(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116320216 A
(43)申请公布日 2023.06.23

(21)申请号 202310252631.1
(22)申请日 2023.03.15
(71)申请人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层
(72)发明人 李鑫 刘芳龙 袁苇航 张琦 李甫 王井东 冯浩城 丁二锐 王海峰 吴甜
(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 孔凡红
(51)Int.Cl.
H04N 5/262(2006.01)
G06T 7/269(2017.01)
G06V 10/82(2022.01)
G06V 10/774(2022.01)
H04N 19/44(2014.01)

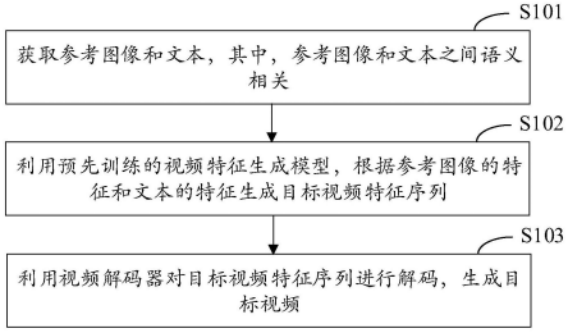
权利要求书6页 说明书12页 附图7页

(54)发明名称

视频生成方法、模型的训练方法、装置、设备和介质

(57)摘要

本公开提供了一种视频生成方法、模型的训练方法、装置、设备和介质，涉及人工智能技术领域，具体为计算机视觉、深度学习等技术领域，可应用于AIGC等场景。具体实现方案为：获取参考图像和文本，其中，所述参考图像和所述文本之间语义相关；利用预先训练的视频特征生成模型，根据所述参考图像的特征和所述文本的特征生成目标视频特征序列；利用视频解码器对所述目标视频特征序列进行解码，生成目标视频。本公开可以提高生成视频的质量。



CN 116320216 A

1.一种视频生成方法,包括:

获取参考图像和文本,其中,所述参考图像和所述文本之间语义相关;

利用预先训练的视频特征生成模型,根据所述参考图像的特征和所述文本的特征生成目标视频特征序列;

利用视频解码器对所述目标视频特征序列进行解码,生成目标视频。

2.根据权利要求1所述的方法,其中,所述参考图像是利用预先训练的文本生成图像模型根据所述文本生成的。

3.根据权利要求1所述的方法,还包括:

从原视频中提取参考帧,对所述参考帧进行图像编辑,得到所述参考图像。

4.根据权利要求1或2所述的方法,其中,所述文本包括用户输入的文本。

5.根据权利要求1或3所述的方法,其中,所述文本是利用预先训练的图像生成文本模型根据所述参考图像生成的。

6.根据权利要求1-3任选一项所述的方法,其中,所述参考图像的特征是通过图像编码器提取,所述文本的特征是通过文本编码器提取。

7.根据权利要求1-3任选一项所述的方法,其中,所述视频特征生成模型是基于扩散模型或者对抗生成网络训练得到。

8.根据权利要求1或7所述的方法,其中,所述视频特征生成模型是由多个视频特征生成子模型组成的级联模型。

9.根据权利要求8所述的方法,其中,所述级联模型包括N级视频特征生成子模型,上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入,其中,N为大于1的自然数。

10.根据权利要求9所述的方法,其中,所述利用预先训练的视频特征生成模型,根据所述参考图像的特征和所述文本的特征生成目标视频特征序列,包括:

通过对所述参考图像进行下采样,获取包括所述参考图像在内的N个不同分辨率级别的参考图像;

分别提取所述N个不同分辨率级别的参考图像的特征;

将第一级分辨率的参考图像的特征和所述文本的特征输入第一级视频特征生成子模型;

针对所述级联模型中除所述第一级视频特征生成子模型之外的任意当前级视频特征生成子模型,按如下方式处理:将所述当前级视频特征生成子模型的上一级视频特征生成子模型输出的视频特征序列进行上采样,并将经所述上采样得到的视频特征序列、当前级分辨率的参考图像的特征和所述文本的特征输入所述当前级视频特征生成子模型;

将第N级视频特征生成子模型的输出作为所述目标视频特征序列。

11.根据权利要求1所述的方法,还包括:

对所述目标视频特征序列进行插帧;

相应的,所述利用视频解码器对所述目标视频特征序列进行解码,生成目标视频,包括:

利用视频解码器对所述插帧后的目标视频特征序列进行解码,生成目标视频。

12.根据权利要求11所述的方法,其中,所述对所述目标视频特征序列进行插帧,包括:

利用预先训练的光流预测模型预测所述目标视频特征序列的光流；

根据所述光流，对所述目标视频特征序列在特征空间进行插帧，得到多个初始化插帧；

利用预先训练的微调模型对所述多个初始化插帧进行微调。

13.根据权利要求1所述的方法，其中，所述视频解码器中添加有时序卷积模块和注意力模块。

14.根据权利要求1所述的方法，其中，在所述视频解码器的训练过程中，是对输入的视频样本进行了退化处理，其中，所述退化处理包括模糊、加噪和压缩。

15.一种视频特征生成模型的训练方法，包括：

从视频训练样本中获取参考图像样本，并获取与所述参考图像样本语义相关的文本样本；

分别提取所述参考图像样本的特征和所述文本样本的特征，并提取所述视频训练样本的视频特征序列样本；

将所述参考图像样本的特征和所述文本样本的特征作为所述视频特征生成模型的输入，将所述视频特征序列样本作为所述视频特征生成模型的输出，对所述视频特征生成模型进行训练；

其中，所述视频特征生成模型为由多个视频特征生成子模型组成的级联模型。

16.根据权利要求15所述的方法，其中，所述级联模型中的各个视频特征生成子模型是分别进行训练的。

17.根据权利要求15所述的方法，其中，所述级联模型包括N级视频特征生成子模型，上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入；

相应的，所述对所述视频特征生成模型进行训练，包括：

利用预先训练的文本生成图像模型初始化第一级视频特征生成子模型，并训练所述第一级视频特征生成子模型；

针对所述级联模型中除所述第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式进行训练：

利用所述当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对所述当前级视频特征生成子模型初始化后再进行训练；

其中，各级视频特征生成子模型的结构相同，N为大于1的自然数。

18.根据权利要求17所述的方法，其中，所述训练所述第一级视频特征生成子模型，包括：

通过对所述参考图像样本进行下采样，获取包括所述参考图像样本在内的N个不同分辨率级别的参考图像样本；

将第一级分辨率的参考图像样本的特征和所述文本样本的特征作为所述第一级视频特征生成子模型的输入，提取所述视频训练样本的第一级分辨率的视频特征作为所述第一级视频特征生成子模型的输出，对所述第一级视频特征生成子模型进行训练。

19.根据权利要求18所述的方法，其中，所述利用所述当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对所述当前级视频特征生成子模型初始化后再进行训练，包括：

利用训练后得到的所述上一级视频特征生成子模型的模型参数初始化所述当前级视

频特征生成子模型；

对所述视频训练样本的上一级分辨率的视频特征进行上采样；

将当前级分辨率的参考图像的特征、所述文本特征和所述上采样得到的视频特征作为所述当前级视频特征生成子模型的输入，提取所述视频训练样本的当前级分辨率的视频特征作为所述当前级视频特征生成子模型的输出，对所述当前级视频特征生成子模型进行训练。

20.根据权利要求15所述的方法，其中，所述从视频训练样本中获取参考图像样本，包括：

从所述视频训练样本中提取参考帧；

对所述参考帧进行图像编辑，得到所述参考图像样本。

21.根据权利要求17-20任选一项所述的方法，其中，所述文本生成图像模型包括卷积模块和空间注意力模块，所述方法还包括：

在所述文本生成图像模型的网络结构中，在每个2D卷积模块后面添加时间维度的1D卷积模块，在每个空间注意力模块之后添加时间注意力模块。

22.根据权利要求17-20任选一项所述的方法，还包括：

获取所述视频训练样本的视频条件信息，其中，所述视频条件信息至少包括深度图和目标关键点图；

将所述视频条件信息作为各级视频特征生成子模型的输入进行训练。

23.一种视频生成装置，包括：

参考图像与文本获取模块，用于获取参考图像和文本，其中，所述参考图像和所述文本之间语义相关；

视频特征序列生成模块，用于利用预先训练的视频特征生成模型，根据所述参考图像的特征和所述文本的特征生成目标视频特征序列；

视频生成模块，用于利用视频解码器对所述目标视频特征序列进行解码，生成目标视频。

24.根据权利要求23所述的装置，其中，所述参考图像是利用预先训练的文本生成图像模型根据所述文本生成的。

25.根据权利要求23所述的装置，还包括：

参考帧编辑模块，用于从原视频中提取参考帧，对所述参考帧进行图像编辑，得到所述参考图像。

26.根据权利要求23或24所述的装置，其中，所述文本包括用户输入的文本。

27.根据权利要求23或25所述的装置，其中，所述文本是利用预先训练的图像生成文本模型根据所述参考图像生成的。

28.根据权利要求23-25任选一项所述的装置，其中，所述参考图像的特征是通过图像编码器提取，所述文本的特征是通过文本编码器提取。

29.根据权利要求23-25任选一项所述的装置，其中，所述视频特征生成模型是基于扩散模型或者对抗生成网络训练得到。

30.根据权利要求23或29所述的装置，其中，所述视频特征生成模型是由多个视频特征生成子模型组成的级联模型。

31.根据权利要求30所述的装置,其中,所述级联模型包括N级视频特征生成子模型,上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入,其中,N为大于1的自然数。

32.根据权利要求31所述的装置,其中,所述视频特征序列生成模块包括:

下采样单元,用于通过对所述参考图像进行下采样,获取包括所述参考图像在内的N个不同分辨率级别的参考图像;

特征提取单元,用于分别提取所述N个不同分辨率级别的参考图像的特征;

第一级视频特征生成子模型处理单元,用于将第一级分辨率的参考图像的特征和所述文本的特征输入第一级视频特征生成子模型;

当前级视频特征生成子模型处理单元,用于针对所述级联模型中除所述第一级视频特征生成子模型之外的任意当前级视频特征生成子模型,按如下方式处理:将所述当前级视频特征生成子模型的上一级视频特征生成子模型输出的视频特征序列进行上采样,并将经所述上采样得到的视频特征序列、当前级分辨率的参考图像的特征和所述文本的特征输入所述当前级视频特征生成子模型;

输出单元,用于将第N级视频特征生成子模型的输出作为所述目标视频特征序列。

33.根据权利要求23所述的装置,还包括:

插帧模块,用于对所述目标视频特征序列进行插帧;

相应的,所述视频生成模块具体用于:

利用视频解码器对所述插帧后的目标视频特征序列进行解码,生成目标视频。

34.根据权利要求33所述的装置,其中,所述插帧模块包括:

光流预测单元,用于利用预先训练的光流预测模型预测所述目标视频特征序列的光流;

插帧单元,用于根据所述光流,对所述目标视频特征序列在特征空间进行插帧,得到多个初始化插帧;

微调单元,用于利用预先训练的微调模型对所述多个初始化插帧进行微调。

35.根据权利要求23所述的装置,其中,所述视频解码器中添加有时序卷积模块和注意力模块。

36.根据权利要求23所述的装置,其中,在所述视频解码器的训练过程中,是对输入的视频样本进行了退化处理,其中,所述退化处理包括模糊、加噪和压缩。

37.一种视频特征生成模型的训练装置,包括:

获取模块,用于从视频训练样本中获取参考图像样本,并获取与所述参考图像样本语义相关的文本样本;

特征提取模块,用于分别提取所述参考图像样本的特征和所述文本样本的特征,并提取所述视频训练样本的视频特征序列样本;

模型训练模块,用于将所述参考图像样本的特征和所述文本样本的特征作为所述视频特征生成模型的输入,将所述视频特征序列样本作为所述视频特征生成模型的输出,对所述视频特征生成模型进行训练;

其中,所述视频特征生成模型为由多个视频特征生成子模型组成的级联模型。

38.根据权利要求37所述的装置,其中,所述级联模型中的各个视频特征生成子模型是

分别进行训练的。

39.根据权利要求37所述的装置，其中，所述级联模型包括N级视频特征生成子模型，上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入；

相应的，所述模型训练模块包括：

第一级视频特征生成子模型训练单元，用于利用预先训练的文本生成图像模型初始化第一级视频特征生成子模型，并训练所述第一级视频特征生成子模型；

当前级视频特征生成子模型训练单元，用于针对所述级联模型中除所述第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式进行训练：

利用所述当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对所述当前级视频特征生成子模型初始化后再进行训练；

其中，各级视频特征生成子模型的结构相同，N为大于1的自然数。

40.根据权利要求39所述的装置，其中，所述第一级视频特征生成子模型训练单元包括：

下采样子单元，用于通过对所述参考图像样本进行下采样，获取包括所述参考图像样本在内的N个不同分辨率级别的参考图像样本；

第一级视频特征生成子模型训练子单元，用于将第一级分辨率的参考图像样本的特征和所述文本样本的特征作为所述第一级视频特征生成子模型的输入，提取所述视频训练样本的第一级分辨率的视频特征作为所述第一级视频特征生成子模型的输出，对所述第一级视频特征生成子模型进行训练。

41.根据权利要求40所述的装置，其中，所述当前级视频特征生成子模型训练单元包括：

初始化子单元，用于利用训练后得到的所述上一级视频特征生成子模型的模型参数初始化所述当前级视频特征生成子模型；

上采样子单元，用于对所述视频训练样本的上一级分辨率的视频特征进行上采样；

当前级视频特征生成子模型训练子单元，用于将当前级分辨率的参考图像的特征、所述文本特征和所述上采样得到的视频特征作为所述当前级视频特征生成子模型的输入，提取所述视频训练样本的当前级分辨率的视频特征作为所述当前级视频特征生成子模型的输出，对所述当前级视频特征生成子模型进行训练。

42.根据权利要求37所述的装置，其中，所述获取模块包括：

参考帧获取单元，用于从所述视频训练样本中提取参考帧；

图像编辑单元，用于对所述参考帧进行图像编辑，得到所述参考图像样本。

43.根据权利要求39-42任选一项所述的装置，其中，所述文本生成图像模型包括卷积模块和空间注意力模块，所述装置还包括：

网络结构处理单元，用于在所述文本生成图像模型的网络结构中，在每个2D卷积模块后面添加时间维度的1D卷积模块，在每个空间注意力模块之后添加时间注意力模块。

44.根据权利要求39-42任选一项所述的装置，还包括视频条件信息处理模块，具体用于：

获取所述视频训练样本的视频条件信息，其中，所述视频条件信息至少包括深度图和目标关键点图；

将所述视频条件信息作为各级视频特征生成子模型的输入进行训练。

45.一种电子设备,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-14中任一项所述的视频生成方法,或权利要求15-22中任一项所述的视频特征生成模型的训练方法。

46.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使计算机执行根据权利要求1-14中任一项所述的视频生成方法,或权利要求15-22中任一项所述的视频特征生成模型的训练方法。

47.一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现根据权利要求1-14中任一项所述的视频生成方法,或权利要求15-22中任一项所述的视频特征生成模型的训练方法。

视频生成方法、模型的训练方法、装置、设备和介质

技术领域

[0001] 本公开涉及人工智能技术领域，具体为计算机视觉、深度学习等技术领域，可应用于AIGC等场景。具体涉及一种视频生成方法、模型的训练方法、装置、设备和介质。

背景技术

[0002] AI生成视频是当前非常热门的话题。通过AI生成视频相比传统人工制作视频，在效率上带来了革命性的提升。而且，AI生成视频可以用于影视、动漫和人机交互等各种不同的应用场景。

[0003] 现有的AI生成视频技术中，生成的视频比较模糊，内容可控性较差，视频内容质量较低，从而影响视频的质量和效果。

发明内容

[0004] 本公开提供了一种视频生成方法、模型的训练方法、装置、设备和介质。

[0005] 根据本公开的一方面，提供了一种视频生成方法，包括：

[0006] 获取参考图像和文本，其中，所述参考图像和所述文本之间语义相关；

[0007] 利用预先训练的视频特征生成模型，根据所述参考图像的特征和所述文本的特征生成目标视频特征序列；

[0008] 利用视频解码器对所述目标视频特征序列进行解码，生成目标视频。

[0009] 根据本公开的另一方面，提供了一种视频特征生成模型的训练方法，

[0010] 根据本公开的另一方面，提供了一种视频生成装置，包括：

[0011] 参考图像与文本获取模块，用于获取参考图像和文本，其中，所述参考图像和所述文本之间语义相关；

[0012] 视频特征序列生成模块，用于利用预先训练的视频特征生成模型，根据所述参考图像的特征和所述文本的特征生成目标视频特征序列；

[0013] 视频生成模块，用于利用视频解码器对所述目标视频特征序列进行解码，生成目标视频。

[0014] 根据本公开的另一方面，提供了一种视频特征生成模型的训练装置，包括：

[0015] 根据本公开的另一方面，提供了一种电子设备，包括：

[0016] 至少一个处理器；以及

[0017] 与所述至少一个处理器通信连接的存储器；其中，

[0018] 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行本公开任意实施例所述的视频生成方法或视频特征生成模型的训练方法。

[0019] 根据本公开的另一方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使计算机执行本公开任意实施例所述的视频生成方法或视频特征生成模型的训练方法。

[0020] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0021] 附图用于更好地理解本方案，不构成对本公开的限定。其中：

[0022] 图1是根据本公开实施例的视频生成方法的流程示意图；

[0023] 图2是根据本公开实施例的视频生成方法的流程示意图；

[0024] 图3是根据本公开实施例的视频特征生成模型的训练方法的流程示意图；

[0025] 图4是根据本公开实施例的N级视频特征生成子模型的训练过程的示意图；

[0026] 图5是根据本公开实施例的视频生成方法的流程示意图；

[0027] 图6是根据本公开实施例的视频生成方法的预测流程图；

[0028] 图7是根据本公开实施例的视频生成装置的结构示意图；

[0029] 图8是根据本公开实施例的视频特征生成模型的训练装置的结构示意图；

[0030] 图9是用来实现本公开实施例的视频生成方法的电子设备的框图。

具体实施方式

[0031] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0032] 图1是根据本公开实施例的视频生成方法的流程示意图，本实施例可适用于自动生成视频的情况，例如根据文本生成视频，根据图像生成视频，或者根据一个输入视频生成一个新视频等，涉及人工智能技术领域，具体为计算机视觉、深度学习等技术领域，可应用于AIGC等场景。该方法可由一种视频生成装置来执行，该装置采用软件和/或硬件的方式实现，优选是配置于电子设备中，例如计算机设备、服务器或智能终端等。如图1所示，该方法具体包括如下：

[0033] S101、获取参考图像和文本，其中，参考图像和文本之间语义相关。

[0034] 在文本生成视频的应用中，本公开可以基于给定文本自动生成一段对应内容的视频。该文本可以是用户输入的一段文本，例如，“往杯子里倒水”、“火柴燃烧”或者“沙滩上行走的企鹅”等，当然，也可以是基于用户说出的语音转化成的文本。本公开对获取文本的形式以及文本的内容不作任何限定。然后，可以利用预先训练的文本生成图像模型来根据这段文本生成参考图像。基于文本生成图像模型也可以生成一个高质量的参考图像，为后续指导视频的生成提供了基础。

[0035] 在图像生成视频的应用中，本公开可以基于给定图像让图像里面的内容动起来，从而自动生成一段视频。其中，给定图像可以作为参考图像，并利用预先训练的图像生成文本模型根据参考图像生成语义相关的文本。

[0036] 在视频生成视频的应用中，本公开可以基于给定视频生成一个新视频，例如将一个真实场景下的视频在内容不改变的情况下转成动漫风格的视频。此种应用中，参考图像可以是从输入的原视频中提取一个参考帧并通过图像编辑得到，例如原视频中的第一帧。

其中，图像编辑可以基于预先训练的文本生成图像的模型来实现，此处不再赘述。通过图像编辑，不仅可以提高参考图像的质量，还可以改变原始图像的风格，或者基于其他应用上的需求改变原始图像，生成一个新图像，以便基于需求生成视频，例如编辑视频的背景或者改变视频中人物的服饰等。

[0037] S102、利用预先训练的视频特征生成模型，根据参考图像的特征和文本的特征生成目标视频特征序列。

[0038] 其中，所述视频特征生成模型例如是基于扩散模型或者对抗生成网络训练得到，用于根据参考图像的特征和文本的特征预测多帧连续的视频特征，即视频特征序列，例如16帧视频特征。训练时，可以从视频训练样本中获取参考图像样本及其语义相关的文本样本，并分别利用图像解码器和文本解码器提取参考图像样本的特征和文本样本的特征，然后，将参考图像样本的特征和文本样本的特征作为视频特征生成模型的输入，将视频训练样本的视频特征序列作为视频特征生成模型的输出，对视频特征生成模型进行训练。其中，从视频训练样本提取视频特征序列可以采用任一种现有技术实现，此处不再赘述。示例性的，该视频特征生成模型Unet结构(一种U型神经网络结构)的模型。

[0039] S103、利用视频解码器对目标视频特征序列进行解码，生成目标视频。

[0040] 获取到要生成的目标视频的目标视频特征序列之后，利用解码器进行解码即可生成目标视频。

[0041] 本公开实施例的技术方案，通过先获取一个高质量的参考图像，然后利用视频特征生成模型基于参考图像指导视频的生成，使得生成的视频内容可控且质量大大提升，继而提升视频效果。

[0042] 图2是根据本公开实施例的视频生成方法的流程示意图，本实施例在上述实施例的基础上，进一步对视频特征生成模型进行优化，其中，视频特征生成模型是由多个视频特征生成子模型组成的级联模型，所述级联模型包括N级视频特征生成子模型，上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入，N为大于1的自然数。相应的，如图2所示，该方法具体包括如下：

[0043] S201、获取参考图像和文本，其中，参考图像和文本之间语义相关。

[0044] S202、通过对参考图像进行下采样，获取包括参考图像在内的N个不同分辨率级别的参考图像。

[0045] 例如，在N个不同分辨率级别的参考图像中，从第一级到第N级，分辨率逐级增大。

[0046] S203、分别提取N个不同分辨率级别的参考图像的特征。

[0047] 例如，对原始的参考图像经图像编码器进行特征提取，得到第N级分辨率的参考图像特征，也即分辨率最大的图像特征。对原始的参考图像进行一次下采样，降低原始参考图像的分辨率，然后再经图像编码器进行特征提取，即可得到第N-1级分辨率的参考图像特征。以此类推，当继续进行第二次下采样和特征提取，即可得到更低分辨率级别的参考图像的特征，第一级分辨率则最低。

[0048] 在一种实施方式中，以三级级联模型为例，将原始参考图像依次经过两次下采样得到第一级分辨率的参考图像，经图像编码器即可得到第一级分辨率的参考图像的特征。将原始参考图像经一次下采样和特征提取，得到的是第二级分辨率的参考图像的特征。将原始参考图像直接经图像编码器提取特征，得到的是第三级分辨率的参考图像的特征。

[0049] S204、将第一级分辨率的参考图像的特征和文本的特征输入第一级视频特征生成子模型。

[0050] 其中，第一级视频特征生成子模型是级联模型中的第一个模型，输入的文本特征与其他视频特征生成子模型相同，输入的图像特征则是分辨率最低的参考图像的特征。第一级视频特征生成子模型根据第一级分辨率的参考图像的特征和文本的特征得到对应分辨率级别的视频特征序列。

[0051] S205、针对级联模型中除第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式处理：将当前级视频特征生成子模型的上一级视频特征生成子模型输出的视频特征序列进行上采样，并将经上采样得到的视频特征序列、当前级分辨率的参考图像的特征和文本的特征输入当前级视频特征生成子模型。

[0052] S206、将第N级视频特征生成子模型的输出作为目标视频特征序列。

[0053] S207、利用视频解码器对目标视频特征序列进行解码，生成目标视频。

[0054] 仍以三级级联模型为例，在得到第一级视频特征生成子模型输出的视频特征序列之后先进行上采样，然后将上采样后的视频特征序列、第二级分辨率的参考图像的特征和文本特征输入第二级视频特征生成子模型，同理，对第二级视频特征生成子模型输出的视频特征序列进行上采样，将经上采样得到的视频特征序列、第三级分辨率的参考图像的特征和文本特征输入第三级视频特征生成子模型，第三级视频特征生成子模型的输出即为最终的目标视频特征序列。其中，通过上采样可以将上一级的视频特征序列转为当前级对应分辨率的视频特征序列，例如可以通过双线性插值的方法实现。

[0055] 当N大于3时，对应级联的视频特征生成子模型数量增多，但处理方式与上述相同，此处不再赘述。在实际应用中，可以根据需求对所需的级联模型进行配置，本公开对此不作任何限定。

[0056] 在一种实施方式中，所述N级视频特征生成子模型中的各个视频特征生成子模型，是分别进行训练的。

[0057] 图3是根据本公开实施例的视频特征生成模型的训练方法的流程示意图。如图3所示，该方法包括：

[0058] S301、从视频训练样本中获取参考图像样本，并获取与参考图像样本语义相关的文本样本。

[0059] S302、分别提取参考图像样本的特征和文本样本的特征，并提取视频训练样本的视频特征序列样本。

[0060] S303、将参考图像样本的特征和文本样本的特征作为视频特征生成模型的输入，将视频特征序列样本作为视频特征生成模型的输出，对视频特征生成模型进行训练；其中，视频特征生成模型为由多个视频特征生成子模型组成的级联模型。

[0061] 其中，本公开并不限定视频训练样本的获取方式。在获取到视频训练样本后，可以从中提取第一帧，并通过图像编辑得到高质量的参考图像样本，同时，也可以通过图像编辑获取到满足不同需求的参考图像样本，例如，更改图像的风格等。

[0062] 在一种实施方式中，所述级联模型包括N级视频特征生成子模型，上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入。训练时，级联模型中的各个视频特征生成子模型可由是分别进行训练的，从而提高模型训练的灵活性。

[0063] 具体的，对视频特征生成模型进行训练，包括：

[0064] 利用预先训练的文本生成图像模型初始化第一级视频特征生成子模型，并训练所述第一级视频特征生成子模型；

[0065] 针对所述级联模型中除所述第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式进行训练：

[0066] 利用所述当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对所述当前级视频特征生成子模型初始化后再进行训练；

[0067] 其中，各级视频特征生成子模型的结构相同，N为大于1的自然数。

[0068] 所述训练第一级视频特征生成子模型，包括：

[0069] 通过对所述参考图像样本进行下采样，获取包括所述参考图像样本在内的N个不同分辨率级别的参考图像样本；

[0070] 将第一级分辨率的参考图像样本的特征和所述文本样本的特征作为所述第一级视频特征生成子模型的输入，提取所述视频训练样本的第一级分辨率的视频特征作为所述第一级视频特征生成子模型的输出，对所述第一级视频特征生成子模型进行训练。

[0071] 图4是根据本公开实施例的N级视频特征生成子模型的训练过程的示意图。如图4所示，所述N级视频特征生成子模型的训练过程包括：

[0072] S401、利用预先训练的文本生成图像模型初始化第一级视频特征生成子模型，并在所述文本生成图像模型的网络结构中，在每个2D卷积模块后面添加时间维度的1D卷积模块，在每个空间注意力模块之后添加时间注意力模块。

[0073] 其中，所述文本生成图像模型包括卷积模块和空间注意力模块。通过添加的时间维度卷积模块和时间注意力模块，可以使得视频特征生成子模型能拟合视频数据，对视频特征序列进行预测。

[0074] S402、对视频训练样本中的第一帧进行图像编辑，得到参考图像样本，并获取与参考图像样本语义相关的文本样本。

[0075] S403、通过对参考图像样本进行下采样，获取包括参考图像样本在内的N个不同分辨率级别的参考图像样本。

[0076] S404、将第一级分辨率的参考图像样本的特征和文本样本的特征作为第一级视频特征生成子模型的输入，提取视频训练样本的第一级分辨率的视频特征作为第一级视频特征生成子模型的输出，对第一级视频特征生成子模型进行训练。

[0077] S405、针对级联模型中除第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式进行训练：利用当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对当前级视频特征生成子模型进行初始化后再进行训练，其中，各级视频特征生成子模型的结构相同。

[0078] 也即，第一级视频特征生成子模型是利用预先训练好的文本生成图像模型的参数进行初始化后进行训练，然后后面各级视频特征生成子模型则是利用其前一级视频特征生成子模型训练后得到的模型参数进行初始化后再进行训练。这样，不仅可以提升模型训练整体的效率，同时，也可以提升其中每一级模型的预测效果。

[0079] 进一步的，利用当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对当前级视频特征生成子模型进行初始化后再进行训练，包括：

[0080] 利用训练后得到的上一级视频特征生成子模型的模型参数初始化当前级视频特征生成子模型；

[0081] 对视频训练样本的上一级分辨率的视频特征进行上采样；

[0082] 将当前级分辨率的参考图像的特征、文本特征和上采样得到的视频特征作为当前级视频特征生成子模型的输入，提取视频训练样本的当前级分辨率的视频特征作为当前级视频特征生成子模型的输出，对当前级视频特征生成子模型进行训练。

[0083] 由此，本公开实施例的技术方案，通过级联模型的实现方式，对预测任务进行拆解，通过各级联的视频特征生成子模型在对应的分辨率级别下对视频特征进行预测，每一级模型都是在上一级模型的基础上预测更多的视频特征的细节，再逐级叠加，从而得到内容准确且质量更高的预测结果。同时，对于每一级模型而言，其训练的难度更小，都在小分辨率的特征空间进行处理，大大减少了模型训练和测试需要的时间，提高了模型训练的效率。

[0084] 此外，在视频生成视频的应用下，还需要获取视频训练样本的视频条件信息，并将视频条件信息也作为各级视频特征生成子模型的输入进行训练。其中，视频条件信息至少包括深度图和目标关键点图等，从而生成原视频指导下的新视频，使得新视频在物体、人物等方面与原视频一致，而只是改变视频风格、背景或人物服饰等。

[0085] 图5是根据本公开实施例的视频生成方法的流程示意图。本实施例在上述实施例的基础上作出进一步优化。如图5所示，该方法包括：

[0086] S501、获取参考图像和文本，其中，参考图像和文本之间语义相关。

[0087] S502、利用预先训练的视频特征生成模型，根据参考图像的特征和文本的特征生成目标视频特征序列。

[0088] S503、对目标视频特征序列进行插帧。

[0089] S504、利用视频解码器对所述插帧后的目标视频特征序列进行解码，生成目标视频。

[0090] 插帧过程主要是指对视频在时间维度进行上采样，例如把一个2秒的视频通过插帧变成4秒甚至更长时间的视频，以增强视频观看效果。为了保证插的帧不模糊，时序上不抖动，本公开实施例采用光流指导的特征插帧方法。

[0091] 具体的，对目标视频特征序列进行插帧，包括：

[0092] 利用预先训练的光流预测模型预测目标视频特征序列的光流；

[0093] 根据光流，对目标视频特征序列在特征空间进行插帧，得到多个初始化插帧；

[0094] 利用预先训练的微调模型对多个初始化插帧进行微调。

[0095] 其中，光流预测模型的训练方式可以是利用一个已有的图像预测光流的模型作为监督。也即，先利用现有的图像预测光流的模型根据视频训练样本得到其光流，然后将视频训练样本的视频特征序列作为光流预测模型的输入，将对应的光流作为光流预测模型的输出，对光流预测模型进行训练，使得光流预测模型可以根据输入的任意视频特征序列预测其光流。

[0096] 接下来，根据光流，对目标视频特征序列在特征空间进行插帧，得到多个初始化插帧。例如，若目标视频特征序列包括16帧视频特征，通过插帧则可以再得到16个插帧，与原16帧即可组成32帧视频特征，从而使得视频边长。

[0097] 进一步，为了提高插帧的质量，避免其中出现的图像空洞或者目标变形等问题，还需要对这些初始化插帧进行微调。具体则可以基于预先训练的微调模型来进行微调，而微调模型则可以是利用扩散模型或对抗生成网络训练得到，此处不再赘述。

[0098] 需要说明的是，本公开利用光流直接在特征空间对视频特征进行插帧，可以大大节省在图像上做插帧的计算量。

[0099] 在一种实施方式中，为了进一步提高视频的质量，可以采用采用增强型视频解码器对插帧后的目标视频特征序列进行解码。具体而言，在训练时，视频解码器可以和图像编码器一起训练，其中，图像编码器可以使用现有的图像自编码器中的编码器，在视频解码器中则添加时序卷积模块和注意力模块，使得模型能够建模视频数据，解码出的视频也更加稳定。训练过程中，可以对输入的视频样本进行退化处理，例如包括模糊、加噪和压缩等，得到一个质量退化的视频样本，然后再将该质量退化的视频样本作为视频解码器的输入，但是视频解码器学习的目标视频却仍然是原始的高质量视频样本，从而使得视频解码器有增强画质的效果。

[0100] 例如，本公开实施例的视频解码器可以生成分辨率是512*512，时长5秒25fps的视频，此外，根据需求，还可以通过现有技术中的其他视频超分或插帧方法最后生成1080p、甚至是4K等更高分辨率的视频。对于经视频解码器生成目标视频的后续处理，根据需求进行配置即可，本公开对此不作任何限定。

[0101] 图6是根据本公开实施例的视频生成方法的预测流程图。如图所示，以文本生成视频并使用三级级联模型为例，其中，三级级联模型包括第一级视频特征生成子模型1、第二级视频特征生成子模型2和第三级视频特征生成子模型3。流程中，先得到一个输入文本，利用文本生成图像模型得到与文本语义相关的参考图像。图像编码器提取的原始参考图像的特征输入第三级视频特征生成子模型3。对原始参考图像进行一次下采样，得到第二级分辨率的参考图像，经图像编码器对其提取的特征输入第二级视频特征生成子模型2。对第二级分辨率的参考图像再次下采样，得到第一级分辨率的参考图像，经图像编码器对其提取的特征输入第一级视频特征生成子模型1。输入文本经文本编码器得到的文本特征则分别输入各个视频特征生成子模型。第一级视频特征生成子模型的输出经上采样输入第二级视频特征生成子模型，第二级视频特征生成子模型的输出经上采样输入第三级视频特征生成子模型，第三级视频特征生成子模型输出的即为目标视频特征序列。该目标视频特征序列经光流指导的特征插帧过程，先对目标视频特征序列的光流进行预测，然后基于预测的光流进行插帧，得到相对粗糙的多个初始化帧，最后通过微调模型进行微调，得到高质量的插帧，并与原视频特征一起组成最终的视频特征序列，经增强型视频解码器解码即可得到目标视频。图中，生成子模型1对应的16×4×16×16，第一个16表示16帧视频特征，代表特征序列的长度，4表示特征维度，即通道数，最后两个16分别表示视频图像的长和宽。同理，生成子模型2与生成子模型1不同的是，生成的视频图像的长和宽均为32，而生成子模型3则为64，分辨率逐级提高。图中光流指导的特征插帧N×4×64×64中的N，则表示经插帧后得到的帧数。

[0102] 本公开实施例的技术方案，利用了成熟的文生图技术先生成一个质量很好的参考图像，然后用这个参考图像指导视频的生成，这使得生成的内容可控且质量大大提升，降低了视频生成模型的难度，模型更多的聚焦在动态合成上。此外，整个方案在特征空间上，大

大减少了训练、测试需要的时间。另外，通过采用视频解码器单独训练的方式使得能利用4K分辨率的超高清视频训练解码器，生成的视频画质更高。

[0103] 图7是根据本公开实施例的视频生成装置的结构示意图，本实施例可适用于自动生成视频的情况，例如根据文本生成视频，根据图像生成视频，或者根据一个输入视频生成一个新视频等，涉及人工智能技术领域，具体为计算机视觉、深度学习等技术领域，可应用于AIGC等场景。该装置可实现本公开任意实施例所述的视频生成方法。如图7所示，该装置700具体包括：

[0104] 参考图像与文本获取模块701，用于获取参考图像和文本，其中，所述参考图像和所述文本之间语义相关；

[0105] 视频特征序列生成模块702，用于利用预先训练的视频特征生成模型，根据所述参考图像的特征和所述文本的特征生成目标视频特征序列；

[0106] 视频生成模块703，用于利用视频解码器对所述目标视频特征序列进行解码，生成目标视频。

[0107] 可选的，所述参考图像是利用预先训练的文本生成图像模型根据所述文本生成的。

[0108] 可选的，所述装置还包括：

[0109] 参考帧编辑模块，用于从原视频中提取参考帧，对所述参考帧进行图像编辑，得到所述参考图像。

[0110] 可选的，所述文本包括用户输入的文本。

[0111] 可选的，所述文本是利用预先训练的图像生成文本模型根据所述参考图像生成的。

[0112] 可选的，所述参考图像的特征是通过图像编码器提取，所述文本的特征是通过文本编码器提取。

[0113] 可选的，所述视频特征生成模型是基于扩散模型或者对抗生成网络训练得到。

[0114] 可选的，所述视频特征生成模型是由多个视频特征生成子模型组成的级联模型。

[0115] 可选的，所述级联模型包括N级视频特征生成子模型，上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入，其中，N为大于1的自然数。

[0116] 可选的，所述视频特征序列生成模块包括：

[0117] 下采样单元，用于通过对所述参考图像进行下采样，获取包括所述参考图像在内的N个不同分辨率级别的参考图像；

[0118] 特征提取单元，用于分别提取所述N个不同分辨率级别的参考图像的特征；

[0119] 第一级视频特征生成子模型处理单元，用于将第一级分辨率的参考图像的特征和所述文本的特征输入第一级视频特征生成子模型；

[0120] 当前级视频特征生成子模型处理单元，用于针对所述级联模型中除所述第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式处理：将所述当前级视频特征生成子模型的上一级视频特征生成子模型输出的视频特征序列进行上采样，并将经所述上采样得到的视频特征序列、当前级分辨率的参考图像的特征和所述文本的特征输入所述当前级视频特征生成子模型；

[0121] 输出单元，用于将第N级视频特征生成子模型的输出作为所述目标视频特征序列。

[0122] 可选的，所述装置还包括：

[0123] 插帧模块，用于对所述目标视频特征序列进行插帧；

[0124] 相应的，所述视频生成模块具体用于：

[0125] 利用视频解码器对所述插帧后的目标视频特征序列进行解码，生成目标视频。

[0126] 可选的，所述插帧模块包括：

[0127] 光流预测单元，用于利用预先训练的光流预测模型预测所述目标视频特征序列的光流；

[0128] 插帧单元，用于根据所述光流，对所述目标视频特征序列在特征空间进行插帧，得到多个初始化插帧；

[0129] 微调单元，用于利用预先训练的微调模型对所述多个初始化插帧进行微调。

[0130] 可选的，所述视频解码器中添加有时序卷积模块和注意力模块。

[0131] 可选的，在所述视频解码器的训练过程中，是对输入的视频样本进行了退化处理，其中，所述退化处理包括模糊、加噪和压缩。

[0132] 图8是根据本公开实施例的视频特征生成模型的训练装置的结构示意图，本实施例可适用于对视频特征生成模型进行训练，以便基于视频特征生成模型自动生成视频的情况，例如根据文本生成视频，根据图像生成视频，或者根据一个输入视频生成一个新视频等，涉及人工智能技术领域，具体为计算机视觉、深度学习等技术领域，可应用于AIGC等场景。该装置可实现本公开任意实施例所述的视频生成方法。如图8所示，该装置800具体包括：

[0133] 获取模块801，用于从视频训练样本中获取参考图像样本，并获取与所述参考图像样本语义相关的文本样本；

[0134] 特征提取模块802，用于分别提取所述参考图像样本的特征和所述文本样本的特征，并提取所述视频训练样本的视频特征序列样本；

[0135] 模型训练模块803，用于将所述参考图像样本的特征和所述文本样本的特征作为所述视频特征生成模型的输入，将所述视频特征序列样本作为所述视频特征生成模型的输出，对所述视频特征生成模型进行训练；

[0136] 其中，所述视频特征生成模型为由多个视频特征生成子模型组成的级联模型。

[0137] 可选的，所述级联模型中的各个视频特征生成子模型是分别进行训练的。

[0138] 可选的，所述级联模型包括N级视频特征生成子模型，上一级视频特征生成子模型的输出作为下一级视频特征生成子模型的输入；

[0139] 相应的，所述模型训练模块包括：

[0140] 第一级视频特征生成子模型训练单元，用于利用预先训练的文本生成图像模型初始化第一级视频特征生成子模型，并训练所述第一级视频特征生成子模型；

[0141] 当前级视频特征生成子模型训练单元，用于针对所述级联模型中除所述第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式进行训练：

[0142] 利用所述当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对所述当前级视频特征生成子模型初始化后再进行训练；

[0143] 其中，各级视频特征生成子模型的结构相同，N为大于1的自然数。

[0144] 可选的，所述第一级视频特征生成子模型训练单元包括：

[0145] 下采样子单元，用于通过对所述参考图像样本进行下采样，获取包括所述参考图像样本在内的N个不同分辨率级别的参考图像样本；

[0146] 第一级视频特征生成子模型训练子单元，用于将第一级分辨率的参考图像样本的特征和所述文本样本的特征作为所述第一级视频特征生成子模型的输入，提取所述视频训练样本的第一级分辨率的视频特征作为所述第一级视频特征生成子模型的输出，对所述第一级视频特征生成子模型进行训练。

[0147] 可选的，所述当前级视频特征生成子模型训练单元包括：

[0148] 初始化子单元，用于利用训练后得到的所述上一级视频特征生成子模型的模型参数初始化所述当前级视频特征生成子模型；

[0149] 上采样子单元，用于对所述视频训练样本的上一级分辨率的视频特征进行上采样；

[0150] 当前级视频特征生成子模型训练子单元，用于将当前级分辨率的参考图像的特征、所述文本特征和所述上采样得到的视频特征作为所述当前级视频特征生成子模型的输入，提取所述视频训练样本的当前级分辨率的视频特征作为所述当前级视频特征生成子模型的输出，对所述当前级视频特征生成子模型进行训练。

[0151] 可选的，所述获取模块包括：

[0152] 参考帧获取单元，用于从所述视频训练样本中提取参考帧；

[0153] 图像编辑单元，用于对所述参考帧进行图像编辑，得到所述参考图像样本。

[0154] 可选的，所述文本生成图像模型包括卷积模块和空间注意力模块，所述装置还包括：

[0155] 网络结构处理单元，用于在所述文本生成图像模型的网络结构中，在每个2D卷积模块后面添加时间维度的1D卷积模块，在每个空间注意力模块之后添加时间注意力模块。

[0156] 可选的，所述装置还包括视频条件信息处理模块，具体用于：

[0157] 获取所述视频训练样本的视频条件信息，其中，所述视频条件信息至少包括深度图和目标关键点图；

[0158] 将所述视频条件信息作为各级视频特征生成子模型的输入进行训练。

[0159] 上述产品可执行本公开任意实施例所提供的方法，具备执行方法相应的功能模块和有益效果。

[0160] 本公开的技术方案中，所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理，均符合相关法律法规的规定，且不违背公序良俗。

[0161] 根据本公开的实施例，本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0162] 图9示出了可以用来实施本公开的实施例的示例电子设备900的示意性框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0163] 如图9所示，设备900包括计算单元901，其可以根据存储在只读存储器(ROM)902中

的计算机程序或者从存储单元908加载到随机访问存储器(RAM)903中的计算机程序，来执行各种适当的动作和处理。在RAM903中，还可存储设备900操作所需的各种程序和数据。计算单元901、ROM 902以及RAM 903通过总线904彼此相连。输入/输出(I/O)接口905也连接至总线904。

[0164] 设备900中的多个部件连接至I/O接口905，包括：输入单元906，例如键盘、鼠标等；输出单元907，例如各种类型的显示器、扬声器等；存储单元908，例如磁盘、光盘等；以及通信单元909，例如网卡、调制解调器、无线通信收发机等。通信单元909允许设备900通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0165] 计算单元901可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元901的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元901执行上文所描述的各个方法和处理，例如视频生成方法。例如，在一些实施例中，视频生成方法可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元908。在一些实施例中，计算机程序的部分或者全部可以经由ROM 902和/或通信单元909而被载入和/或安装到设备900上。当计算机程序加载到RAM 903并由计算单元901执行时，可以执行上文描述的视频生成方法的一个或多个步骤。备选地，在其他实施例中，计算单元901可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行视频生成方法。

[0166] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、现场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、复杂可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0167] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0168] 在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0169] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0170] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)、区块链网络和互联网。

[0171] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与VPS服务中,存在的管理难度大,业务扩展性弱的缺陷。服务器也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0172] 人工智能是研究使计算机来模拟人的某些思维过程和智能行为(如学习、推理、思考、规划等)的学科,既有硬件层面的技术也有软件层面的技术。人工智能硬件技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理等技术;人工智能软件技术主要包括计算机视觉技术、语音识别技术、自然语言处理技术及机器学习/深度学习技术、大数据处理技术、知识图谱技术等几大方向。

[0173] 云计算(cloud computing),指的是通过网络接入弹性可扩展的共享物理或虚拟资源池,资源可以包括服务器、操作系统、网络、软件、应用和存储设备等,并可以按需、自服务的方式对资源进行部署和管理的技术体系。通过云计算技术,可以为人工智能、区块链等技术应用、模型训练提供高效强大的数据处理能力。

[0174] 此外,根据本公开的实施例,本公开还提供了另一种电子设备、另一种可读存储介质和另一种计算机程序产品,用于执行本公开任意实施例所述的视频特征生成模型的训练方法的一个或多个步骤。其具体的结构和程序代码可参见如图9所示的实施例的内容描述,此处不再赘述。

[0175] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本公开提供的技术方案所期望的结果,本文在此不进行限制。

[0176] 上述具体实施方式,并不构成对本公开保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等,均应包含在本公开保护范围之内。

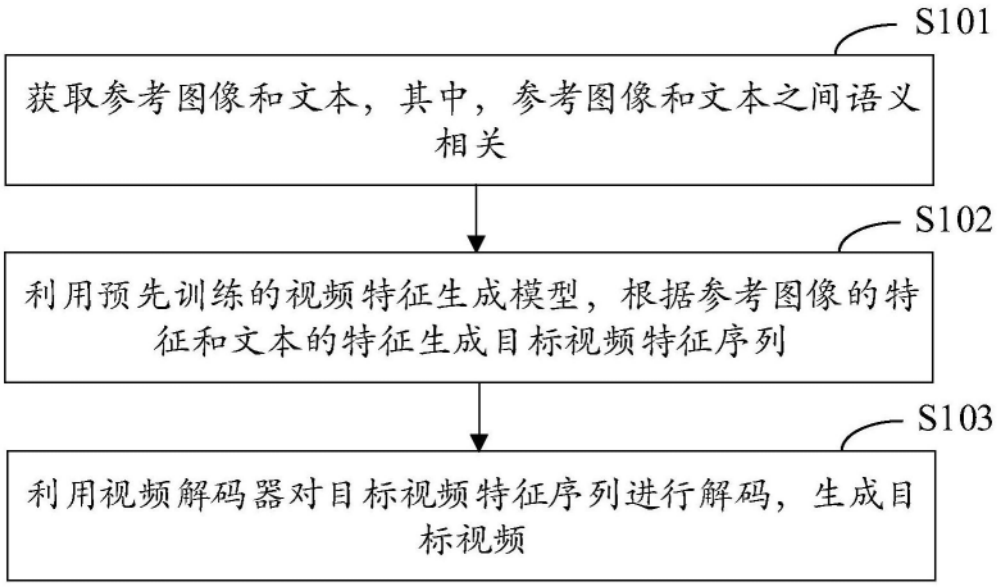

S101
获取参考图像和文本，其中，参考图像和文本之间语义相关
S102
利用预先训练的视频特征生成模型，根据参考图像的特征和文本的特征生成目标视频特征序列
S103
利用视频解码器对目标视频特征序列进行解码，生成目标视频

图1

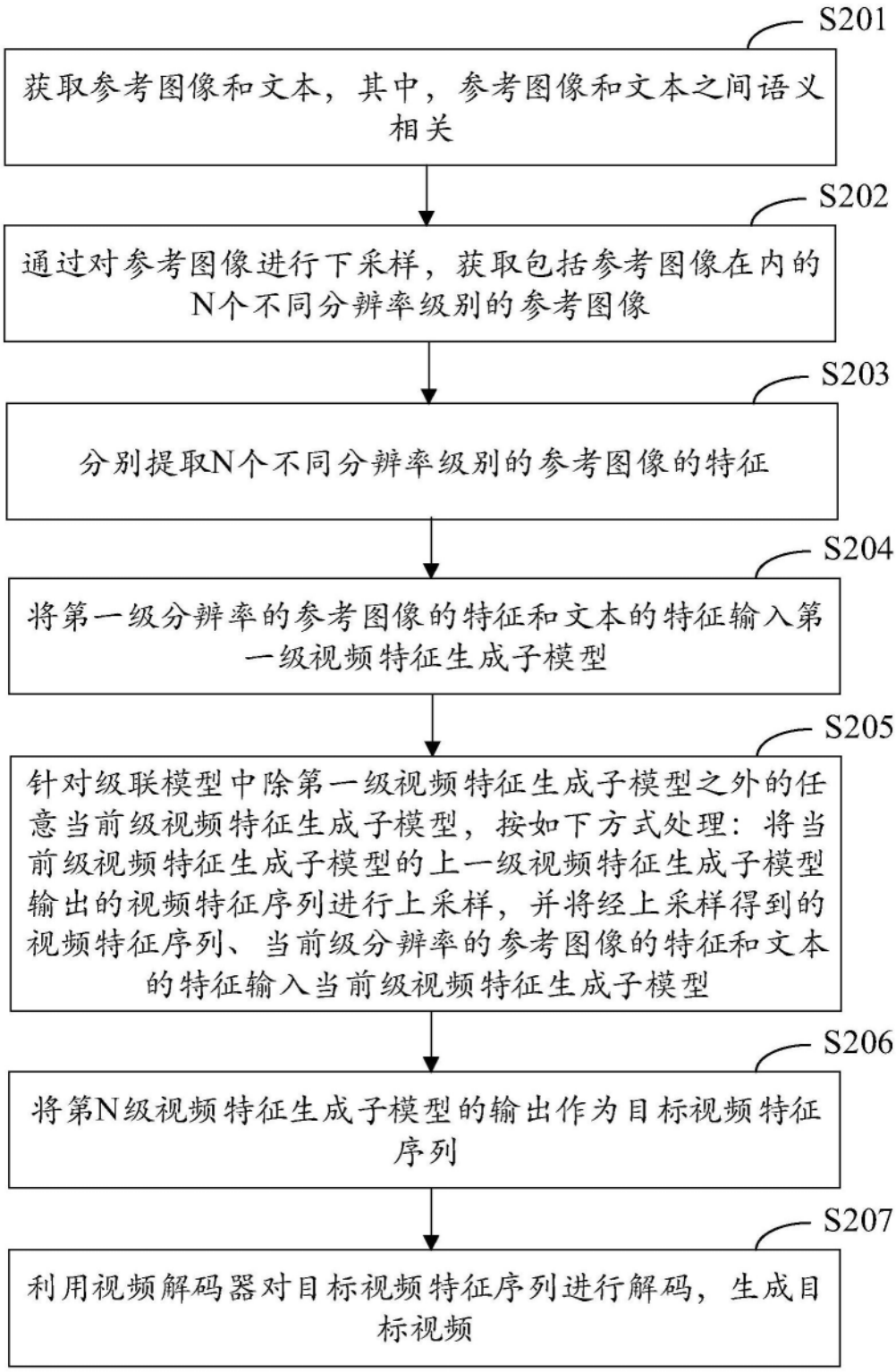

S201
获取参考图像和文本，其中，参考图像和文本之间语义相关
S202
通过对参考图像进行下采样，获取包括参考图像在内的N个不同分辨率级别的参考图像
S203
分别提取N个不同分辨率级别的参考图像的特征
S204
将第一级分辨率的参考图像的特征和文本的特征输入第一级视频特征生成子模型
S205
针对级联模型中除第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式处理：将当前级视频特征生成子模型的上一级视频特征生成子模型输出的视频特征序列进行上采样，并将经上采样得到的视频特征序列、当前级分辨率的参考图像的特征和文本的特征输入当前级视频特征生成子模型
S206
将第N级视频特征生成子模型的输出作为目标视频特征序列
S207
利用视频解码器对目标视频特征序列进行解码，生成目标视频

图2

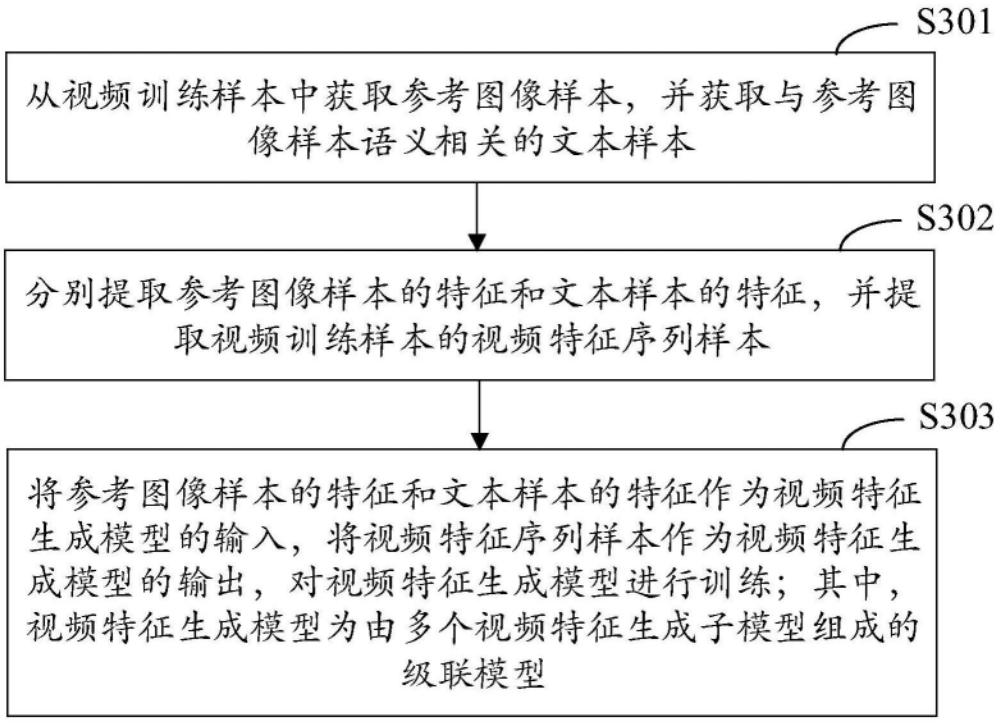

S301
从视频训练样本中获取参考图像样本，并获取与参考图像样本语义相关的文本样本
S302
分别提取参考图像样本的特征和文本样本的特征，并提取视频训练样本的视频特征序列样本
S303
将参考图像样本的特征和文本样本的特征作为视频特征生成模型的输入，将视频特征序列样本作为视频特征生成模型的输出，对视频特征生成模型进行训练；其中，视频特征生成模型为由多个视频特征生成子模型组成的级联模型

图3

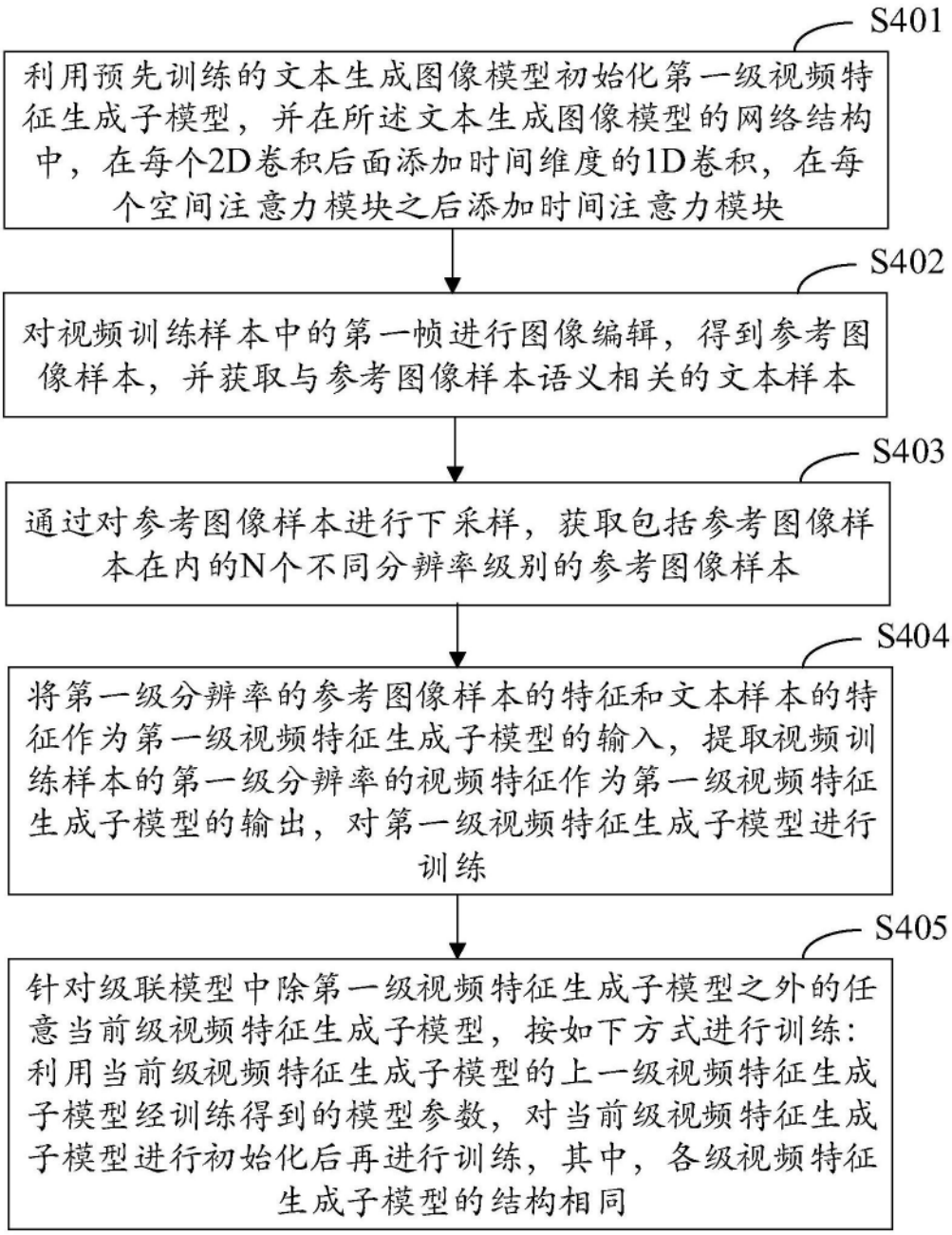

S401
利用预先训练的文本生成图像模型初始化第一级视频特征生成子模型，并在所述文本生成图像模型的网络结构中，在每个2D卷积后面添加时间维度的1D卷积，在每个空间注意力模块之后添加时间注意力模块
S402
对视频训练样本中的第一帧进行图像编辑，得到参考图像样本，并获取与参考图像样本语义相关的文本样本
S403
通过对参考图像样本进行下采样，获取包括参考图像样本在内的N个不同分辨率级别的参考图像样本
S404
将第一级分辨率的参考图像样本的特征和文本样本的特征作为第一级视频特征生成子模型的输入，提取视频训练样本的第一级分辨率的视频特征作为第一级视频特征生成子模型的输出，对第一级视频特征生成子模型进行训练
S405
针对级联模型中除第一级视频特征生成子模型之外的任意当前级视频特征生成子模型，按如下方式进行训练：利用当前级视频特征生成子模型的上一级视频特征生成子模型经训练得到的模型参数，对当前级视频特征生成子模型进行初始化后再进行训练，其中，各级视频特征生成子模型的结构相同

图4

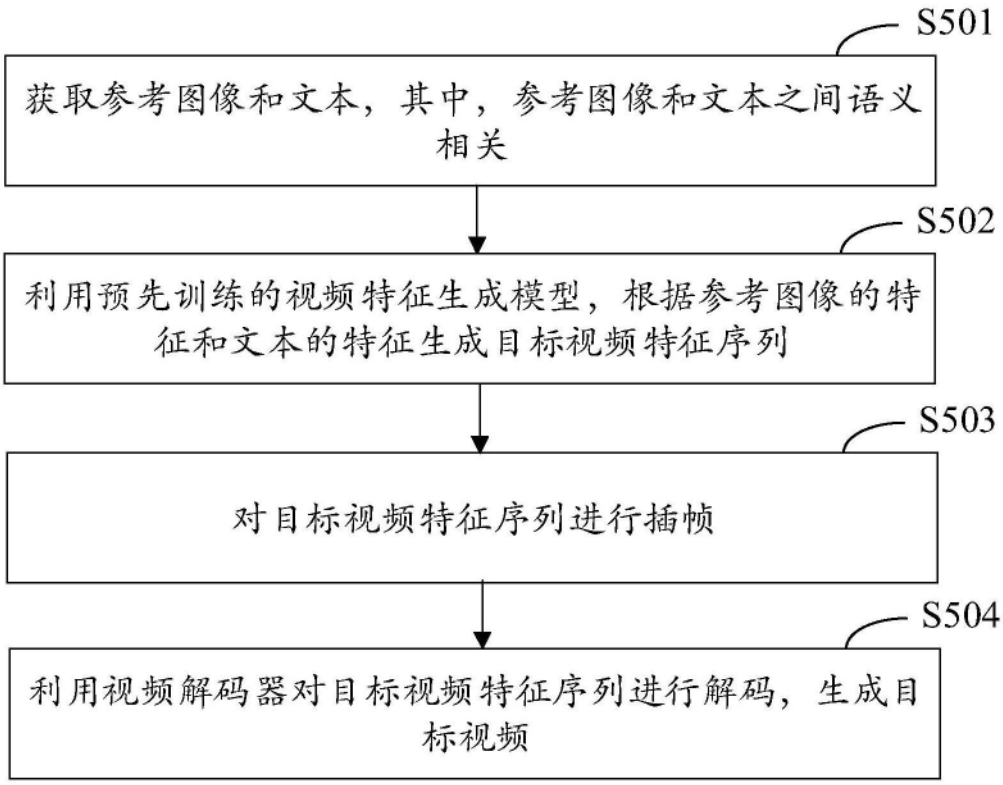

S501
获取参考图像和文本，其中，参考图像和文本之间语义相关
S502
利用预先训练的视频特征生成模型，根据参考图像的特征和文本的特征生成目标视频特征序列
S503
对目标视频特征序列进行插帧
S504
利用视频解码器对目标视频特征序列进行解码，生成目标视频

图5

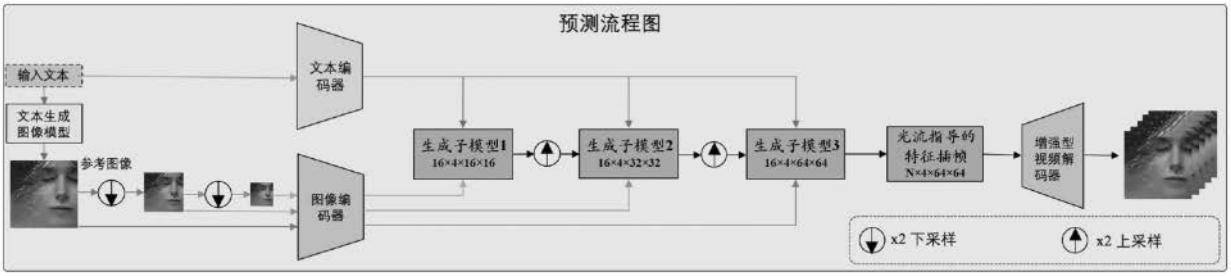

预测流程图
输入文本
文本生成图像模型
参考图像
文本编码器
图像编码器
生成子模型1
16×4×16×16
生成子模型2
16×4×32×32
生成子模型3
16×4×64×64
光流指导的特征插帧
N×4×64×64
增强型视频解码器
x2 下采样
x2 上采样

图6

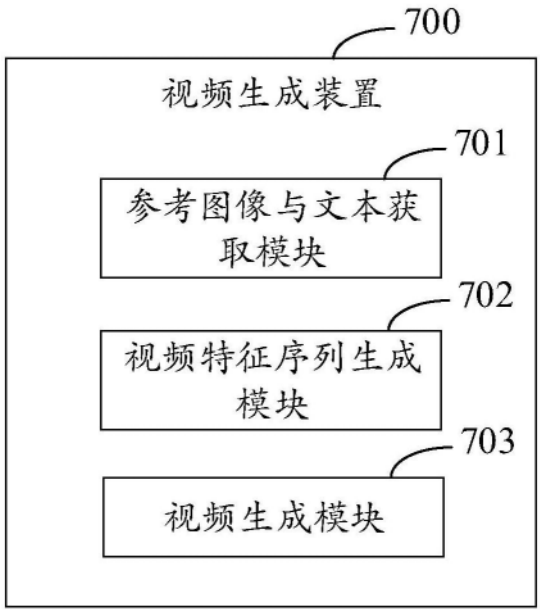

700
视频生成装置
701
参考图像与文本获取模块
702
视频特征序列生成模块
703
视频生成模块

图7

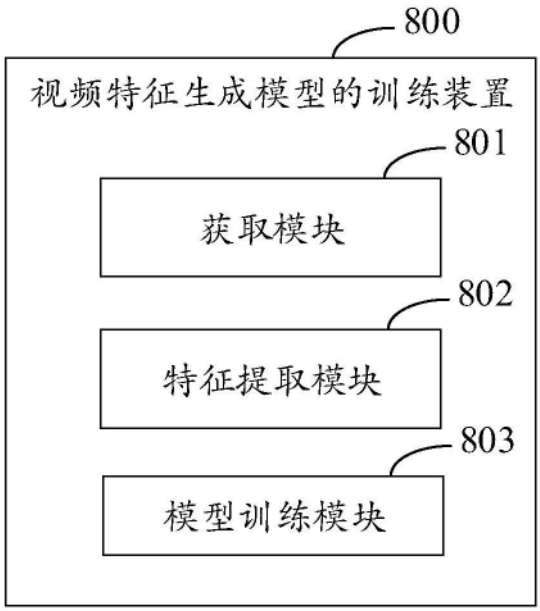

800
视频特征生成模型的训练装置
801
获取模块
802
特征提取模块
803
模型训练模块

图8

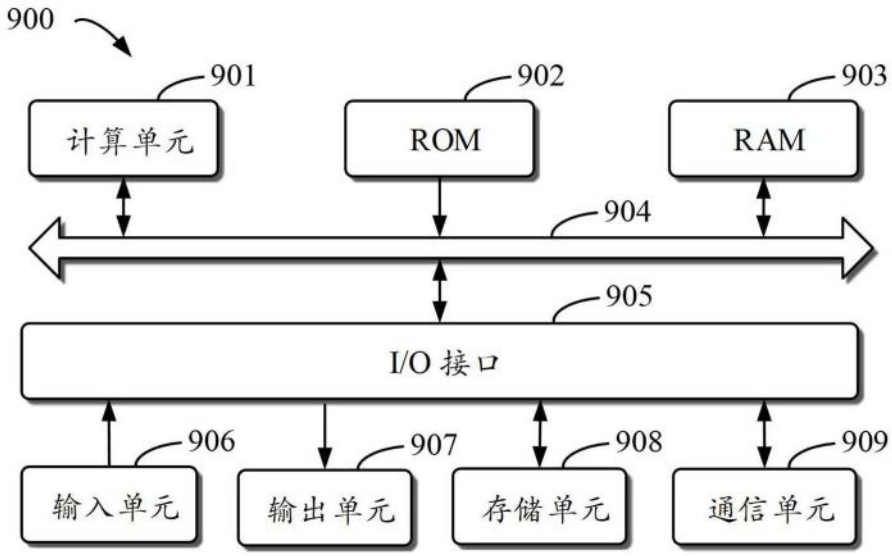

900
901
计算单元
902
ROM
903
RAM
904
905
I/O 接口
906
输入单元
907
输出单元
908
存储单元
909
通信单元

图9