

用于生成目标视频的方法、装置、服务器和介质

相关申请的交叉引用

5 本申请要求于 2020 年 8 月 12 日提交中国专利局、申请号为 202010806612.5、发明名称为“用于生成目标视频的方法、装置、服务器和介质”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。

技术领域

10 本申请实施例涉及计算机技术领域，具体涉及用于生成目标视频的方法、装置、服务器和介质。

背景技术

随着互联网技术的飞速发展，视频直播的应用也越来越广泛。
15 相关的方式通常是首先将直播流数据保存成长视频文件，再利用人工从长视频文件中截取所需要的片段，生成短视频。

发明内容

本申请实施例提出了用于生成目标视频的方法、装置、服务器和介质。
20 第一方面，本申请实施例提供了一种用于生成目标视频的方法，该方法包括：获取直播流数据，其中，直播流数据包括语音数据、直播交互数据中的至少一项和视频数据；对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值、交互度量值中的至少一项和视频度量值；根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值；响应于确定直播流数据的综合度量值满足预设条件，基于直播流数据生成目标视频。
25 在一些实施例中，上述对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的视频度量值，包括：对视频数据中的视频帧进行图像识别，分别确定属于第一预设类别图像和属于第二预设类别图像的数目；根据所确定的属于第一预设类别图像和属于第二预设类别图像的数目，生成视频度量值。
30 在一些实施例中，上述根据所确定的属于第一预设类别图像和属于第二预设类别图像的数目，生成视频度量值，包括：获取与第一预设类别图像和第二预设类别图像分别对应的预设图像权重值；将所确定的属于第一预设类别图像和属于第二预设类别图像的数目与分别对应的图像预设权重值进行加权求和，生成上述视频度量值。
35 在一些实施例中，上述直播流数据包括语音数据；以及上述对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值，包括：对语音数据进行语音识别，生成语音识别文本；分别确定语音识别文本中所包括的文本属于第一预设类别文本和属于第二预设类别文本的数目；根据所确定的属于第一预设类别文本和属于第二预设类别文本的数目，生成语音度量值。

在一些实施例中，上述直播流数据包括直播交互数据；以及上述对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的交互度量值，包括：确定直播交互数据所指示的目标交互行为的数目；根据所确定的目标交互行为的数目，生成交互度量值。

5 在一些实施例中，上述目标交互行为包括第一预设交互行为、第二预设交互行为、第三预设交互行为中的至少两者；以及上述根据所确定的目标交互行为的数目，生成交互度量值，包括：根据所确定的目标交互行为的数目和第一预设交互行为、第二预设交互行为、第三预设交互行为对应的预设交互权重进行加权求和，生成交互度量值。

10 在一些实施例中，上述根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值，包括：获取与所生成的语音度量值、交互度量值中的至少一项和视频度量值各自对应的预设度量权重；对所生成的语音度量值、交互度量值中的至少一项和视频度量值进行归一化；将归一化后的语音度量值、交互度量值中的至少一项和视频度量值进行加权求和，生成直播流数据的综合度量值。

15 在一些实施例中，上述预设条件包括：与直播流数据关联的直播流切片集合中满足综合度量值条件的直播流切片的数目大于目标数目，其中，综合度量值条件包括直播流切片对应的综合度量值小于直播流数据的综合度量值。

在一些实施例中，上述直播流数据包括语音数据；以及上述基于直播流数据生成目标视频，包括：根据语音数据对应的识别文本的语句完整性，确定直播流数据的剪辑起止位置；基于剪辑后的直播流数据生成目标视频。

20 在一些实施例中，上述基于剪辑后的直播流数据生成目标视频，包括：向剪辑后的视频流数据添加特效，生成目标视频。

第二方面，本申请实施例提供了一种用于生成目标视频的装置，该装置包括：获取单元，被配置成获取直播流数据，其中，直播流数据包括语音数据、直播交互数据中的至少一项和视频数据；处理单元，被配置成对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值、交互度量值中的至少一项和视频度量值；第一生成单元，被配置成
25 根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值；第二生成单元，被配置成响应于确定直播流数据的综合度量值满足预设条件，基于直播流数据生成目标视频。

30 在一些实施例中，上述处理单元包括：第一识别子单元，被配置成对视频数据中的视频帧进行图像识别，分别确定属于第一预设类别图像和属于第二预设类别图像的数目；第一生成子单元，被配置成根据所确定的属于第一预设类别图像和属于第二预设类别图像的数目，生成视频度量值。

35 在一些实施例中，上述第一生成子单元包括：获取模块，被配置成获取与第一预设类别图像和第二预设类别图像分别对应的预设图像权重值；生成模块，被配置成将所确定的属于第一预设类别图像和属于第二预设类别图像的数目与分别对应的图像预设权重值进行加权求和，生成上述视频度量值。

在一些实施例中，上述直播流数据包括语音数据；以及上述处理单元包括：第二识别子单元，被配置成对语音数据进行语音识别，生成语音识别文本；第一确定子单元，被配置成分别确定语音识别文本中所包括的文本属于第一预设类别文本和属于第二预设类别文本的数目；第二生成子单元，被配置成根据所确定的属于第一预设类别文本和属于第二预设类别文

本的数目，生成语音度量值。

在一些实施例中，上述直播流数据包括直播交互数据；以及上述处理单元包括：第二确定子单元，被配置成确定直播交互数据所指示的目标交互行为的数目；第三生成子单元，被配置成根据所确定的目标交互行为的数目，生成交互度量值。

- 5 在一些实施例中，上述目标交互行为包括第一预设交互行为、第二预设交互行为、第三预设交互行为中的至少两者；以及上述第三生成子单元进一步被配置成：根据所确定的目标交互行为的数目和第一预设交互行为、第二预设交互行为、第三预设交互行为对应的预设交互权重进行加权求和，生成交互度量值。

- 10 在一些实施例中，上述第一生成单元包括：获取子单元，被配置成获取与所生成的语音度量值、交互度量值中的至少一项和视频度量值各自对应的预设度量权重；归一化子单元，被配置成对所生成的语音度量值、交互度量值中的至少一项和视频度量值进行归一化；第四生成子单元，被配置成将归一化后的语音度量值、交互度量值中的至少一项和视频度量值进行加权求和，生成直播流数据的综合度量值。

- 15 在一些实施例中，上述预设条件包括：与直播流数据关联的直播流切片集合中满足综合度量值条件的直播流切片的数目大于目标数目，其中，综合度量值条件包括直播流切片对应的综合度量值小于直播流数据的综合度量值。

在一些实施例中，上述直播流数据包括语音数据；以及上述第二生成单元包括：第三确定子单元，被配置成根据语音数据对应的识别文本的语句完整性，确定直播流数据的剪辑起止位置；第五生成子单元，被配置成基于剪辑后的直播流数据生成目标视频。

- 20 在一些实施例中，上述第五生成子单元进一步被配置成：向剪辑后的视频流数据添加特效，生成目标视频。

第三方面，本申请实施例提供了一种服务器，该服务器包括：一个或多个处理器；存储装置，其上存储有一个或多个程序；当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现如第一方面中任一实现方式描述的方法。

- 25 第四方面，本申请实施例提供了一种计算机可读介质，其上存储有计算机程序，该程序被处理器执行时实现如第一方面中任一实现方式描述的方法。

第五方面，本申请实施例还提供了一种计算机程序产品，包括计算机程序指令，该计算机程序指令使得计算机执行如第一方面中任一实现方式描述的方法。

- 30 第六方面，本申请实施例还提供了一种计算机程序，当计算机程序在计算机上运行时，使得计算机执行如第一方面中任一实现方式描述的方法。

- 35 本申请实施例提供的用于生成目标视频的方法、装置、服务器和介质，通过对获取的直播流数据中包括的语音数据、直播交互数据中的至少一项和视频数据分别进行处理，生成由语音度量值、交互度量值中的至少一项和视频度量值而得到的综合度量值，最终生成目标视频。从而一方面实现了目标视频的自动生成，另一方面从语音、直播交互中的至少一项和视频等多方面综合选定目标视频的生成基础，提高了所生成的目标视频的质量和生成效率。

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更加明显：

- 图 1 是本申请的一个实施例可以应用于其中的示例性系统架构图；
图 2 是根据本申请的用于生成目标视频的方法的一个实施例的流程图；
图 3 是根据本申请的实施例的用于生成目标视频的方法的一个应用场景的示意图；
图 4 是根据本申请的用于生成目标视频的方法的又一个实施例的流程图；
5 图 5 是根据本申请的用于生成目标视频的装置的一个实施例的结构示意图；
图 6 是适于用来实现本申请的实施例的电子设备的结构示意图。

具体实施方式

- 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本公开，而非对本公开的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关本公开相关的部分。

需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

- 图 1 示出了可以应用本申请的用于生成目标视频的方法或用于生成目标视频的装置的示例性架构 100。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103，网络 104 和服务器 105。网络 104 用以在终端设备 101、102、103 和服务器 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 通过网络 104 与服务器 105 交互，以接收或发送消息等。终端设备 101、102、103 上可以安装有各种通讯客户端应用，例如网页浏览器应用、购物类应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件、文本编辑类应用、视频直播类应用等。

- 终端设备 101、102、103 可以是硬件，也可以是软件。当终端设备 101、102、103 为硬件时，可以是具有显示屏并且支持音视频传输的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。当终端设备 101、102、103 为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块（例如用来提供分布式服务的软件或软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

- 服务器 105 可以是提供各种服务的服务器，例如为终端设备 101、102、103 上视频直播应用提供支持的后台服务器。后台服务器可以对接收的直播流数据进行分析等处理，并将处理结果（如目标视频）反馈给终端设备。

需要说明的是，上述直播流数据也可以直接存储在服务器 105 的本地，服务器 105 可以直接提取本地所存储的直播流数据并进行处理，此时，可以不存在终端设备 101、102、103 和网络 104。

- 需要说明的是，服务器可以是硬件，也可以是软件。当服务器为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器为软件时，可以实现成多个软件或软件模块（例如用来提供分布式服务的软件或软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

需要说明的是，本申请实施例所提供的用于生成目标视频的方法一般由服务器 105 执行，相应地，用于生成目标视频的装置一般设置于服务器 105 中。

需要说明的是，终端 101、102、103 可用于执行该用于生成目标视频的方法；终端 101 也可以采集直播流数据，将采集到的直播流数据发送至服务器 105，使得服务器 105 执行该用于生成目标视频的方法。

- 应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，
5 可以具有任意数目的终端设备、网络和服务器。

继续参考图 2，示出了根据本申请的用于生成目标视频的方法的一个实施例的流程 200。该用于生成目标视频的方法包括以下步骤：

步骤 201，获取直播流数据。

- 在本实施例中，用于生成目标视频的方法的执行主体（如图 1 所示的服务器 105）可以通过有线连接方式或者无线连接方式获取直播流数据。其中，上述直播流数据可以包括语音数据、直播交互数据中的至少一项和视频数据。从而，上述直播流数据可以包括语音数据和视频数据，也可以包括直播交互数据和视频数据，还可以包括语音数据、直播交互数据和视频数据。上述语音数据通常与上述视频数据在时间上同步。上述直播交互数据可以包括用于记录直播中主播与观众的互动情况的数据。上述直播交互数据可以包括但不限于以下至少一项：
10 预设时段（例如每分钟）中弹幕的数目，预设时段（例如每分钟）中表征赞同主播（例如点赞、送礼物）的数目，预设时段（例如每分钟）中评论、留言的数目。
15

- 作为示例，上述执行主体可以从与之通信连接的电子设备（例如图 1 所示的终端设备）实时获取直播流数据。作为又一示例，上述执行主体可以获取预先存储于本地的直播流数据。其中，上述直播流数据可以是预先存储的、对历史直播流数据进行视频切片所得到的。上述
20 视频切片还可以对应有在上述历史直播流数据中的起始时间和结束时间。

步骤 202，对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值、交互度量值中的至少一项和视频度量值。

- 在本实施例中，上述执行主体可以通过各种方式对上述步骤 201 所获取的直播流数据进行处理；以及根据处理结果所包括的目标对象，上述执行主体可以生成对应的语音度量值、
25 交互度量值中的至少一项和视频度量值。

- 在本实施例中，作为示例，上述执行主体可以通过各种方式从所获取的直播流数据中的语音数据提取声学特征。其中，上述声学特征可以包括但不限于以下至少一项：梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient, MFCC），线性预测倒谱系数（Linear Prediction Cepstrum Coefficient, LPCC），音调，音色，响度。而后，上述执行主体可以利用各种方式生成与上述所提取的声学特征对应的语音度量值。例如，上述执行主体可以利用预先训练的人工神经网络生成上述所提取的声学特征对应的语音度量值。其中，上述人工神经网络可以通过历史直播流数据的精彩片段对应的语音数据作为正样本、普通片段对应的语音数据作为负样本进行训练。上述语音度量值可以是 0~1 之间的数值。再例如，上述执行主体可以将上述所提取的声学特征与各声学特征对应的预设阈值进行比较，而后根据大于对应的预设阈值的
30 数目生成上述所提取的声学特征对应的语音度量值。
35

在本实施例中，作为示例，上述执行主体可以通过各种方式对所获取的直播流数据中的直播交互数据进行处理，生成对应的交互度量值。例如，上述执行主体可以将弹幕或评论数超过预设阈值的时段数作为上述交互度量值。例如，上述直播流数据包括 5 分钟的数据。第 0~1 分钟的弹幕数目为 15 条，第 1~2 分钟的弹幕数目为 28 条，第 2~3 分钟的弹幕数目为 85

条，第 3~4 分钟的弹幕数目为 66 条，第 4~5 分钟的弹幕数目为 32 条。假定上述预设阈值为 50 条，则上述执行主体可以将上述交互度量值确定为 2。

在本实施例中，作为示例，上述执行主体可以通过各种方式对所获取的直播流数据中的视频数据进行处理，生成对应的视频度量值。例如，上述执行主体可以确定上述直播流数据中包括目标图像的视频帧的数目。根据所确定的包括目标图像的视频帧的数目生成上述直播流数据对应的视频度量值。

在本实施例的一些可选的实现方式中，上述执行主体可以按照如下步骤对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的视频度量值：

第一步，对视频数据中的视频帧进行图像识别，分别确定属于第一预设类别图像和属于第二预设类别图像的数目。

在这些实现方式中，上述执行主体可以利用各种图像识别方法对上述视频数据中的视频帧进行图像识别，分别确定属于第一预设类别图像和属于第二预设类别图像的数目。其中，上述第一预设类别图像和第二预设类别图像可以包括预先设定的与应用场景相关联的图像。作为示例，上述第一预设类别图像可以为扣篮图像，上述第二预设类别图像可以为在三分线外的投篮图像。

可选地，上述执行主体可以从上述视频数据中进行抽帧，例如每 10 帧抽取 1 帧。而后，上述执行主体所抽取的视频帧进行图像识别。从而节省计算资源。

可选地，上述第一预设类别图像可以包括用于表征商品售卖的图像，例如商品图像、价格标签等。上述第二预设类别图像可以包括预设人物图像。其中，上述预设人物例如可以是主播。从而，上述方法可以从图像识别的角度为识别直播售卖视频中的高光时刻提供技术基础。

第二步，根据所确定的属于第一预设类别图像和属于第二预设类别图像的数目，生成视频度量值。

在这些实现方式中，根据所确定的属于第一预设类别图像和属于第二预设类别图像的数目，上述执行主体可以通过各种方式生成视频度量值。作为示例，上述执行主体可以选取上述所确定的属于第一预设类别图像和属于第二预设类别图像的数目中较大的值作为上述视频度量值。作为又一示例，上述执行主体还可以将上述所选取的较大的值与进行图像识别的视频帧的比值作为上述视频度量值。

可选地，上述执行主体还可以首先获取与上述第一预设类别图像和第二预设类别图像分别对应的图像预设权重值（例如 1 和 0.5）。而后，上述执行主体可以将上述所确定的属于第一预设类别图像和属于第二预设类别图像的数目与上述分别对应的图像预设权重值进行加权求和，生成上述视频度量值。

在本实施例的一些可选的实现方式中，基于上述直播流数据包括的语音数据，上述执行主体可以按照如下步骤对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值：

第一步，对语音数据进行语音识别，生成语音识别文本。

在这些实现方式中，上述执行主体可以通过各种语音识别技术对上述步骤 201 所获取的直播流数据包括的语音数据进行识别，生成对应的语音识别文本。

第二步，分别确定语音识别文本中所包括的文本属于第一预设类别文本和属于第二预设

类别文本的数目。

在这些实现方式中，上述执行主体可以通过各自方式分别确定语音识别文本中所包括的文本属于第一预设类别文本和属于第二预设类别文本的数目。其中，上述第一预设类别文本和第二预设类别文本可以包括预先设定的与应用场景相关联的文本。作为示例，上述第一预设类别文本可以包括“好球”、“漂亮”、“真精彩”等预设的描述词；上述第二预设类别文本可以包括“大家看一下”、“注意到”等预设提示词。

可选地，上述第一预设类别文本可以包括商品描述信息。其中，上述商品描述信息可以包括商品名称、商品评价信息（例如“真好吃”、“好用不贵”等）等。上述第二预设类别文本可以包括预设的售卖关键词。其中，上述预设的售卖关键词例如可以包括“上链接”、“快来买”等。从而，上述方法可以从语音识别的角度为识别直播售卖视频中的高光时刻提供技术基础。

第三步，根据所确定的属于第一预设类别文本和属于第二预设类别文本的数目，生成语音度量值。

在这些实现方式中，根据所确定的属于第一预设类别文本和属于第二预设类别文本的数目，上述执行主体可以通过各种方式生成语音度量值。作为示例，上述执行主体可以选取上述所确定的属于第一预设类别文本和属于第二预设类别文本的数目中较大的值作为上述语音度量值。作为又一示例，上述执行主体还可以将上述所选取的较大的值与识别文本中所包括的词的数目的比值作为上述语音度量值。

可选地，上述执行主体还可以首先获取与上述第一预设类别文本和第二预设类别文本分别对应的文本预设权重值。而后，上述执行主体可以将上述所确定的属于第一预设类别文本和属于第二预设类别文本的数目与上述分别对应的文本预设权重值进行加权求和，生成上述语音度量值。

可选地，基于上述第一预设类别文本包括的商品描述信息和上述第二预设类别文本包括的预设的售卖关键词，与上述第一预设类别文本对应的文本预设权重值（例如1）通常小于上述与第二预设类别文本对应的文本预设权重值（例如5）。

在本实施例的一些可选的实现方式中，基于上述直播流数据包括的直播交互数据，上述执行主体可以按照如下步骤对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的交互度量值：

第一步，确定直播交互数据所指示的目标交互行为的数目。

在这些实现方式中，上述执行主体可以根据各种方式确定上述步骤201所获取的直播流数据包括的直播交互数据所指示的目标交互行为的数目。其中，上述目标交互行为可以包括但不限于以下至少一项：发送弹幕，表征赞同主播的行为（例如点赞、送礼物），发送评论，写留言。

第二步，根据所确定的目标交互行为的数目，生成交互度量值。

在这些实现方式中，根据上述第一步所确定的目标交互行为的数目，上述执行主体可以通过各种方式生成交互度量值。作为示例，上述执行主体可以直接将上述所确定的目标交互行为的数目作为上述交互度量值。作为又一示例，上述执行主体还可以将上述所确定的目标交互行为的数目与预设数值的比值作为上述交互度量值。

可选地，上述目标交互行为可以包括第一预设交互行为、第二预设交互行为、第三预设

交互行为中的至少两者。根据所确定的目标交互行为的数目，上述执行主体可以根据所确定的目标交互行为的数目和上述第一预设交互行为、第二预设交互行为、第三预设交互行为对应的预设交互权重进行加权求和，生成上述交互度量值。其中，第一预设交互行为、第二预设交互行为、第三预设交互行为可以包括预先设定的与应用场景相关联的交互行为。

- 5 可选地，上述第一预设交互行为、第二预设交互行为、第三预设交互行为可以分别包括直播画面中出现商品链接、通过上述直播流数据提供的商品链接的生成订单、发送弹幕。从而，上述方法可以从交互行为的角度为识别直播售卖视频中的高光时刻提供技术基础。

步骤 203，根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值。

- 10 在本实施例中，根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，上述执行主体可以通过各种方式生成直播流数据的综合度量值。作为示例，上述执行主体可以从上述所生成的语音度量值、交互度量值中的至少一项和视频度量值中选取最大值作为上述直播流数据的综合度量值。

在本实施例的一些可选的实现方式中，根据所生成的语音度量值、交互度量值中的至少

- 15 一项和视频度量值，上述执行主体可以按照如下步骤生成直播流数据的综合度量值：

第一步，获取与所生成的语音度量值、交互度量值中的至少一项和视频度量值各自对应的预设度量权重。

在这些实现方式中，上述执行主体可以首先获取与所生成的语音度量值、交互度量值中的至少一项和视频度量值各自对应的预设度量权重。其中，上述预设度量权重例如可以是 0.3、

- 20 0.3、0.4。

第二步，对所生成的语音度量值、交互度量值中的至少一项和视频度量值进行归一化。

在这些实现方式中，上述执行主体可以对上述第一步所生成的语音度量值、交互度量值中的至少一项和视频度量值进行归一化。从而使得将归一化后的第一步所生成的语音度量值、交互度量值中的至少一项和视频度量值属于同一个数量级。

- 25 第三步，将归一化后的语音度量值、交互度量值中的至少一项和视频度量值进行加权求和，生成直播流数据的综合度量值。

在这些实现方式中，上述执行主体可以将上述第二步所得到的归一化后的语音度量值、交互度量值中的至少一项和视频度量值进行加权求和，生成直播流数据的综合度量值。

步骤 204，响应于确定直播流数据的综合度量值满足预设条件，基于直播流数据生成目标

- 30 视频。

在本实施例中，响应于确定直播流数据的综合度量值满足预设条件，上述执行主体可以通过各种方式基于直播流数据生成目标视频。作为示例，上述预设条件可以包括上述直播流数据的综合度量值大于预设度量值阈值。作为示例，上述执行主体可以直接将上述直播流数据作为上述目标视频。作为又一示例，上述执行主体可以对上述直播流数据进行后处理，从而得到上述目标视频。其中，上述后处理例如可以包括添加滤镜、调整亮度、调整对比度等。

- 35 在本实施例的一些可选的实现方式中，上述预设条件可以包括：与上述直播流数据关联的直播流切片集合中满足综合度量值条件的直播流切片的数目大于目标数目。其中，上述综合度量值条件可以包括直播流切片对应的综合度量值小于上述直播流数据的综合度量值。上述目标数目可以是根据实际的应用需求预先指定的任意数目。上述目标数目也可以是根据规

则而定的数目，例如上述关联的直播流切片集合中所包括的直播流切片数目乘以预设比例而得到的数目。

作为示例，上述与上述直播流数据关联的直播流切片集合可以包括从对应于同一直播流信息源（例如直播间 id）获取的分时段的直播流数据切片。假设与上述直播流数据关联的直播流切片集合中包括 10 条直播流切片。上述目标数目为 6。若上述直播流切片集合对应的综合度量值小于上述直播流数据的综合度量值的数目大于 6，则满足上述预设条件。

继续参见图 3，图 3 是根据本申请实施例的用于生成目标视频的方法的应用场景的一个示意图。在图 3 的应用场景中，用户 301 使用终端设备 302 进行直播。终端设备 302 将直播流数据 303 发送至后台服务器 304。其中，上述直播流数据可以包括语音数据和视频数据。后台服务器 304 对直播流数据 303 进行处理，根据所包括的表征精彩程度的对象（例如“好球”的语音、“扣篮图像”），生成语音度量值 80 和视频度量值 70（如图 3 中 305 所示）。而后，后台服务器 304 根据所生成的语音度量值和视频度量值进行求平均，生成综合度量值 75（如图 3 中 306 所示）。之后，根据综合度量值 75 大于预设阈值（例如 70），后台服务器 304 可以基于直播流数据 303 生成精彩剪辑视频 307。可选地，后台服务器 304 还可以将所生成的精彩剪辑视频 307 发送至终端设备 302。

目前，现有技术之一通常是首先将直播流数据保存成长视频文件，再利用人工从长视频文件中截取所需要的片段，生成短视频，导致需要耗费大量的人力成本。而本申请的上述实施例提供的方法，通过对获取的直播流数据中包括的语音数据、直播交互数据中的至少一项和视频数据分别进行处理，生成由语音度量值、交互度量值中的至少一项和视频度量值而得到的综合度量值，最终生成目标视频。与通过人工截取的方法生成目标视频相比，本申请提供的方案实现了目标视频的自动生成，有效地减少了人工成本。与仅根据单一维度如音频或视频来生成目标视频的方法相比，本申请提供的方案从语音、直播交互中的至少一项和视频等多方面综合选定目标视频的生成基础，提高了所生成的目标视频的质量和生成效率。

进一步参考图 4，其示出了用于生成目标视频的方法的又一个实施例的流程 400。该用于生成目标视频的方法的流程 400，包括以下步骤：

步骤 401，获取直播流数据。

在本实施例中，上述直播流数据可以包括语音数据和视频数据。

步骤 402，对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值、交互度量值中的至少一项和视频度量值。

步骤 403，根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值。

上述步骤 401、步骤 402、步骤 403 分别与前述实施例中的步骤 201、步骤 202、步骤 203 及其可选的实现方式一致，上文针对步骤 201、步骤 202 和步骤 203 及其可选的实现方式的描述也适用于步骤 401、步骤 402 和步骤 403，此处不再赘述。

步骤 404，响应于确定直播流数据的综合度量值满足预设条件，根据语音数据对应的识别文本的语句连续性，确定直播流数据的剪辑起止位置；基于剪辑后的直播流数据生成目标视频。

在本实施例中，响应于确定直播流数据的综合度量值满足预设条件，用于生成目标视频的方法的执行主体（例如图 1 所示的服务器 105）可以按照如下步骤生成目标视频：

第一步，根据语音数据对应的识别文本的语句完整性，确定直播流数据的剪辑起止位置。

在本实施例中，上述执行主体可以首先根据语音数据对应的识别文本确定语句完整性。而后，根据所确定的语句完整性，上述执行主体可以通过各种方式确定上述直播流数据的剪辑起止位置。其中，上述剪辑的起止位置可以包括剪辑的起始位置和结束位置。作为示例，
5 响应于确定上述语音数据对应的识别文本语句完整（例如“XX 真的很好吃”），上述执行主体可以将上述语音数据的起止位置确定为剪辑的起止位置。作为又一示例，响应于确定上述语音数据对应的识别文本语句不完整（例如“个球太精彩”、“接下来请大家关注第”），上述执行主体可以将仅具有后半句的语句的结束位置确定为剪辑的起始位置，将仅具有前半句的语句的起始位置确定为剪辑的结束位置。

10 第二步，基于剪辑后的直播流数据生成目标视频。

在本实施例中，上述执行主体可以通过各种方式基于剪辑后的直播流数据生成目标视频。作为示例，上述执行主体可以直接将剪辑后的直播流数据确定为上述目标视频。作为又一示例，上述执行主体可以对上述剪辑后的直播流数据进行后处理，根据后处理后的直播流数据生成目标视频。

15 在本实施例的一些可选的实现方式中，上述执行主体还可以向剪辑后的视频流数据添加特效，生成目标视频。其中，上述特效可以包括但不限于以下至少一项：字幕，贴纸，转场效果。

从图 4 中可以看出，本实施例中的用于生成目标视频的方法的流程 400 体现了根据语音数据对应的识别文本的语句完整性，确定直播流数据的剪辑起止位置的步骤，以及基于剪辑
20 后的直播流数据生成目标视频的步骤。由此，本实施例描述的方案可以根据语音数据对应的识别文本的语句完整性生成目标视频，从而保证了目标视频中语句的完整性。

进一步参考图 5，作为对上述各图所示方法的实现，本申请提供了用于生成目标视频的装置的一个实施例，该装置实施例与图 2 或图 4 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

25 如图 5 所示，本实施例提供的用于生成目标视频的装置 500 包括获取单元 501、处理单元 502、第一生成单元 503 和第二生成单元 504。其中，获取单元 501，被配置成获取直播流数据，其中，直播流数据包括语音数据、直播交互数据中的至少一项和视频数据；处理单元 502，被配置成对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值、交互度量值中的至少一项和视频度量值；第一生成单元 503，被配置成根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值；第二生成单元
30 504，被配置成响应于确定直播流数据的综合度量值满足预设条件，基于直播流数据生成目标视频。

在本实施例中，用于生成目标视频的装置 500 中：获取单元 501、处理单元 502、第一生成单元 503 和第二生成单元 504 的具体处理及其所带来的技术效果可分别参考图 2 对应实施例中的步骤 201、步骤 202、步骤 203 和步骤 204 的相关说明，在此不再赘述。
35

在本实施例的一些可选的实现方式中，上述处理单元 502 可以包括第一识别子单元（图中未示出）、第一生成子单元（图中未示出）。其中，上述第一识别子单元，可以被配置成对视频数据中的视频帧进行图像识别，分别确定属于第一预设类别图像和属于第二预设类别图像的数目。上述第一生成子单元，可以被配置成根据所确定的属于第一预设类别图像和属于

第二预设类别图像的数目，生成视频度量值。

在本实施例的一些可选的实现方式中，上述第一生成子单元可以包括获取模块（图中未示出）、生成模块（图中未示出）。其中，上述获取模块，可以被配置成获取与第一预设类别图像和第二预设类别图像分别对应的预设图像权重值。上述生成模块，可以被配置成将所确定的属于第一预设类别图像和属于第二预设类别图像的数目与分别对应的图像预设权重值进行加权求和，生成上述视频度量值。

在本实施例的一些可选的实现方式中，上述直播流数据可以包括语音数据。上述处理单元 502 可以包括第二识别子单元（图中未示出）、第一确定子单元（图中未示出）、第二生成子单元（图中未示出）。其中，上述第二识别子单元，可以被配置成对语音数据进行语音识别，生成语音识别文本。上述第一确定子单元，可以被配置成分别确定语音识别文本中所包括的文本属于第一预设类别文本和属于第二预设类别文本的数目。上述第二生成子单元，可以被配置成根据所确定的属于第一预设类别文本和属于第二预设类别文本的数目，生成语音度量值。

在本实施例的一些可选的实现方式中，上述直播流数据可以包括直播交互数据。上述处理单元 502 可以包括第二确定子单元（图中未示出）、第二生成子单元（图中未示出）。其中，上述第二确定子单元，可以被配置成确定直播交互数据所指示的目标交互行为的数目。上述第三生成子单元，可以被配置成根据所确定的目标交互行为的数目，生成交互度量值。

在本实施例的一些可选的实现方式中，上述目标交互行为可以包括第一预设交互行为、第二预设交互行为、第三预设交互行为中的至少两者。上述第三生成子单元可以进一步被配置成：根据所确定的目标交互行为的数目和第一预设交互行为、第二预设交互行为、第三预设交互行为对应的预设交互权重进行加权求和，生成交互度量值。

在本实施例的一些可选的实现方式中，上述第一生成单元 503 可以包括获取子单元（图中未示出）、归一化子单元（图中未示出）、第四生成子单元（图中未示出）。其中，上述获取子单元，可以被配置成获取与所生成的语音度量值、交互度量值中的至少一项和视频度量值各自对应的预设度量权重。上述归一化子单元，可以被配置成对所生成的语音度量值、交互度量值中的至少一项和视频度量值进行归一化。上述第四生成子单元，可以被配置成将归一化后的语音度量值、交互度量值中的至少一项和视频度量值进行加权求和，生成直播流数据的综合度量值。

在本实施例的一些可选的实现方式中，上述预设条件可以包括：与直播流数据关联的直播流切片集合中满足综合度量值条件的直播流切片的数目大于目标数目，其中，综合度量值条件包括直播流切片对应的综合度量值小于直播流数据的综合度量值。

在本实施例的一些可选的实现方式中，上述直播流数据可以包括语音数据。上述第二生成单元 504 可以包括第三确定子单元（图中未示出）、第五生成子单元（图中未示出）。其中，上述第三确定子单元，可以被配置成根据语音数据对应的识别文本的语句完整性，确定直播流数据的剪辑起止位置。上述第五生成子单元，可以被配置成基于剪辑后的直播流数据生成目标视频。

在本实施例的一些可选的实现方式中，上述第五生成子单元可以被进一步配置成向剪辑后的视频流数据添加特效，生成目标视频。

本申请的上述实施例提供的装置，通过获取单元 501 获取直播流数据。其中，直播流数

据包括语音数据、直播交互数据中的至少一项和视频数据。而后，处理单元 502 对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值、交互度量值中的至少一项和视频度量值。接下来，第一生成单元 503 根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值。最后，第二生成单元 504 响应于

5 确定直播流数据的综合度量值满足预设条件，基于直播流数据生成目标视频。从而一方面实现了目标视频的自动生成，另一方面从语音、直播交互中的至少一项和视频等多方面综合选定目标视频的生成基础，提高了所生成的目标视频的质量和生成效率。

下面参考图 6，其示出了适于用来实现本申请实施例的电子设备（例如图 1 中的服务器）600 的结构示意图。本申请实施例中的终端设备可以包括但不限于诸如移动电话、笔记本电脑、

10 数字广播接收器、PDA（个人数字助理，personal digital assistant）、PAD（平板电脑，portable android device）、PMP（便携式多媒体播放器，personal multimedia player）、车载终端（例如车载导航终端）等等的移动终端以及诸如数字 TV、台式计算机等等的固定终端。图 6 示出的服务器仅仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

如图 6 所示，电子设备 600 可以包括处理装置（例如中央处理器、图形处理器等）601，

15 其可以根据存储在只读存储器（read-only memory，ROM）602 中的程序或者从存储装置 608 加载到随机访问存储器（random access memory，RAM）603 中的程序而执行各种适当的动作和处理。在 RAM 603 中，还存储有电子设备 600 操作所需的各种程序和数据。处理装置 601、ROM 602 以及 RAM 603 通过总线 604 彼此相连。输入/输出（input/output，I/O）接口 605 也连接至总线 604。

20 通常，以下装置可以连接至 I/O 接口 605：包括例如触摸屏、触摸板、键盘、鼠标、摄像头等的输入装置 606；包括例如液晶显示器（LCD，Liquid Crystal Display）、扬声器、振动器等等的输出装置 607；包括例如磁带、硬盘等的存储装置 608；以及通信装置 609。通信装置 609 可以允许电子设备 600 与其他设备进行无线或有线通信以交换数据。虽然图 6 示出了具有各种装置的电子设备 600，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地

25 实施或具备更多或更少的装置。图 6 中示出的每个方框可以代表一个装置，也可以根据需要代表多个装置。

特别地，根据本申请的实施例，上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，

30 该计算机程序可以通过通信装置 609 从网络上被下载和安装，或者从存储装置 608 被安装，或者从 ROM 602 被安装。在该计算机程序被处理装置 601 执行时，执行本申请的实施例的方法中限定的上述功能。

需要说明的是，本申请的实施例所述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（electrical programmable ROM，EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（compact disc ROM，CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。

35

在本申请的实施例中，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。而在本申请的实施例中，计算机可读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，该计算机可读信号介质可以发送、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：电线、光缆、RF（Radio Frequency，射频）等等，或者上述的任意合适的组合。

- 5 上述计算机可读介质可以是上述服务器中所包含的；也可以是单独存在，而未装配入该服务器中。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该服务器执行时，使得该服务器：获取直播流数据，其中，直播流数据包括语音数据、直播交互数据中的至少一项和视频数据；对直播流数据进行处理，根据处理结果所包括的目标对象，生成对应的语音度量值、交互度量值中的至少一项和视频度量值；根据所生成的语音度量值、交互度量值中的至少一项和视频度量值，生成直播流数据的综合度量值；响应于确定直播流数据的综合度量值满足预设条件，基于直播流数据生成目标视频。

本公开还提供了一种计算机程序，该计算机程序使得计算机执行上述实施例所提供的用于生成目标视频的方法。

- 20 可以以一种或多种程序设计语言或其组合来编写用于执行本申请实施例的操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言、Python 或类似的设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(local area network, LAN)或广域网(wide area network, WAN)—一连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

- 25 附图中的流程图和框图，图示了按照本申请的各种实施例的系统、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，该模块、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，框图和/或流程图中的每个方框、以及框图和/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。

- 30 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，例如，可以描述为：一种处理器，包括获取单元、处理单元、第一生成单元、第二生成单元。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，例如，获取单元还可以被描述为“获取直播流数据的单元，其

中，直播流数据包括语音数据、直播交互数据中的至少一项和视频数据”。

- 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请的实施例中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述本公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请实施例中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。
- 5