

虚拟视频直播处理方法及装置、存储介质、电子设备

本申请要求于2020年11月26日提交中国专利局、申请号202011356995.7、申请名称为“虚拟视频直播处理方法及装置、存储介质、电子设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及虚拟视频直播技术领域，具体而言，涉及虚拟视频直播处理技术。

背景技术

10 随着通信技术的发展，网络通信带宽得到很大的提高，视频直播技术也日趋成熟，在各个方面得到应用。同时随着人工智能技术的发展，从文本到语音（Text To Speech, TTS）技术与图像合成技术也成为了人们的研究热点。视频直播技术与人工智能技术的结合（例如虚拟视频直播技术）可以在许多地方发挥作用，比如代替真人进行新闻虚拟播报、代替游戏主播进行游戏虚拟解说等，具有广阔的应用前景。

15 在虚拟视频直播技术中，音频与图片的生成需要耗费大量的计算时间，为保证虚拟视频直播的实时性，虚拟视频实时推流的实现成为影响最终直播视频质量的重要因素。现有的视频实时直播方法大都是针对已有稳定的音频、图片数据输入（如本地视频推流），或者是能够快速获取到音频、图片数据（如摄像头获取数据）等应用场景。

需要说明的是，在上述背景技术部分公开的信息仅用于加强对本申请的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

20 本申请的目的在于提供一种虚拟视频直播处理方法、虚拟视频直播处理装置、电子设备以及计算机可读存储介质，即使没有稳定的数据源也可以实时生成直播视频流，在一定程度上克服现有的虚拟视频直播技术采用硬切换方法进行视频中断处理而导致视频卡顿等问题，提高虚拟视频直播的流畅性。

25 根据本申请的第一方面，提供一种虚拟视频直播处理方法，所述方法由电子设备执行，所述方法包括：获取文本数据，并确定与所述文本数据对应的待合成视频数据；根据所述待合成视频数据实时合成直播视频流并将所述直播视频流推送至直播客户端；在接收到直播中断请求时，响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定目标视频数据；根据所述目标视频数据合成中断过渡视频流，并将所述中断过渡视频流推送至所述直播客户端。

30 根据本申请的第二方面，提供一种虚拟视频直播处理装置，所述装置部署在电子设备上，所述装置包括：数据确定模块，用于获取文本数据，并确定与所述文本数据对应的待合成视频数据；第一视频合成模块，用于根据所述待合成视频数据实时合成直播视频流并将所述直播视频流推送至直播客户端；中断响应模块，用于在接收到直播中断请求时，响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定目标视频数据；
35 第二视频合成模块，用于根据所述目标视频数据合成中断过渡视频流，并将所述中断过渡视频流推送至所述直播客户端。

-2-

在本申请的一种示例性实施例中，所述数据确定模块包括：目标文本确定单元，用于确定与所述文本数据相关的关联文本数据，将所述文本数据和所述关联文本数据作为目标文本；音频特征生成单元，用于将所述目标文本转化为对应的目标词向量，对所述目标词向量进行编码处理和解码处理，以生成所述目标词向量的音频特征；音频数据生成单元，用于对所述音频特征进行合成处理，以生成音频数据；图像生成单元，用于确定与所述目标文本对应的表情特征数据，并根据所述表情特征数据生成视频输出图像。

在本申请的一种示例性实施例中，图像生成单元包括：脸部图像生成单元，用于对所述表情特征数据进行渲染处理以生成虚拟对象对应的脸部图像；背景图像确定单元，用于确定与所述脸部图像匹配的背景图像；所述背景图像包括所述虚拟对象的肢体动作信息和环境背景信息；图像融合单元，用于对所述脸部图像与所述背景图像进行图像融合处理以生成所述视频输出图像。

在本申请的一种示例性实施例中，背景图像确定单元包括：音频时长确定子单元，用于确定与所述脸部图像时序对齐的所述音频数据的音频时长；视频时长确定子单元，用于根据所述音频时长确定输出视频的视频时长；背景图像确定子单元，用于根据所述视频时长从预先存储的背景图像集中确定与所述脸部图像匹配的背景图像。

在本申请的一种示例性实施例中，背景图像确定子单元被配置为：确定所述视频时长下所述输出视频中包含的所述虚拟对象的对象动作数量；根据所述对象动作数量从所述背景图像集中确定同样数量组别的背景图像，作为与所述脸部图像匹配的背景图像。

在本申请的一种示例性实施例中，中断响应模块包括：目标图像确定单元，用于响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定剩余视频输出数据，并根据所述剩余视频输出数据确定目标视频图像；目标音频确定单元，用于响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定剩余音频数据，并根据所述剩余音频数据和所述目标视频图像确定目标音频数据；视频数据确定单元，用于将所述目标视频图像和所述目标音频数据作为所述目标视频数据。

在本申请的一种示例性实施例中，目标图像确定单元包括：时长判断子单元，用于确定所述剩余视频输出图像对应的剩余图像帧时长；目标图像确定子单元，用于如果所述剩余图像帧时长大于时长阈值，则确定跳帧步长，根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像。

在本申请的一种示例性实施例中，目标图像确定子单元包括：跳帧步长确定子单元，用于获取所述剩余视频输出图像对应的图像帧率，并确定所述剩余视频输出图像的剩余图像数量；根据所述剩余图像数量和所述图像帧率确定所述跳帧步长；图像确定子单元，用于按照时间顺序并根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像。

在本申请的一种示例性实施例中，图像确定子单元被配置为：获取所述剩余视频输出图像中的首帧图像和末帧图像；按照所述时间顺序从所述首帧图像开始，每隔所述跳帧步长从所述剩余视频输出图像中确定中间图像；将所述首帧图像、所述中间图像和所述末帧图像作为所述目标视频图像。

在本申请的一种示例性实施例中，目标图像确定单元还可以被配置为：根据所述跳帧步长确定所述目标视频图像的目标图像数量；确定所述剩余视频输出图像中两两相邻图像

-3-

之间的图像相似度,根据所述图像相似度从所述剩余视频输出图像中确定不超过所述目标图像数量的所述目标视频图像。

- 5 在本申请的一种示例性实施例中,目标音频确定单元被配置为:确定所述剩余音频数据的剩余音频时长;如果所述剩余音频时长大于所述时长阈值,则删除所述剩余音频数据;确定所述目标视频图像对应的目标图像帧时长,根据所述目标图像帧时长生成所述目标音频数据。

- 10 在本申请的一种示例性实施例中,所述中断响应模块还包括动画生成子单元,动画生成子单元被配置为:响应于所述直播中断请求将尚未合成直播视频流的所述待合成视频数据中的第一个视频输出图像作为第一中断图像;响应于所述直播中断请求将尚未合成直播视频流的所述待合成视频数据中的最后一个视频输出图像作为第二中断图像;根据所述第一中断图像与所述第二中断图像生成预设时长的中断过渡动画。

根据本申请的第三方面,提供一种用于虚拟视频直播处理的电子设备,包括:处理器;以及存储器,用于存储所述处理器的可执行指令;其中,所述处理器配置为经由执行所述可执行指令来执行上述任意一项所述的方法。

- 15 根据本申请的第四方面,提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现上述任意一项所述的方法。

根据本申请的第五方面,提供一种计算机程序产品,当所述计算机程序产品被执行时,用于实现上述任意一项所述的方法。

本申请示例性实施例可以具有以下部分或全部有益效果:

- 20 在本申请的一示例实施方式所提供的虚拟视频直播处理中,根据与文本数据对应的待合成视频数据实时合成直播视频流并将直播视频流推送至直播客户端的过程中,如果接收到直播中断请求,则从尚未合成直播视频流的待合成视频数据中确定目标视频数据,以根据目标视频数据合成中断过渡视频流。一方面,通过文本数据生成对应的待合成视频数据,在视频直播过程中,即使没有稳定的数据源也可以实时生成直播视频流。另一方面,避免采用硬切换进行视频中断处理时存在视频卡顿的问题,可以有效进行直播视频被中断时的平滑过渡处理,提高虚拟视频直播的流畅性。
- 25

应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本申请。

附图说明

- 30 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本申请的实施例,并与说明书一起用于解释本申请的原理。显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,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。

- 35 图1示出了可以应用本申请实施例的一种图像融合方法及装置的示例性系统架构的示意图;

图2示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图;

图3示意性示出了根据本申请的一个实施例的虚拟视频直播处理方法的流程图;

图 4 示意性示出了根据本申请的一个实施例的虚拟视频直播服务的整体框架图；

图 5 示意性示出了根据本申请的一个实施例的从剩余视频输出图像中确定目标视频图像的流程圖；

5 图 6 示意性示出了根据本申请的另一实施例的从剩余视频输出图像中确定目标视频图像的流程圖；

图 7 示意性示出了根据本申请的一个实施例的根据目标图像帧时长生成目标音频数据的流程圖；

图 8 示意性示出了根据本申请的一个实施例的生成中断过渡动画的流程圖；

图 9 示意性示出了根据本申请的一个实施例的虚拟视频推流服务的实现框架圖；

10 图 10 示意性示出了根据本申请的一个实施例的虚拟视频直播处理装置的结构圖。

具体实施方式

15 现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本申请将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中，提供许多具体细节从而给出对本申请的实施方式的充分理解。然而，本领域技术人员将意识到，可以实践本申请的技术方案而省略所述特定细节中的一个或更多，或者可以采用其它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知技术方案以避免喧宾夺主而使得本申请的各方面变得模糊。

20 此外，附图仅为本申请的示意性圖解，并非一定是按比例绘制。圖中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附圖中所示的一些方框圖是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理装置和/或微控制器装置中实现这些功能实体。

25 人工智能（Artificial Intelligence，AI）是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

30 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础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、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

本申请实施例提供的方案涉及人工智能的虚拟视频直播处理等技术，具体通过如下实施例进行说明：

图 1 示出了可以应用本申请实施例的一种虚拟视频直播处理方法及装置的示例性应用环境的系统架构的示意图。

5 如图 1 所示，系统架构 100 可以包括终端设备 101、102、103 中的一个或多个，网络 104 和服务 105。网络 104 用以在终端设备 101、102、103 和服务 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。终端设备 101、102、103 可以是具有显示屏的各种电子设备，包括但不限于台式计算机、便携式计算机、智能手机和平板电脑等等。应该理解，图 1 中的终端设备、网络和服务 10 的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务。比如服务器 105 可以是多个服务器组成的服务器集群等。

本申请实施例所提供的虚拟视频直播处理方法一般由服务器 105 执行，相应地，虚拟视频直播处理装置一般设置于服务器 105 中。但本领域技术人员容易理解的是，本申请实施例所提供的虚拟视频直播处理方法也可以由终端设备 101、102、103 执行，为其他终端 15 设备提供虚拟直播处理服务，相应的，虚拟视频直播处理装置也可以设置于终端设备 101、102、103 中，本示例性实施例中对此不做特殊限定。举例而言，在一种示例性实施例中，可以是用户通过终端设备 101、102、103 将文本数据上传至服务器 105，服务器通过本申请实施例所提供的虚拟视频直播处理方法根据文本数据对应的待合成视频数据生成直播视频流和中断过渡视频流，并将直播视频流和中断过渡视频流传输给终端设备 101、102、103 20 等。

图 2 示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

需要说明的是，图 2 示出的电子设备的计算机系统 200 仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

25 如图 2 所示，计算机系统 200 包括中央处理单元（CPU）201，其可以根据存储在只读存储器（ROM）202 中的程序或者从存储部分 208 加载到随机访问存储器（RAM）203 中的程序而执行各种适当的动作和处理。在 RAM 203 中，还存储有系统操作所需的各种程序和数据。CPU 201、ROM 202 以及 RAM 203 通过总线 204 彼此相连。输入/输出（I/O）接口 205 也连接至总线 204。

30 以下部件连接至 I/O 接口 205：包括键盘、鼠标等的输入部分 206；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部分 207；包括硬盘等的存储部分 208；以及包括诸如 LAN 卡、调制解调器等的网络接口卡的通信部分 209。通信部分 209 经由诸如因特网的网络执行通信处理。驱动器 210 也根据需要连接至 I/O 接口 205。可拆卸介质 211，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器 210 上，以便于从其上读出的计算机程序根据需要被安装入存储部分 208。

特别地，根据本申请的实施例，下文参考流程图描述的过程可以被实现为计算机软件程序。例如，本申请的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分 209 从网络上被下载和安装，和/或从可拆卸介质 211 被安装。在该计算机程序被中央处理单元 (CPU) 201 执行时，执行本申请的方法和装置中限定的各种功能。在一些实施例中，计算机系统 200 还可以包括 AI (Artificial Intelligence, 人工智能) 处理器，该 AI 处理器用于处理有关机器学习的计算操作。

需要说明的是，本申请所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器 (RAM)、只读存储器 (ROM)、可擦式可编程只读存储器 (EPROM 或闪存)、光纤、便携式紧凑磁盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF 等等，或者上述的任意合适的组合。

附图中的流程图和框图，图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图或流程图中的每个方框、以及框图或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现，所描述的单元也可以设置在处理器中。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定。

作为另一方面，本申请还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该电子设备中。上

述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被一个该电子设备执行时，使得该电子设备实现如下述实施例中所述的方法。例如，所述的电子设备可以实现如图 3~图 9 所示的各个步骤等。

以下对本申请实施例的技术方案进行详细阐述：

- 5 在虚拟视频直播技术中，主要是根据音频与图像生成对应的直播视频流，并将直播视频流实时推送至直播客户端，其中直播客户端可以安装在如图 1 所示的终端设备 101、102、103 上。在音频与图像的生成过程中，需要耗费大量的计算时间，为保证虚拟视频直播的实时性，虚拟视频实时推流的实现成为影响最终直播视频质量的重要因素。现有的视频实时直播方法大都是针对已有稳定的音频、图片数据输入（如本地视频推流），或者是能够快速获取到音频、图片数据（如摄像头获取数据）等应用场景。
- 10

然而，这些稳定的数据源无法在虚拟视频直播中得到很好应用；同时，在不影响直播实时性的前提下，现有的虚拟直播视频技术大都不支持视频中断平滑过渡处理。在虚拟视频直播过程中，如果接收到直播中断请求，通常是通过硬切换的方式处理直播中断请求；或者要么不进行中断处理，直到当前视频播放完成。

- 15 在虚拟视频直播过程中，如果采用硬切换的方式处理中断请求时，将导致直播视频出现卡顿现象，无法实现当前视频动作与下一个视频动作之间的平滑过渡；另外，在视频直播过程中，如果不进行中断处理，那么用户将无法感受到直播视频的实时响应。

- 基于上述一个或多个问题，本示例实施方式提供了一种虚拟视频直播处理方法。该虚拟视频直播处理方法可以应用于上述服务器 105，也可以应用于上述终端设备 101、102、103 中的一个或多个，本示例性实施例中对此不做特殊限定。参考图 3 所示，该虚拟视频直播处理方法可以包括以下步骤 S310 至步骤 S340：
- 20

 步骤 S310. 获取文本数据，并确定与文本数据对应的待合成视频数据。

 步骤 S320. 根据待合成视频数据实时合成直播视频流并将直播视频流推送至直播客户端。

- 25 步骤 S330. 在接收到直播中断请求时，响应于直播中断请求从尚未合成直播视频流的待合成视频数据中确定目标视频数据。

 步骤 S340. 根据目标视频数据合成中断过渡视频流，并将中断过渡视频流推送至直播客户端。

- 在本示例实施方式所提供的虚拟视频直播处理方法中，根据与文本数据对应的待合成视频数据实时合成直播视频流并将直播视频流推送至直播客户端的过程中，如果接收到直播中断请求，则从尚未合成直播视频流的待合成视频数据中确定目标视频数据，以根据目标视频数据合成中断过渡视频流。一方面，通过文本数据生成对应的待合成视频数据，在视频直播过程中，即使没有稳定的数据源也可以实时生成直播视频流。另一方面，避免采
- 30

用硬切换进行视频中断处理时存在视频卡顿的问题，可以有效进行直播视频被中断时的平滑过渡处理，提高虚拟视频直播的流畅性。

下面，对于本示例实施方式的上述步骤进行更加详细的说明。

在步骤 S310 中，获取文本数据，并确定与文本数据对应的待合成视频数据。

- 5 本示例实施方式中，文本数据可以是生成待合成视频数据所依据的文本数据，文本数据可以是虚拟对象在直播视频中将说出的文字。例如，用户可以通过直播客户端自行输入文本数据，用户还可以从直播客户端提供的众多可选文本中选择出相应的文本数据；文本数据还可以是预先存储于服务器中的文本数据。待合成视频数据可以在虚拟视频直播过程中，生成直播视频流所依赖的相关视频数据。例如，待合成视频数据可以包括图像数据、
10 音频数据和动画数据等；其中，图像数据可以是包括虚拟对象的脸部特征、肢体动作以及
与虚拟对象对应的环境背景的图像。音频数据可以是模拟的人声，用于匹配虚拟对象的视频输出图像。动画数据可以是根据一些特定事件生成的动画。

- 在进行虚拟视频直播时，虚拟视频直播服务端（例如服务器）可以先获取文本数据，在获取到文本数据后，根据文本数据确定对应的待合成视频数据。参考图 4，图 4 示意性
15 示出了根据本申请的一个实施例的虚拟视频直播服务的整体框架图。虚拟视频直播服务可以包括直播客户端 410 和虚拟视频直播服务端 420，虚拟视频直播服务端 420 中可以包括
文本转语音模块 421、三维渲染模块 422 和视频推流模块 423。在获取文本数据时，虚拟视频直播服务端 420 可以通过直播客户端 410 中获取用户输入或选择的文本数据；虚拟视频直播服务端 420 还可以获取预先存储在服务端数据库中的文本数据。虚拟视频直播服务端
20 420 在获取到文本数据后，可以根据文本数据生成对应的待合成视频数据。

例如，可以通过下述步骤确定与文本数据对应的待合成视频数据。

- 本申请的一个示例实施例中，确定与文本数据相关的关联文本数据，将文本数据和关联文本数据作为目标文本；将目标文本转化为对应的目标词向量，对目标词向量进行编码处理和解码处理，以生成目标词向量的音频特征；对音频特征进行合成处理，以生成音频
25 数据；确定与目标文本对应的表情特征数据，并根据表情特征数据生成视频输出图像。

- 其中，关联文本数据可以是与文本数据相关的文本数据。例如，在一些人机交互场景中，当文本数据为“产品 A 如何使用”，关联文本数据可以是产品 A 具体使用说明对应的文本。目标文本可以包括文本数据以及与之相关的关联文本数据，例如，将文本数据和关联文本数据进行拼接可以生成目标文本。目标词向量可以是对目标文本进行向量化处理后
30 生成的向量。音频特征可以是目标词向量转换为音频数据所对应的关键特征，例如，音频特征可以包括短时过零率、短时能量、频谱质心等。音频数据可以是文本数据经过文本语音转换处理后所生成的对应音频。表情特征数据可以是文本数据对应于直播视频中的虚拟对象的脸部表情特征。视频输出图像可以是直播视频所对应的视频图像。

在获取到文本数据后，可以确定与该文本数据相关的关联文本数据，并将文本数据与关联文本数据作为目标文本。参考图 4，虚拟视频直播服务端 420 可以通过文本转语音模块 421 将目标文本进行文本转语音处理，文本转语音模块中可以包含一个训练后的文本转语音模型，即预先构建一个初始的深度学习文本转语音模型，采用深度学习方法对该初始的文本转语音模型进行训练，训练完成后，得到训练后的文本转语音模型。将目标文本输入至训练好的文本转语音模型后，可以先把目标文本转化为对应的目标词向量，例如，目标文本可以是一句话，对目标文本进行分词，可以确定各个词对应的词向量即目标词向量。在得到目标词向量后，可以对目标词向量进行编码处理得到中间语义标识，再通过解码处理得到目标词向量的音频特征，采用音频合成算法，对得到的音频特征进行合成处理，可以生成目标文本对应的音频数据。

另外，文本转语音模块 421 还可以根据目标文本生成对应的表情特征数据，生成对应的表情特征数据可以根据下述技术方案实现，根据目标文本的音频数据预测直播视频中虚拟对象的嘴部关键点，对嘴部关键点进行归一化处理，以使嘴部关键点适应于标准面部模板；对经过归一化处理的嘴部关键点进行降维处理，得到对应虚拟对象的口型特征；对目标文本进行语义分析，得到目标文本所表征的语义；根据目标文本所表征的语义，确定与语义匹配的表情特征数据，以便后续根据表情特征数据可以生成对应的视频输出图像。

在本申请的一种示例性实施例中，对表情特征数据进行渲染处理以生成虚拟对象对应的脸部图像；确定与脸部图像匹配的背景图像；背景图像包括虚拟对象的肢体动作信息和环境背景信息；对脸部图像与背景图像进行图像融合处理以生成视频输出图像。

其中，脸部图像可以是直播视频中虚拟对象对应的脸部图像。背景图像可以是与虚拟对象脸部图像对应且包含虚拟对象其他特征的图像，例如，背景图像可以包括虚拟对象的肢体动作信息以及生成视频输出图像对应的环境背景信息。视频输出图像可以是包含虚拟对象的脸部图像、肢体动作以及当前所对应的环境背景的图像。

参考图 4，当文本转语音模块 421 将文本数据进行文本转化处理并生成待合成视频数据后，可以将生成的待合成视频数据推送至三维渲染模块 422，三维渲染模块 422 可以对待合成视频数据中的表情特征数据进行渲染处理，即结合虚拟对象的口型特征以及表情特征数据形成虚拟对象的脸部特征，根据脸部特征生成虚拟对象对应的脸部图像。在三维渲染模块中，由于三维渲染与融合是十分耗时的，所以当输入所有的表情特征数据时，无法马上就能得到所有融合后的图像，而是流式得到一组组融合后的图像。

在对表情特征数据进行渲染处理生成脸部图像时，可以采用开放图形库(Open Graphics Library, OpenGL)工具实现。例如，可以从 OpenGL 的几何图元中设置数据，用于构建虚拟对象的脸部轮廓形状；对输入的口型特征、表情特征数据等图元数据进行计算，判断脸部关键点的位置、颜色等其他渲染属性；将输入的图元的数学描述转化为与屏幕位置对应的像素片元，即进行光栅化处理；对光栅化产生的每个片元，执行片元着色器，从而确定

每隔片元的颜色及位置;另外,可以通过一些额外操作判断每个片元对应的对象是否可见,或者将该片元的颜色与当前屏幕的颜色进行融合。

在生成脸部图像之后,可以根据脸部图像确定出与之对应的背景图像。将虚拟对象的脸部图像与各自对应的背景图像进行图像融合处理后,可以生成直播视频中虚拟对象对应的视频输出图像,每个视频输出图像可以对应虚拟对象在一个时刻的肢体动作以及在当前时刻下对应的环境背景。脸部图像与背景图像融合为一张完整的视频输出图像可以由卷积神经网络(Convolutional Neural Networks, CNN)实现。

在本申请的一种示例性实施例中,确定与脸部图像时序对齐的音频数据的音频时长;根据音频时长确定输出视频的视频时长;根据视频时长从预先存储的背景图像集中确定与脸部图像匹配的背景图像。

其中,音频时长可以是根据目标文本生成的音频数据的播放时长。输出视频可以是待合成视频数据进行视频合成处理后并推流至直播客户端的直播视频。视频时长可以是输出视频的播放时长。背景图像集可以是预先设置的用于存储各种不同类别的背景图像的图像集。

由于生成的音频数据与根据表情数据生成的视频输出图像是时序对齐的,因此,在确定与脸部图像对应的背景图像时,可以先确定与脸部图像时序对齐的音频数据,并确定该音频数据对应的音频时长。将音频数据和视频输出图像进行合成处理后可以生成输出视频,因此,可以根据音频时长计算出输出视频的视频时长,输出视频的视频时长与音频数据的音频时长相等。在确定出视频时长后,可以获取预先存储的背景图像集,根据视频时长从背景图像集中选取背景图像。

在本申请的一种示例性实施例中,确定视频时长下输出视频中包含的虚拟对象的对象动作数量;根据对象动作数量从背景图像集中确定同样数量组别的背景图像,作为与脸部图像匹配的背景图像。

其中,对象动作数量可以是根据待合成视频数据确定出的直播视频中虚拟对象所进行的肢体动作的数量。根据脸部图像确定背景图像的具体过程可以包括:根据输出视频的视频时长可以确定该输出视频中包含的虚拟对象的对象动作数量,例如,输出视频中包含 n 个虚拟对象的动作;可以从背景图像集中预先保存的通用背景图像中选择 n 组合适的背景图像与脸部图像进行匹配。由于每组背景图像均可以对应一个完整的动作, n 组背景图像可以对应虚拟对象所要完成的 n 个动作, n 组背景图像正好可以在虚拟视频结束时完成 n 个动作。每一组动作对应的背景图像可以是同一张图像,以保证虚拟视频的自然衔接。

在步骤 S320 中,根据待合成视频数据实时合成直播视频流并将直播视频流推送至直播客户端。

本示例实施方式中,直播视频流可以是待合成视频数据进行视频合成处理后生成的视频流,直播视频流可以通过视频推流模块实时推送至直播客户端。虚拟直播视频服务端

在生成直播视频流后，可以将直播视频流推送至直播客户端进行实时播放。直播客户端可以用于接收直播视频流，并根据直播视频流在直播客户端实时呈现虚拟直播视频。直播中断请求可以是在虚拟直播过程中虚拟视频直播服务端接收到的中断请求。直播中断请求可以用于中断当前实时直播的虚拟直播视频。

- 5 参考图 4，虚拟视频直播服务端 420 中的视频推流模块 423，在接收到待合成视频数据后，可以对待合成视频数据进行视频合成处理，实时合成直播视频流，并将直播视频流推送至直播客户端 410 中，以便直播客户端 410 根据直播视频流实时呈现虚拟直播视频。直播客户端 410 获取直播视频流，并实时呈现与直播视频流对应的虚拟直播视频，可以通过以下技术方案实现，虚拟视频直播服务端 420 将直播视频流推送至相应流媒体播放地址的流媒体接口，以使直播客户端 410 基于流媒体播放地址拉取直播视频流，并通过直播客户端 410 的用户界面实时呈现实时合成的虚拟直播视频。

- 15 在虚拟视频直播服务端向直播客户端推送直播视频流的同时，用户可以通过直播客户端向虚拟视频直播服务端发送直播中断请求，当虚拟视频直播服务端接收到直播中断请求，可以进行对直播中断请求进行处理的过程。用户可以与虚拟视频直播中的虚拟对象进行互动，当虚拟对象正在进行直播时，用户可以通过直播客户端录入语音或输入文字，当直播客户端接收到用户反馈时，将生成直播中断请求。例如，用户询问虚拟对象“产品 A 如何使用”，当虚拟对象通过实时直播介绍“产品 A 的使用说明”时，用户通过直播客户端回答“我知道了”，此时，将直播客户端将生成一个直播中断请求，并将该直播中断请求发送至虚拟视频直播服务端。

- 20 在步骤 S330 中，在接收到直播中断请求时，响应于所述直播中断请求从尚未合成直播视频流的待合成视频数据中确定目标视频数据。

- 本示例实施方式中，尚未合成直播视频流的待合成视频数据可以是虚拟视频直播服务端在接收到直播中断请求后，还未进行视频合成处理的待合成视频数据。尚未合成直播视频流的待合成视频数据可以是经过三维渲染模块的渲染处理但还未经视频推流模块的视频合成处理的视频数据。目标视频数据可以从尚未合成直播视频流的待合成视频数据确定出的视频数据，目标视频数据可以用于生成直播中断请求后对应的中断过渡视频流的视频数据，例如，目标视频数据可以包括目标视频图像和目标音频数据等。

- 30 虚拟视频直播服务端在接收到直播中断请求后，三维渲染模块可以从尚未合成直播视频流的待合成视频数据中确定出目标视频数据，以根据目标视频数据生成对应的直播视频流。

- 在本申请的一种示例性实施例，响应于直播中断请求从尚未合成直播视频流的待合成视频数据中确定剩余视频输出数据，并根据剩余视频输出数据确定目标视频图像；响应于直播中断请求从尚未合成直播视频流的待合成视频数据中确定剩余音频数据，并根据剩余音频数据和目标视频图像确定目标音频数据；将目标视频图像和目标音频数据作为目标视频数据。

其中，剩余视频输出数据可以是尚未合成直播视频流的待合成视频数据中包含的视频输出数据，剩余视频输出数据可以包括剩余视频输出图像。剩余音频数据可以是尚未合成直播视频流的待合成视频数据中包含的音频数据。剩余视频输出数据和剩余音频数据均是还未进行视频合成处理的待合成视频数据。

5 虚拟视频直播服务端在接收到直播中断请求后，可以从尚未合成直播视频流的待合成视频分别确定出剩余视频输出数据和剩余音频数据，根据剩余视频输出数据确定出目标视频图像，根据剩余音频数据和目标视频图像确定目标音频数据，并将目标视频图像和目标音频数据作为目标视频数据。

10 在本申请的一种示例性实施例中，确定剩余视频输出图像对应的剩余图像帧时长，判断剩余图像帧时长是否大于时长阈值；如果剩余图像帧时长大于时长阈值，则确定跳帧步长，根据跳帧步长从剩余视频输出图像中确定目标视频图像。

其中，剩余视频输出图像可以是还未合成直播视频流的视频输出图像，可以将剩余视频输出图像的数量记为 F 。剩余图像帧时长可以是播放这些剩余视频输出图像所对应的时间长度，可以记为 T 。时长阈值可以是预先设定的用于与剩余图像帧时长进行对比的数值。
15 跳帧步长可以是从小于剩余视频输出图像中选择目标视频图像中所采用的视频帧的步长，可以记为 S 。目标视频图像可以是用于生成直播中断请求后的虚拟直播视频所采用的视频图像。

当接收到直播中断请求后，虚拟视频直播服务端中的三维渲染模块可以对剩余视频输出图像进行跳帧压缩处理，以确定出目标视频数据。参考图 5，图 5 示意性示出了根据本申请的一个实施例的从剩余视频输出图像中确定目标视频图像的流程图。图 5 示出了对剩余视频输出图像进行跳帧压缩处理的流程。具体可以包括下述步骤：在步骤 S510 中，确定
20 视频输出图像对应的剩余图像帧时长，并将确定出的剩余图像帧时长与预先设定的时长阈值进行对比。在步骤 S520 中，如果剩余图像帧时长大于时长阈值，则确定跳帧步长，以根据跳帧步长从剩余视频输出图像中确定出目标视频图像。如果剩余图像帧时长小于时长阈值，说明当前视频动作可以在规定时长内归为到初始位置，可以与下一个文本视频动作之间进行平滑过渡，则不进行跳帧压缩处理，结束跳帧压缩处理步骤。
25

在本申请的一种示例性实施例中，确定跳帧步长的可能实现方式可以是获取剩余视频输出图像对应的图像帧率，并确定剩余视频输出图像的剩余图像数量；根据剩余图像数量和图像帧率确定跳帧步长；根据跳帧步长从剩余视频输出图像中确定目标视频图像的实现方式可以是按照时间顺序并根据跳帧步长从剩余视频输出图像中确定目标视频图像。其中，
30 图像帧率可以是直播客户端每秒可以播放的图像的帧数，图像帧率可以采用 Fps 表示。剩余图像数量可以是剩余视频输出图像对应的图像数量。

举例而言，可以将时长阈值设定为 500 毫秒，计算剩余视频输出图像对应的剩余图像帧时长。剩余图像帧时长可以根据图像帧率确定，如剩余图像帧时长的计算公式为： $T = F / Fps * 1000$ （毫秒）。因此，如果计算出的剩余图像帧时长小于 500 毫秒，则不对剩余
35 视频输出图像进行跳帧压缩处理。如果计算出的剩余图像帧时长大于 500 毫秒，则对剩余

视频输出图像执行跳帧压缩处理步骤。具体的，跳帧压缩处理步骤可以包括：确定剩余视频输出图像的剩余图像数量，并获取剩余输出视频图像对应的图像帧率，根据图像帧率和剩余图像数量可以确定出跳帧步长，跳帧步长的计算公式可以为： $S = [F / (Fps / 2)]$ 。在计算出跳帧步长后，可以按照时间顺序，并且每隔跳帧步长从剩余视频输出图像中确定一目标视频图像。

在本申请的一种示例性实施例中，按照时间顺序并根据跳帧步长从剩余视频输出图像中确定目标视频图像的可能实现方式可以是获取剩余视频输出图像中的首帧图像和末帧图像；按照时间顺序从首帧图像开始，每隔跳帧步长从剩余视频输出图像中确定中间图像；将首帧图像、中间图像和末帧图像作为目标视频图像。

10 其中，首帧图像可以是按照时间顺序从剩余视频输出图像中确定出的第一帧图像。末帧图像可以是按照时间顺序从剩余视频输出图像中确定出的最后一帧图像。中间图像可以是按照时间顺序处于剩余视频输出图像中间部分的图像。

在计算出跳帧步长后，可以根据跳帧步长从剩余视频输出图像中确定目标视频图像，具体过程为：首先，获取剩余视频输出图像中的第一帧图像（即首帧图像）和最后一张图像（即末帧图像）。其次，按照时间顺序，每隔 S 帧（即跳帧步长）从剩余输出图像中抽取一张图像作为中间图像；另外，还可以每隔 S 帧，在这 S 帧中随机选取一张图像作为中间图像。重复此步骤，直至对剩余视频输出图像均进行上述处理。最后，可以将获取到的首帧图像、一个或多个中间图像以及末帧图像作为目标视频图像。

20 在本申请的一种示例性实施例中，如果剩余图像帧时长大于时长阈值，则确定跳帧步长，根据跳帧步长从剩余视频输出图像中确定目标视频图像的方式还可以是根据跳帧步长确定目标视频图像的目标图像数量；确定剩余视频输出图像中两两相邻图像之间的图像相似度，根据图像相似度从剩余视频输出图像中确定不超过目标图像数量的目标视频图像。其中，目标图像数量可以从剩余视频输出图像中确定出的目标视频图像的数量，可以记为 N 。图像相似度可以是剩余视频输出图像中两个相邻图像之间的相似程度。

25 参考图 6，图 6 示意性示出了根据本申请的另一实施例的从剩余视频输出图像中确定目标视频图像的流程图。图 6 中示出了根据相邻图像之间的图像相似度从剩余视频输出图像中确定目标视频输出图像的过程。在步骤 S610 中，可以对比剩余视频输出图像对应的剩余图像帧时长与时长阈值的大小关系，如果剩余图像帧时长大于时长阈值，则确定跳帧步长，跳帧步长的计算方式与上文中相同，此处不再进行赘述。在步骤 S620 中，可以根据跳帧步长确定目标视频图像对应的目标图像数量，例如，目标图像数量 N 的计算公式可以为 $N = [F / S]$ 。在步骤 S630 中，可以计算剩余视频输出图像中两两相邻图像之间的图像相似度，并根据图像相似度从剩余视频输出图像中筛选出不超过目标图像数量个目标视频图像。例如，在连续第一数量的剩余视频输出图像中，两两连续图像之间的图像相似度均大于 95%，则可以从这些连续第一数量的剩余视频输出图像中筛选一个图像作为目标视频图像。重复

执行该步骤，直至对所有的剩余视频输出图像完成上述判断，使得确定出的目标视频图像的数量不超过目标图像数量。

在本申请的一种示例性实施例中，根据剩余音频数据和目标视频图像确定目标音频数据的可能实现方式可以是确定剩余音频数据的剩余音频时长，判断剩余音频时长是否大于
5 时长阈值；如果剩余音频时长大于时长阈值，则删除剩余音频数据；确定目标视频图像对应的目标图像帧时长，根据目标图像帧时长生成目标音频数据。其中，剩余音频时长可以是剩余音频数据对应的时间长度。时长阈值可以是预先设定的用于与剩余音频时长进行对比的阈值。目标音频数据可以是生成中断过渡视频流所采用的音频数据。

参考图 7，图 7 示意性示出了根据本申请的一个实施例的根据目标图像帧时长生成目标
10 音频数据的流程图。在步骤 S710 中，可以确定剩余音频数据对应的剩余音频时长，并将剩余音频时长与时长阈值进行对比，判断剩余音频时长与时长阈值的大小关系。在步骤 S720 中，如果剩余音频时长大于时长阈值，则删除剩余音频数据。在步骤 S730 中，确定目标视频图像对应的目标图像帧时长，根据目标图像帧时长生成对应的目标音频数据。例如，在删除剩余音频数据后，可以根据目标图像帧时长生成一段相等时长的静音音频 A，并将静
15 音音频 A 作为目标音频数据。还例如，可以按照时间顺序从剩余音频数据中截取一段时长等于目标图像帧时长的音频片段，并将进行音量调整处理后的该音频片段作为目标音频数据。又例如，可以在虚拟直播视频服务端预先配置一个目标音频库，可以从目标音频库中选取一段时长等于目标图像帧时长的音频片段作为目标音频数据。

在本申请的一种示例性实施例中，若目标视频数据包括中断过渡动画，则响应于直播
20 中断请求从尚未合成直播视频流的待合成视频数据中确定目标视频数据的可能实现方式可以是响应于直播中断请求将尚未合成直播视频流的待合成视频数据中的第一个视频输出图像作为第一中断图像；响应于直播中断请求将尚未合成直播视频流的待合成视频数据中的最后一个视频输出图像作为第二中断图像；根据所述第一中断图像与所述第二中断图像生成预设时长的中断过渡动画。其中，第一中断图像可以是按照时间顺序尚未合成直播视频
25 流的待合成视频数据（即剩余视频输出图像）中确定的第一帧图像。第二中断图像可以是按照时间顺序从剩余视频输出图像中确定的最后一帧图像。中断过渡动画可以是根据第一中断图像和第二中断图像生成的动画，中断过渡动画可以反映由第一中断图像转变为第二中断图像的过渡状态。

参考图 8，图 8 示意性示出了根据本申请的一个实施例的生成中断过渡动画的流程图。
30 在步骤 S810 中，当虚拟视频直播服务端接收到直播中断请求后，可以从尚未合成直播视频流的待合成视频数据中确定第一中断图像，即将剩余视频输出图像中的第一个视频输出图像作为第一中断图像。在步骤 S820 中，虚拟视频直播服务端可以从尚未合成直播视频流的待合成视频数据中确定第二中断图像，即将剩余视频输出图像中的最后一个视频输出图像作为第二中断图像。在步骤 S830 中，虚拟视频直播服务端可以根据第一中断图像和第二中
35 断图像生成预设时长的中断过渡动画，以便将生成的中断过渡动画和目标音频数据作为目标视频数据。举例而言，可以在虚拟视频直播服务端配置过渡动画库，过渡动画库中可以

预先存储一些关键表情特征图像对应的过渡动画，在确定出第一中断图像和第二中断图像后，可以从过渡动画库中匹配出与这两个图像对应且播放时长为预设时长的过渡动画作为中断过渡动画。另外，在确定出第一中断图像和第二中断图像后，还可以实时生成一个预设时长的中断过渡动画。

5 在步骤 S340 中，根据目标视频数据合成中断过渡视频流，并将中断过渡视频流推送至直播客户端。

本示例实施方式中，中断过渡视频流可以是对目标视频数据进行视频流合成处理后所生成的视频流。当虚拟视频直播服务端确定出目标视频数据后，可以根据目标视频数据合成中断过渡视频流，并将中断过渡视频流推送至直播客户端，以便直播客户端将获取到的中断过渡视频流对应的虚拟直播视频实时呈现在用户界面中。

参考图 9，图 9 示意性示出了根据本申请的一个实施例的虚拟视频推流服务的实现框架图。在图 9 中，虚拟视频推流服务可以包括文本转语音模块、三维渲染模块和视频推流模块。

15 当直播客户端向虚拟视频直播服务端发送文本数据时，直播客户端发送的文本数据可以是虚拟直播视频中虚拟对象将说出的文字；另外，在某些场景下，直播客户端发送的文本数据可以是虚拟对象将说出的部分文字，因此，可以确定与该文本数据对应的关联文本数据，并确定出目标文本。

20 当虚拟视频直播服务端确定出目标文本后，文本转语音模块可以根据目标文本生成对应的音频数据和表情特征数据，将音频数据和表情特征数据打包为一个待合成视频数据的数据包，并将待合成视频数据包推送至三维渲染模块。例如，文本转语音模块可以流式地获取目标文本相对应的音频数据与表情特征数据，文本转语音模块每获取到一个包含音频数据与表情特征数据的数据包，就将该数据包推送给三维渲染模块；当虚拟视频直播服务端接收到直播中断请求或直播结束请求后，可以将对应的中断信号或结束信号转发至三维渲染模块。

25 三维渲染模块接收到待合成视频数据包时，可以提取其中的表情特征数据进行三维渲染获取一组对应脸部图像，同时将一张脸部图像与一张背景图像进行图像融合处理合成为一张完整的图像，得到一组完整的视频输出图像，并与音频数据打包，一起推送给视频推流模块。当接收到中断信号或结束信号时，三维渲染模块可以将尚未发送至视频推流模块的图像数据和音频数据进行跳帧压缩处理，并将经过跳帧压缩处理后生成目标视频数据发送

30 送至视频推流模块。

视频推流模块每获取到一个三维渲染模块推送过来的数据包，可以提取其中的音频数据与图像数据，通过快进动态图像专家组 (Fast Forward Moving Picture Expert Group, FFMPEG) 工具将音频数据、图像帧数据进行合成处理后生成直播视频流或中断过渡视频流，并将直播视频流或中断过渡视频流同步推送至直播客户端。当视频推流模块接收到一个中

断信号后，可以跳转至获取文本数据的步骤；当视频推流模块接收到结束信号后，可以结束本次视频推流服务，断开与直播客户端之间的通信连接。

综上所述，本申请中的虚拟视频直播处理方法，获取文本数据，并确定与文本数据对应的待合成视频数据；根据待合成视频数据实时合成直播视频流并将直播视频流推送至直播客户端；在接收到直播中断请求时，响应于直播中断请求从尚未合成直播视频流的待合成视频数据中确定目标视频数据；根据目标视频数据合成中断过渡视频流，并将中断过渡视频流推送至直播客户端。一方面，通过文本数据生成对应的待合成视频数据，在视频直播过程中，即使没有稳定的数据源也可以实时生成直播视频流。另一方面，避免采用硬切换进行视频中断处理时存在视频卡顿的问题，可以有效进行直播视频被中断时的平滑过渡处理，提高虚拟视频直播的流畅性。又一方面，本申请在解决直播视频被中断的平滑过渡问题的同时，考虑了直播视频的实时性，将平滑过渡过程控制在一个设定的时长内，使得当前视频动作可以与下一文本视频动作平滑过渡，同时可以减小对视频实时性的影响。

应当注意，尽管在附图中以特定顺序描述了本申请中方法的各个步骤，但是，这并非要求或者暗示必须按照该特定顺序来执行这些步骤，或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的，可以省略某些步骤，将多个步骤合并为一个步骤执行，以及/或者将一个步骤分解为多个步骤执行等。

进一步的，本示例实施方式中，还提供了一种虚拟视频直播处理装置。该虚拟视频直播处理装置可以应用于服务器或终端设备。参考图 10 所示，该虚拟视频直播处理装置 1000 可以包括数据确定模块 1010、第一视频合成模块 1020、中断响应模块 1030 以及第二视频合成模块 1040。其中：

数据确定模块 1010，用于获取文本数据，并确定与所述文本数据对应的待合成视频数据；

第一视频合成模块 1020，用于根据所述待合成视频数据实时合成直播视频流并将所述直播视频流推送至直播客户端；

中断响应模块 1030，用于在接收到直播中断请求时，响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定目标视频数据；

第二视频合成模块 1040，用于根据所述目标视频数据合成中断过渡视频流，并将所述中断过渡视频流推送至所述直播客户端。

在本申请的一种示例性实施例中，所述数据确定模块包括：目标文本确定单元，用于确定与所述文本数据相关的关联文本数据，将所述文本数据和所述关联文本数据作为目标文本；音频特征生成单元，用于将所述目标文本转化为对应的目标词向量，对所述目标词向量进行编码处理和解码处理，以生成所述目标词向量的音频特征；音频数据生成单元，用于对所述音频特征进行合成处理，以生成音频数据；图像生成单元，用于确定与所述目标文本对应的表情特征数据，并根据所述表情特征数据生成视频输出图像。

在本申请的一种示例性实施例中，图像生成单元包括：脸部图像生成单元，用于对所述表情特征数据进行渲染处理以生成虚拟对象对应的脸部图像；背景图像确定单元，用于确定与所述脸部图像匹配的背景图像；所述背景图像包括所述虚拟对象的肢体动作信息和环境背景信息；图像融合单元，用于对所述脸部图像与所述背景图像进行图像融合处理以生成所述视频输出图像。

在本申请的一种示例性实施例中，背景图像确定单元包括：音频时长确定子单元，用于确定与所述脸部图像时序对齐的所述音频数据的音频时长；视频时长确定子单元，用于根据所述音频时长确定输出视频的视频时长；背景图像确定子单元，用于根据所述视频时长从预先存储的背景图像集中确定与所述脸部图像匹配的背景图像。

10 在本申请的一种示例性实施例中，背景图像确定子单元被配置为：确定所述视频时长下所述输出视频中包含的所述虚拟对象的对象动作数量；根据所述对象动作数量从所述背景图像集中确定同样数量组别的背景图像，作为与所述脸部图像匹配的背景图像。

15 在本申请的一种示例性实施例中，中断响应模块包括：目标图像确定单元，用于响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定剩余视频输出数据，并根据所述剩余视频输出数据确定目标视频图像；目标音频确定单元，用于响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定剩余音频数据，并根据所述剩余音频数据和所述目标视频图像确定目标音频数据；视频数据确定单元，用于将所述目标视频图像和所述目标音频数据作为所述目标视频数据。

20 在本申请的一种示例性实施例中，目标图像确定单元包括：时长判断子单元，用于确定所述剩余视频输出图像对应的剩余图像帧时长；目标图像确定子单元，用于如果所述剩余图像帧时长大于时长阈值，则确定跳帧步长，根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像。

25 在本申请的一种示例性实施例中，目标图像确定子单元包括：跳帧步长确定子单元，用于获取所述剩余视频输出图像对应的图像帧率，并确定所述剩余视频输出图像的剩余图像数量；根据所述剩余图像数量和所述图像帧率确定所述跳帧步长；图像确定子单元，用于按照时间顺序并根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像。

30 在本申请的一种示例性实施例中，图像确定子单元被配置为：获取所述剩余视频输出图像中的首帧图像和末帧图像；按照所述时间顺序从所述首帧图像开始，每隔所述跳帧步长从所述剩余视频输出图像中确定中间图像；将所述首帧图像、所述中间图像和所述末帧图像作为所述目标视频图像。

在本申请的一种示例性实施例中，目标图像确定单元还可以被配置为：根据所述跳帧步长确定所述目标视频图像的目标图像数量；确定所述剩余视频输出图像中两两相邻图像之间的图像相似度，根据所述图像相似度从所述剩余视频输出图像中确定不超过所述目标图像数量的所述目标视频图像。

在本申请的一种示例性实施例中，目标音频确定单元被配置为：确定所述剩余音频数据的剩余音频时长；如果所述剩余音频时长大于所述时长阈值，则删除所述剩余音频数据；确定所述目标视频图像对应的目标图像帧时长，根据所述目标图像帧时长生成所述目标音频数据。

5 在本申请的一种示例性实施例中，所述中断响应模块还包括动画生成子单元，动画生成子单元被配置为：响应于所述直播中断请求将尚未合成直播视频流的所述待合成视频数据中的第一个视频输出图像作为第一中断图像；响应于所述直播中断请求将尚未合成直播视频流的所述待合成视频数据中的最后一个视频输出图像作为第二中断图像；根据所述第一中断图像与所述第二中断图像生成预设时长的中断过渡动画。

10 上述虚拟视频直播处理装置中各模块或单元的具体细节已经在对应的虚拟视频直播处理方法中进行了详细的描述，因此此处不再赘述。

 应当注意，尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元，但是这种划分并非强制性的。实际上，根据本申请的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

15 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本申请的真正范围和精神由下面的权利要求指出。

20 应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

权 利 要 求

- 1、一种虚拟视频直播处理方法，所述方法由电子设备执行，所述方法包括：
- 获取文本数据，并确定与所述文本数据对应的待合成视频数据；
- 根据所述待合成视频数据实时合成直播视频流并将所述直播视频流推送至直播客户端；
- 5 在接收到直播中断请求时，响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定目标视频数据；
- 根据所述目标视频数据合成中断过渡视频流，并将所述中断过渡视频流推送至所述直播客户端。
- 2、根据权利要求1所述的虚拟视频直播处理方法，所述待合成视频数据包括音频数据和视频输出图像，所述确定与所述文本数据对应的待合成视频数据，包括：
- 10 确定与所述文本数据相关的关联文本数据，将所述文本数据和所述关联文本数据作为目标文本；
- 将所述目标文本转化为对应的目标词向量，对所述目标词向量进行编码处理和解码处理，以生成所述目标词向量的音频特征；
- 15 对所述音频特征进行合成处理，以生成所述音频数据；
- 确定与所述目标文本对应的表情特征数据，并根据所述表情特征数据生成所述视频输出图像。
- 3、根据权利要求2所述的虚拟视频直播处理方法，所述根据所述表情特征数据生成所述视频输出图像，包括：
- 20 对所述表情特征数据进行渲染处理以生成虚拟对象对应的脸部图像；
- 确定与所述脸部图像匹配的背景图像；所述背景图像包括所述虚拟对象的肢体动作信息和环境背景信息；
- 对所述脸部图像与所述背景图像进行图像融合处理以生成所述视频输出图像。
- 4、根据权利要求3所述的虚拟视频直播处理方法，所述确定与所述脸部图像匹配的背景图像，包括：
- 25 确定与所述脸部图像时序对齐的所述音频数据的音频时长；
- 根据所述音频时长确定输出视频的视频时长；
- 根据所述视频时长从预先存储的背景图像集中确定与所述脸部图像匹配的背景图像。

5、根据权利要求4所述的虚拟视频直播处理方法，所述根据所述视频时长从预先存储的背景图像集中确定与所述脸部图像匹配的背景图像，包括：

确定所述视频时长下所述输出视频中包含的所述虚拟对象的对象动作数量；

5 根据所述对象动作数量从所述背景图像集中确定同样数量组别的背景图像，作为与所述脸部图像匹配的背景图像。

6、根据权利要求1所述的虚拟视频直播处理方法，所述响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定目标视频数据，包括：

响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定剩余视频输出数据，并根据所述剩余视频输出数据确定目标视频图像；

10 响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定剩余音频数据，并根据所述剩余音频数据和所述目标视频图像确定目标音频数据；

将所述目标视频图像和所述目标音频数据作为所述目标视频数据。

7、根据权利要求6所述的虚拟视频直播处理方法，所述剩余视频输出数据包括剩余视频输出图像，所述根据所述剩余视频输出数据确定目标视频图像，包括：

15 确定所述剩余视频输出图像对应的剩余图像帧时长；

如果所述剩余图像帧时长大于时长阈值，则确定跳帧步长；

根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像。

8、根据权利要求7所述的虚拟视频直播处理方法，所述确定跳帧步长，包括：

20 获取所述剩余视频输出图像对应的图像帧率，并确定所述剩余视频输出图像的剩余图像数量；

根据所述剩余图像数量和所述图像帧率确定所述跳帧步长；

所述根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像，包括：

按照时间顺序并根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像。

25 9、根据权利要求8所述的虚拟视频直播处理方法，所述按照时间顺序并根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像，包括：

获取所述剩余视频输出图像中的首帧图像和末帧图像；

按照所述时间顺序从所述首帧图像开始，每隔所述跳帧步长从所述剩余视频输出图像中确定中间图像；

将所述首帧图像、所述中间图像和所述末帧图像作为所述目标视频图像。

10、根据权利要求 7 所述的虚拟视频直播处理方法，所述根据所述跳帧步长从所述剩余视频输出图像中确定所述目标视频图像，包括：

根据所述跳帧步长确定所述目标视频图像的目标图像数量；

确定所述剩余视频输出图像中两两相邻图像之间的图像相似度；

5 根据所述图像相似度从所述剩余视频输出图像中确定不超过所述目标图像数量的所述目标视频图像。

11、根据权利要求 6 所述的虚拟视频直播处理方法，所述根据所述剩余音频数据和所述目标视频图像确定目标音频数据，包括：

确定所述剩余音频数据的剩余音频时长；

10 如果所述剩余音频时长大于所述时长阈值，则删除所述剩余音频数据；

确定所述目标视频图像对应的目标图像帧时长；

根据所述目标图像帧时长生成所述目标音频数据。

12、根据权利要求 1 所述的虚拟视频直播处理方法，所述目标视频数据包括中断过渡动画，所述响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定
15 目标视频数据，包括：

响应于所述直播中断请求将尚未合成直播视频流的所述待合成视频数据中的第一个视频输出图像作为第一中断图像；

响应于所述直播中断请求将尚未合成直播视频流的所述待合成视频数据中的最后一个视频输出图像作为第二中断图像；

20 根据所述第一中断图像与所述第二中断图像生成预设时长的所述中断过渡动画。

13、一种虚拟视频直播处理装置，所述装置部署在电子设备中，所述装置包括：

数据确定模块，用于获取文本数据，并确定与所述文本数据对应的待合成视频数据；

第一视频合成模块，用于根据所述待合成视频数据实时合成直播视频流并将所述直播视频流推送至直播客户端；

25 中断响应模块，用于在接收到直播中断请求时，响应于所述直播中断请求从尚未合成直播视频流的所述待合成视频数据中确定目标视频数据；

第二视频合成模块，用于根据所述目标视频数据合成中断过渡视频流，并将所述中断过渡视频流推送至所述直播客户端。

14、一种用于虚拟视频直播处理的电子设备，包括：

处理器；以及

存储器，所述存储器上存储有计算机可读指令，所述计算机可读指令被所述处理器执行时实现如权利要求 1 至 12 中任意一项所述的虚拟视频直播处理方法。

5 15、一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求 1 至 12 中任意一项所述的虚拟视频直播处理方法。

16、一种计算机程序产品，当所述计算机程序产品被执行时，用于实现如权利要求 1 至 12 中任一项所述的虚拟视频直播处理方法。