实现上述图像生成方法。该电子设备可以是服务器,也可以是终端设备。

[0088] 参见图8所示,该电子设备包括处理器100和存储器101,该存储器101存储有能够被处理器100执行的机器可执行指令,该处理器100执行机器可执行指令以实现上述图像生成方法。

[0089] 进一步地,图8所示的电子设备还包括总线102和通信接口103,处理器100、通信接口103和存储器101通过总线102连接。

[0090] 其中,存储器101可能包含高速随机存取存储器(RAM,Random Access Memory),也可能还包括非不稳定的存储器(non-volatile memory),例如至少一个磁盘存储器。通过至少一个通信接口103(可以是有线或者无线)实现该系统网元与至少一个其他网元之间的通信连接,可以使用互联网,广域网,本地网,城域网等。总线102可以是ISA总线、PCI总线或EISA总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示,图8中仅用一个双向箭头表示,但并不表示仅有一根总线或一种类型的总线。

[0091] 处理器100可能是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器100中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器100可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、网络处理器(Network Processor,简称NP)等;还可以是数字信号处理器(Digital Signal Processor,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)、现场可编程门阵列(Field-Programmable Gate Array,简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实现或者执行本

发明实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本发明实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器101，处理器100读取存储器101中的信息，结合其硬件完成前述实施例的方法的步骤。

[0092] 本实施例还提供一种机器可读存储介质，机器可读存储介质存储有机器可执行指令，机器可执行指令在被处理器调用和执行时，机器可执行指令促使处理器实现上述图像生成方法。

[0093] 本发明实施例所提供的图像生成方法、装置、电子设备及存储介质的计算机程序产品，包括存储了程序代码的计算机可读存储介质，所述程序代码包括的指令可用于执行前面方法实施例中所述的方法，具体实现可参见方法实施例，在此不再赘述。

[0094] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统和装置的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0095] 另外，在本发明实施例的描述中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。

[0096] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0097] 在本发明的描述中，需要说明的是，术语“中心”、“上”、“下”、“左”、“右”、“竖直”、“水平”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”、“第二”、“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性。

[0098] 最后应说明的是：以上实施例，仅为本发明的具体实施方式，用以说明本发明的技术方案，而非对其限制，本发明的保护范围并不局限于此，尽管参照前述实施例对本发明进行了详细的说明，本领域技术人员应当理解：任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化，或者对其中部分技术特征进行等同替换；而这些修改、变化或者替换，并不使相应技术方案的本质脱离本发明实施例技术方案的精神和范围，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

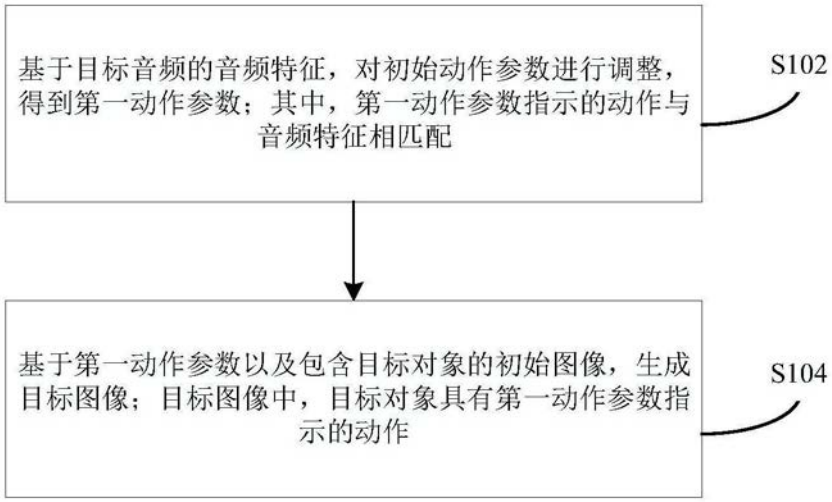

基于目标音频的音频特征，对初始动作参数进行调整，得到第一动作参数；其中，第一动作参数指示的动作与音频特征相匹配
S102
基于第一动作参数以及包含目标对象的初始图像，生成目标图像；目标图像中，目标对象具有第一动作参数指示的动作
S104

图1

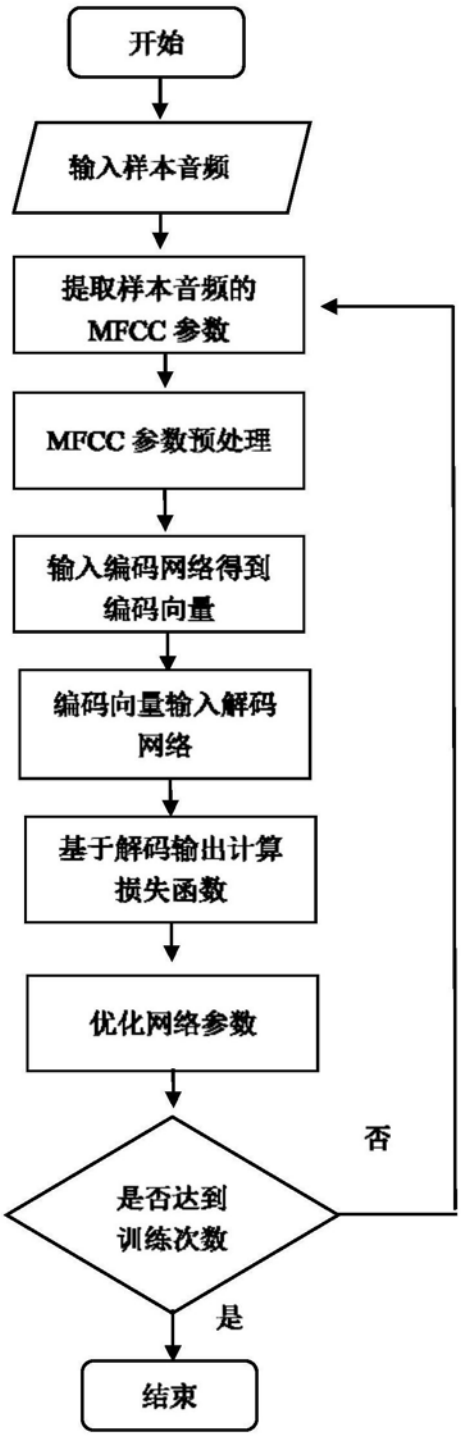

开始
输入样本音频
提取样本音频的MFCC参数
MFCC参数预处理
输入编码网络得到编码向量
编码向量输入解码网络
基于解码输出计算损失函数
优化网络参数
是否达到训练次数
否
是
结束

图2

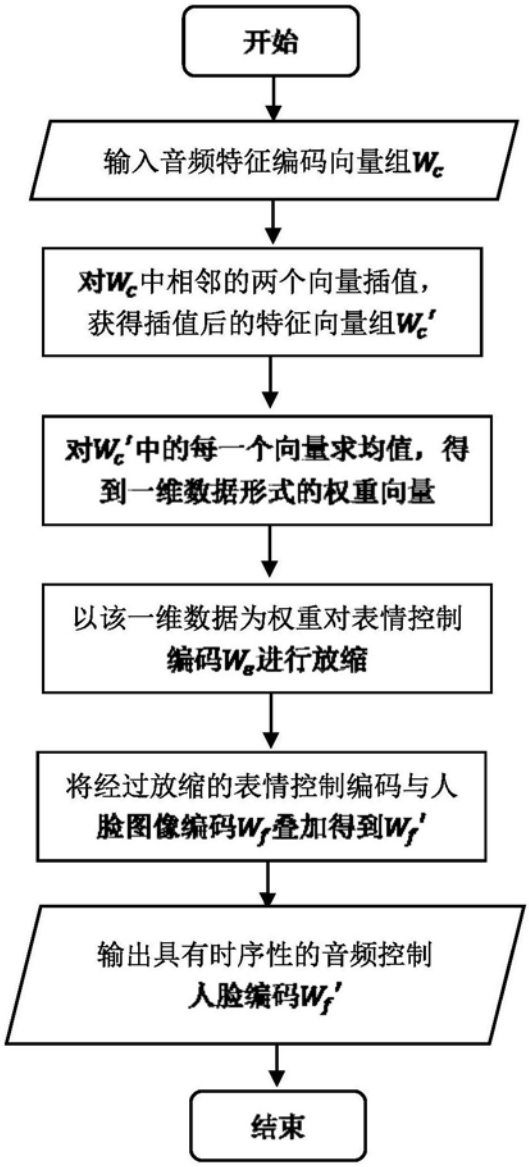

开始
输入音频特征编码向量组W_c
对W_c中相邻的两个向量插值，获得插值后的特征向量组W_c'
对W_c'中的每一个向量求均值，得到一维数据形式的权重向量
以该一维数据为权重对表情控制编码W_e进行放缩
将经过放缩的表情控制编码与人脸图像编码W_f叠加得到W_f'
输出具有时序性的音频控制人脸编码W_f'
结束

图3

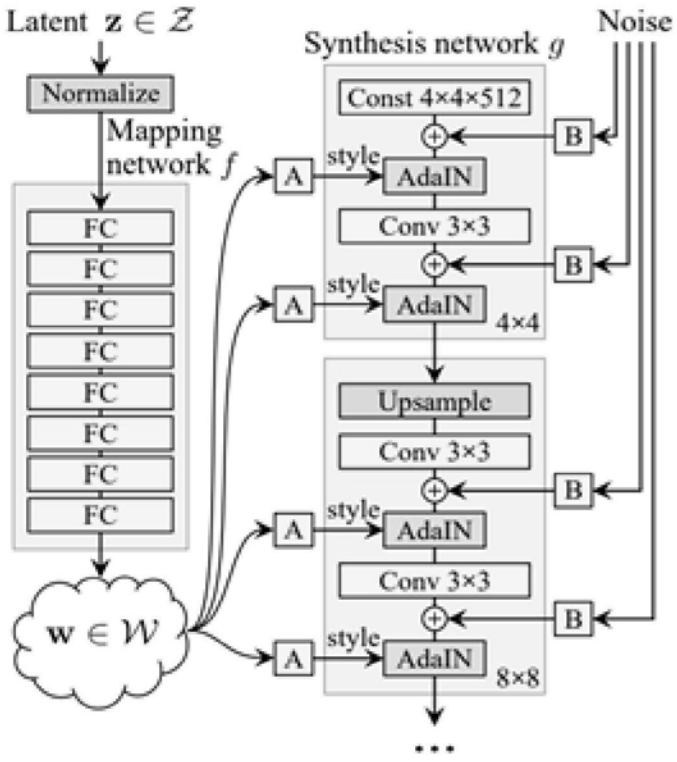

Latent $\mathbf{z} \in \mathcal{Z}$
Normalize
Mapping network f
FC
FC
FC
FC
FC
FC
FC
FC
$\mathbf{w} \in \mathcal{W}$
Synthesis network g
Noise
Const 4×4×512
B
style
A
AdaIN
Conv 3×3
4×4
Upsample
8×8

图4

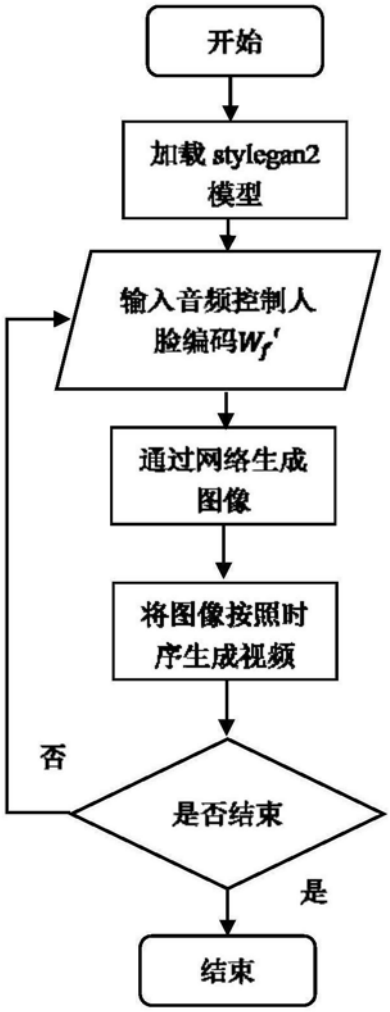

开始
加载 stylegan2 模型
输入音频控制人脸编码W_f'
通过网络生成图像
将图像按照时序生成视频
否
是否结束
是
结束

图5

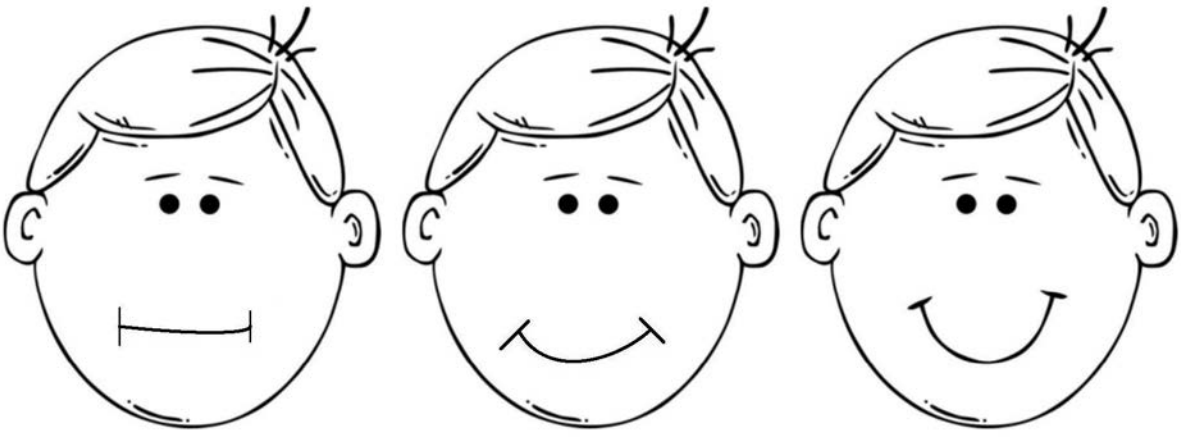


图6

图7

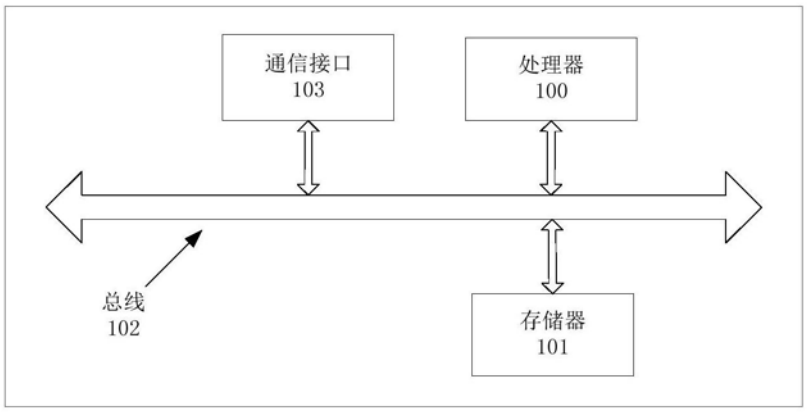


图8