

METHODS, APPARATUSES, DEVICES FOR GENERATING VIDEO AND TRAINING MODEL, AND STORAGE MEDIUM

TECHNICAL FIELD

[0001] The disclosure relates, but is not limited, to the technical field of image processing, and in particular to a methods, apparatuses and devices for generating a video and training a model, a storage medium and a computer program product.

BACKGROUND

[0002] Text-driven content creation has developed into a transformational technology of innovative creativity. In some application scenes, a video containing an object executing a specific action may be generated by text driving. Compared with the general text-driven video generation, the action executed by the object may be complex, and consistency of appearance of the object needs to be kept in a process of executing the specific action by the object, and therefore, text-driven generation of the video containing the object executing the specific action faces a greater challenge. In the video generated by using a text-driven video generation solution in the related art, it is difficult for the object to keep the consistency of the appearance in the process of executing the corresponding action.

SUMMARY

[0003] In view of this, an embodiment of the disclosure provides methods, apparatuses, devices for generating a video and for training a model, storage media and computer program products.

[0004] The technical solution of the embodiment of the disclosure is realized below.

[0005] In one aspect, an embodiment of the disclosure provides a method for generating a video, which includes the following operations.

[0006] An appearance description text and an action description text of an object are obtained.

[0007] A first appearance feature set of the object is sampled from an appearance codebook based on the appearance description text; the appearance codebook includes a plurality of appearance features; and the first appearance feature set includes first

appearance features, each corresponding to at least one pixel of an image frame in a first video to be generated.

[0008] A first posture feature set sequence of the object is sampled from a posture codebook based on the action description text; the posture codebook includes a plurality of posture features; the first posture feature set sequence includes the first posture feature sets, each respectively corresponding to at least one image frame in the first video to be generated; and each of the first posture feature sets includes the first posture features, each corresponding to at least one pixel in the corresponding image frame.

[0009] The first video containing the object is generated based on the first appearance feature set and the first posture feature set sequence.

[0010] In another aspect, an embodiment of the disclosure provides a method for training a model. The model includes an appearance codebook, a posture codebook and a decoding network. The method includes the following operations.

[0011] A video sample containing a sample object is obtained, and the video sample includes a first image frame and a second image frame.

[0012] Feature extraction is performed on the first image frame by using the appearance codebook to obtain a third appearance feature set, and feature extraction is performed on the second image frame after being augmented by using the posture codebook to obtain a third posture feature set; the appearance codebook includes the plurality of appearance features; the posture codebook includes the plurality of posture features; and the augmenting is configured to reduce appearance information in the second image frame.

[0013] A third image frame is generated based on the third appearance feature set and the third posture feature set by using the decoding network.

[0014] A first loss value is determined based on the third image frame and the second image frame before being augmented .

[0015] The appearance codebook, the posture codebook and network parameters of the decoding network are updated based on the first loss value to obtain the updated appearance codebook, posture codebook and decoding network.

[0016] In another aspect, an embodiment of the disclosure provides an apparatus for generating a video, which includes a first obtaining module, a first sampling module, a second sampling module and a first generation module.

- [0017] The first obtaining module is configured to obtain the appearance description text and the action description text of an object.
- [0018] The first sampling module is configured to sample the first appearance feature set of the object from the appearance codebook based on the appearance description text; the appearance codebook includes a plurality of appearance features; the first appearance feature set includes the first appearance features, each corresponding to at least one pixel of the image frame in the first video to be generated.
- [0019] The second sampling module is configured to sample the first posture feature set sequence of the object from the posture codebook based on the action description text; the posture codebook includes a plurality of posture features; the first posture feature set sequence includes the first posture feature sets, each corresponding to at least one image frame in the first video to be generated; and each of the first posture feature sets includes the first posture features, each corresponding to at least one pixel in the corresponding image frame.
- [0020] The first generation module is configured to generate the first video containing the object based on the first appearance feature set and the first posture feature set sequence.
- [0021] In still another aspect, an embodiment of the disclosure provides an apparatus for training a model. The model includes an appearance codebook, a posture codebook and a decoding network. The apparatus includes a second obtaining module, a first extraction module, a second generation module, a first determination module and a first updating module.
- [0022] The second obtaining module is configured to obtain a video sample containing a sample object, and the video sample includes the first image frame and the second image frame.
- [0023] The first extraction module is configured to perform feature extraction on the first image frame by using the appearance codebook to obtain the third appearance feature set, and perform feature extraction on the second image frame after being augmented by using the posture codebook to obtain a third posture feature set; the appearance codebook includes a plurality of appearance features; the posture codebook includes a plurality of posture features; and the augmenting is configured to reduce appearance information in the second image frame.

- [0024] The second generation module is configured to generate the third image frame based on the third appearance feature set and the third posture feature set by using the decoding network.
- [0025] The first determination module is configured to determine the first loss value based on the third image frame and the second image frame before being augmented .
- [0026] The first updating module is configured to update the appearance codebook, the posture codebook and network parameters of the decoding network based on the first loss value, to obtain the updated appearance codebook, posture codebook and decoding network.
- [0027] In still another aspect, an embodiment of the disclosure provides a computer device, which includes a memory and a processor. The memory stores a computer program capable of running on the processor, and the processor, when executing the program, implements part or all steps of the above method.
- [0028] In yet another aspect, an embodiment of the disclosure provides a computer-readable storage medium, in which a computer program is stored. When the computer program, being executed by a processor, implements part or all steps of the above method.
- [0029] In yet another aspect, an embodiment of the disclosure provides a computer program, which includes a computer-readable code. The computer-readable code, when running in a computer device, causes a processor in the computer device to implement part or all steps in the above method.
- [0030] In still another aspect, an embodiment of the disclosure provides a computer program product, which includes a non-transitory computer-readable storage medium, in which a computer program is stored. The computer program, when being read and executed by a computer, implements part or all steps in the above method.
- [0031] In the embodiment of the disclosure, the distribution space of the appearance feature of the object may be separated from the distribution space of the posture feature by providing one appearance codebook including a plurality of appearance features and one posture codebook including a plurality of posture features, thus the first appearance feature set and the first posture feature set sequence of the object may be sampled separately in the appearance codebook and the posture codebook, furthermore, the same first appearance feature set may be adopted in a process of generating each image frame

of the video so as to reduce drift of the appearance feature of the object along the image frames, such that the appearance of the object in the generated video is well kept between the image frames. In addition, compared with a solution in which the distribution space of the appearance feature and the distribution space of the posture feature are not separated, in the embodiments of the disclosure, as the first appearance feature set and the first posture feature set sequence of the object are sampled separately, on one hand, the posture features of the object are not considered in a process of sampling the appearance features, thereby reducing the complexity for sampling of the appearance features, and sampling the first appearance feature set more accurately; on the other hand, the appearance features of the object are not considered in a process of sampling the posture features, thereby simplifying motion modeling for the object, reducing the complexity for sampling of the posture features, and sampling the first posture feature set sequence more accurately. In this way, the appearance and the action of the object in the generated video are more accurate and the quality and fidelity for generation of the video are effectively improved.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0032] FIG. 1 illustrates an implementation flowchart of a method for generating a video according to an embodiment of the disclosure.
- [0033] FIG. 2 illustrates an implementation flowchart of a method for generating a video according to an embodiment of the disclosure.
- [0034] FIG. 3 illustrates an implementation flowchart of a method for generating a video according to an embodiment of the disclosure.
- [0035] FIG. 4A illustrates an implementation flowchart of a method for generating a video according to an embodiment of the disclosure.
- [0036] FIG. 4B illustrates a block diagram of a composition structure of a video generating model according to an embodiment of the disclosure.
- [0037] FIG. 4C illustrates a schematic diagram of training an appearance codebook, a posture codebook and a decoding network according to an embodiment of the disclosure.
- [0038] FIG. 4D illustrates a schematic diagram of training a first sampling network according to an embodiment of the disclosure.
- [0039] FIG. 4E illustrates a schematic diagram of training a second sampling network

according to an embodiment of the disclosure.

[0040] FIG. 4F illustrates a schematic diagram of an image frame of a human body video generated by a method for generating a video according to an embodiment of the disclosure.

[0041] FIG. 5 illustrates a schematic diagram of a composition structure of an apparatus for generating a video according to an embodiment of the disclosure.

[0042] FIG. 6 illustrates a block diagram of an apparatus for training a model according to an embodiment of the disclosure.

[0043] FIG. 7 illustrates a schematic diagram of a hardware entity of a computer device according to an embodiment of the disclosure.

DETAILED DESCRIPTION

[0044] For making the purposes, technical solutions and advantages of the disclosure clearer, the technical solutions of the disclosure will further be described below in combination with the drawings and embodiments in detail. The described embodiments should not be considered as limits to the disclosure. All other embodiments obtained by those of ordinary skill in the art without creative work shall fall within the scope of protection of the disclosure.

[0045] "Some embodiments" involved in the following descriptions describes a subset of all possible embodiments. However, it can be understood that "some embodiments" may be the same subset or different subsets of all the possible embodiments, and may be combined without conflicts. Term "first/second/third" involved is only for distinguishing similar objects and does not represent a specific sequence of the objects. It can be understood that "first/second/third" may be interchanged to specific sequences or orders if allowed to implement the embodiments of the disclosure described herein in sequences except the illustrated or described ones.

[0046] Unless otherwise defined, all technical and scientific terms in the specification have the same meaning as those skilled in the art, belonging to the disclosure, usually understand. Terms used in the specification are only used for describing the purpose of the disclosure, but not intended to limit the disclosure.

[0047] In order to better understand the embodiment of the disclosure, a text-driven video generating scheme in the related art is described first.

[0048] In the related art, the video containing the object may be synthesized according to the text for describing the appearance and/or action of the object, that is, the video containing the object executing the specific action is generated by text driving. Compared with the general text-driven video generation, the action executed by the object may be complex, and consistency of appearance of the object needs to be kept in a process of executing the specific action by the object, and therefore, text-driven generation of the video containing the object executing the specific action faces a greater challenge. Taking the video containing human body motions as an example, on one hand, as a human body structure is articulated, motions of joints of different body parts form many out-of-plane motions (such as, rotation), which are complex; on the other hand, when executing the complex action, consistency of the appearance of a synthesized person in the video should be kept, for example, the appearance and clothes of the synthesized person after back turning should be kept consistent with the appearance and the clothes of the synthesized person before back turning.

[0049] In implementation modes of the related art, images are generally encoded into uniform feature representation, and a uniform feature codebook is established on this basis; in a process of generating the video, an image feature may be sampled from the uniform feature codebook according to the text for describing the appearance and/or action of the object, and thus the feature is decoded to generate the video. However, since the images are encoded into the uniform feature representation, both appearance information and posture information of the object need to be represented through the feature representation, resulting in feature drift of the appearance of the object among the image frames, such that it is difficult to keep the consistency of appearance of the object in each image frame in the generated video, and the quality and fidelity of the generated video are reduced.

[0050] An embodiment of the disclosure provides a method for generating a video, which may be executed by a processor of a computer device. The computer device refers to a server, a notebook computer, a tablet computer, a desktop computer, a smart TV, a set-top box, a mobile device (such as a mobile phone, a portable video player, a personal digital assistant, a special messaging device and a portable game device) and other devices with data processing capabilities. FIG. 1 illustrates an implementation flowchart of a method for generating a video according to an embodiment of the disclosure. As

illustrated in FIG. 1, the method includes S101 to S104 below.

[0051] At S101, an appearance description text and an action description text of an object are obtained.

[0052] Here, the object refers to an object contained in a first video to be generated. The object may include at least one of creatures and stationary objects, such as, a person, a cat, a dog, a tree, a flower and a car and the like.

[0053] The appearance description text of the object refers to a text for describing the appearance of the object to be contained in the first video to be generated. The action description text of the object refers to a text for describing an action in the video. For example, in a case that the object includes a person, the appearance description text may include a text for describing face and clothes of the person, such as "a woman with long hair", "a person in a white shirt", etc.; the action description text may include a text for describing an action of the person, such as moving to the left, turning backwards, etc. For another example, in the case that the object includes a cat, the appearance description text may include text for describing at least one of hair color, tail length, ear shapes and the like of the cat, such as "a tabby cat with a short tail", "a white cat with blue eyes", and the like; the action description text may include text for describing the action of the cat, such as "jumping upwards", "circling" and the like.

[0054] The above appearance description text and the action description text of the object may be set in advance. The language adopted may include, but is not limited to, at least one of Chinese, English, Japanese, Korean and the like.

[0055] At S102, a first appearance feature set of the object is sampled from an appearance codebook based on the appearance description text; the appearance codebook includes a plurality of appearance features; and the first appearance feature set includes first appearance features, each corresponding to at least one pixel of an image frame in the first video to be generated.

[0056] Here, the appearance codebook is a codebook for storing the appearance feature. The appearance feature in the appearance codebook may be obtained by learning from distribution space of the appearance feature of the object. The distribution space of the appearance feature may be determined according to an encoding mode of the appearance feature in a practical application scene. A value range of each code corresponding to the encoding mode is the distribution space of the appearance feature.

- [0057] During implementation, the first appearance features respectively corresponding to pixels of the image frames in the first video to be generated are sampled from the appearance codebook according to semantic information of the appearance description text to obtain the first appearance feature set.
- [0058] At S103, a first posture feature set sequence of the object is sampled from a posture codebook based on the action description text; the posture codebook includes a plurality of posture features; the first posture feature set sequence includes first posture feature sets, each corresponding to at least one image frame in the first video to be generated; and each of the first posture feature sets includes first posture features, each corresponding to at least one pixel in the corresponding image frame.
- [0059] Here, the posture codebook is a codebook for storing the posture feature. The posture feature in the appearance codebook may be obtained by learning from the distribution space of the posture feature of the object. The distribution space of the posture feature may be determined according to the encoding mode of the posture feature in a practical application scene. A value range of each code corresponding to the encoding mode is the distribution space of the posture feature.
- [0060] During implementation, the first posture features respectively corresponding to pixels of each image frame in the first video to be generated may be sampled from the posture codebook according to the semantic information of the posture description text, to obtain the first posture feature set sequence.
- [0061] At S104, the first video containing the object is generated based on the first appearance feature set and the first posture feature set sequence.
- [0062] Here, the first appearance feature set may represent appearance information of the object in the first video to be generated, and each of the first posture feature sets in the first posture feature set sequence may represent posture information of the object in one image frame of the first video to be generated. Therefore, for each of the first posture feature sets in the first posture feature set sequence, one image frame may be generated based on the first appearance feature set and the first posture feature set. The first video of the object may be obtained according to each generated image frame.
- [0063] In some implementation modes, the first video containing the object may be generated based on the first appearance feature set and the first posture feature set sequence by using any appropriate video synthesis network, for example, the video

synthesis network may include, but is not limited to, at least one of a trained decoder, an image generation network and the like.

[0064] In the embodiment of the disclosure, the distribution space of the appearance feature of the object may be separated from the distribution space of the posture feature of the object by providing one appearance codebook including the plurality of appearance features and one posture codebook including the plurality of posture features, thus the first appearance feature set and the first posture feature set sequence of the object may be sampled separately in the appearance codebook and the posture codebook, and thereby the same first appearance feature set may be adopted in a process of generating the image frames of the video, so as to reduce drift of the appearance feature of the object among the image frames, such that the appearance of the object in the generated video is well kept among the image frames. In addition, compared with a solution in which the distribution space of the appearance feature and the distribution space of the posture feature are not separated, in the embodiments of the disclosure, as the first appearance feature set and the first posture feature set sequence of the object are sampled separately, on one hand, the posture feature of the object does not need to be considered in a process of sampling the appearance feature, thereby reducing the complexity for sampling of the appearance feature and facilitating sampling the first appearance feature set more accurately; on the other hand, the appearance feature of the object does not need to be considered in a process of sampling the posture features, thereby simplifying action modeling for the object, reducing the complexity for sampling of the posture feature and facilitating sampling the first posture feature set sequence more accurately. Thus, the appearance and the action of the object in the generated video are more accurate and the quality and fidelity for generation of the video are effectively improved.

[0065] In some embodiments, the above step S102 may include S111 to S112 below.

[0066] At S111, feature extraction is performed on the appearance description text, to obtain a first text feature.

[0067] At S112, the first appearance feature set is sampled from the appearance codebook based on the first text feature.

[0068] The above S103 may include S113 to S114 below.

[0069] At S113, feature extraction is performed on the action description text, to obtain a second text feature.

[0070] At S114, the first posture feature set sequence is sampled from the posture codebook based on the second text feature.

[0071] Here, feature extraction may be performed on the appearance description text by using any appropriate text feature extraction network, to obtain the first text feature, and feature extraction may be performed on the action description text, to obtain the second text feature. During implementation, the text feature extraction network adopted to perform feature extraction on the appearance description text and the text feature extraction network adopted to perform feature extraction on the action description text may be the same or different. For example, the text feature extraction network may include, but is not limited to, at least one of a Bi-directional Long Short Term Memory (Bi-LSTM), a transformer-based encoder network and the like.

[0072] In some implementation modes, the first video includes the plurality of image frames, each corresponding to a respective set of the first posture feature sets; the above S112 may include S121. At S121, the first appearance feature set of the object is sampled from the appearance codebook based on the first text feature by using a first sampling network. Here, the first sampling network may be any appropriate network capable of performing feature sampling on the text feature. It is to be understood that the first sampling network herein and a second sampling network or other sampling network in the following paragraph may include, but is not limited to, at least one of a convolutional neural network, a recurrent neural network, a Transformer sampler, a diffusion-based Transformer sampler and the like.

[0073] The above S114 may include S122. At S122, the first posture feature set sequence of the object is sampled from the posture codebook based on the second text feature by using the second sampling network. Here, the first sampling network may be any appropriate network capable of performing feature sampling on the text feature.

[0074] The above S104 may include S123. At S123, for each of the first posture feature sets in the first posture feature set sequence, one image frame corresponding to the first posture feature set in the first video is generated based on the first appearance feature set and the first posture feature set by using a decoding network. Herein, those skilled in the art may select an appropriate decoding network according to an actual situation. For example, the decoding network may include, but is not limited to, at least one of a trained decoder, an image generation network and the like.

- [0075] In some implementation modes, the first sampling network, the second sampling network and the decoding network may be trained by using any proper method for training a model, to obtain the trained first sampling network, the second sampling network and the decoding network which are configured to execute the above steps S121 to S123.
- [0076] In some implementation modes, each of the appearance features in the appearance codebook is provided with an index value; first index values, each corresponding to at least one pixel of the image frame in the first video to be generated may be predicted based on the first text feature, and the appearance features, each corresponding to a respective one of the first index values in the appearance codebook are determined as the first appearance features, thereby obtaining the first appearance feature set.
- [0077] Similarly, each of the posture features in the posture codebook is provided with an index value; second index values, each corresponding to at least one pixel of each image frame in the first video to be generated may be predicted based on the second text feature, the posture features in the posture codebook, each corresponding to a respective one of the second index values, are further determined as the first posture features, thereby obtaining the first posture feature set sequence.
- [0078] In some embodiments, S112 may include S131 to S132 below.
- [0079] At S131, the appearance features of the object are predicted based on the first text feature, to obtain a second appearance feature set.
- [0080] At S132, the first appearance feature set is determined from the appearance codebook based on the second appearance feature set.
- [0081] Here, the appearance features corresponding to the pixels in the image frame of a video to be synthesized may be predicted first based on the first text feature by using any appropriate sampler, to obtain the second appearance feature set; and then, the appearance features (that is, the first appearance features), each corresponding to a respective one of the second appearance features in the second appearance feature set are determined from the appearance codebook, thereby obtaining the first appearance feature set.
- [0082] In some implementation modes, the appearance features matched with the second appearance features may be determined from the appearance codebook by using

any appropriate matching mode, and are determined as the first appearance features corresponding to the second appearance features. For example, the appearance features, nearest to or the most similar to the second appearance features, in the appearance codebook may be determined as the first appearance features corresponding to the second appearance features.

[0083] In the above embodiment, on one hand, the appearance feature of the object are predicted directly based on the first text feature, to obtain the second appearance feature set, that is, the second appearance features corresponding to the first text feature is predicted directly in the distribution space of the continuous appearance features, a one-to-many mapping prediction problem resulted from feature redundancy in the codebook may be reduced, such that the predicted appearance features are more meaningful, thereby improving quality of the generated video; on the other hand, after the second appearance feature set is obtained by prediction in the distribution space of the continuous appearance features, the first appearance features corresponding to the second appearance features in the second appearance feature set are determined from the appearance codebook, therefore, as each of the first appearance features in the first appearance feature set is determined from the appearance codebook, the rich appearance features stored in the appearance codebook may be fully utilized, thereby improving the accuracy for sampling of the first appearance feature set and further improving the quality of the quality of the generated video. In addition, compared with a solution in which the first index values, each corresponding to at least one pixel of the image frame in the first video to be generated may be predicted based on the first text feature, a process of predicting the second appearance features corresponding to the first text feature in the distribution space of the continuous appearance features is more liable to converge, thereby reducing prediction time and improving the efficiency of video generation.

[0084] In some embodiments, the above S132 may include S141 to S142 below.

[0085] At S141, for each of the second appearance features in the second appearance feature set, the first appearance features matched with the second appearance features are selected from the appearance codebook.

[0086] At S142, the first appearance feature set is obtained based on the first appearance features.

[0087] In some implementation modes, the appearance features, nearest to or the most similar to the second appearance features may be selected from the appearance codebook as the first appearance features matched with the second appearance features. The distance may include, but is not limited to, at least one of the Euclidean distance, Pearson's correlation coefficients, the cosine similarity, the Manhattan distance and the like.

[0088] For example, for the second appearance feature set, the first appearance feature $\hat{f}_{ij}$ matched with each second appearance feature f_{ij} in the second appearance feature set f_{a1} in the appearance codebook C_a may be determined in a way as shown in the formula 1-1:

$$\hat{f}_{ij} = \{\operatorname{argmin}_{c_k \in C_a} \|f_{ij} - c_k\|_2\} \quad (1-1)$$

[0089] Where c_k is the appearance feature in the appearance codebook C_a , the first appearance feature $\hat{f}_{ij}$ is the appearance feature nearest to the second appearance feature f_{ij} in the appearance codebook C_a .

[0090] In some implementation modes, a plurality of appearance features, nearest to or the most similar to each second appearance feature may be selected from the appearance codebook, and one of the plurality of appearance features is selected as the first appearance feature matched with the second appearance feature.

[0091] In the above embodiment, for each of the second appearance features in the second appearance feature set, a first appearance feature matched with the second appearance feature is selected from the appearance codebook, and the first appearance feature set is obtained based on the first appearance features, therefore, the first appearance feature set may be determined rapidly and accurately from the appearance codebook based on the second appearance feature set.

[0092] In some embodiments, the dimension of posture features in the posture codebook is lower than the dimension of appearance features in the appearance codebook. In this way, as the posture feature is a feature for representing overall features of the object, compared with the appearance feature, information concerned by the posture feature is less. Since the dimension of the posture feature is lower than the dimension of the appearance feature, redundant information in the posture feature may be

reduced, it facilitates extraction of useful information in the posture feature, thus improving the accuracy of the first posture feature set sequence obtained by sampling, and reducing the complexity in sampling of the first posture feature set sequence.

[0093] An embodiment of the disclosure provides a method for generating a video, which may be executed by a processor of a computer device. As illustrated in FIG. 2, the method includes S201 to S207 below.

[0094] At S201, an appearance description text and an action description text of an object are obtained.

[0095] At S202, feature extraction is performed on the appearance description text, to obtain a first text feature.

[0096] At S203, a first appearance feature set of the object is sampled from an appearance codebook based on the first text feature; the appearance codebook includes a plurality of appearance features; and the first appearance feature set includes the first appearance features, each corresponding to at least one pixel of an image frame in a first video to be generated.

[0097] Here, the above S201 to S203 correspond to the above S101, S111 to S112 respectively. During implementation, the implementation modes of the above corresponding steps may be referred to.

[0098] At S204, a first reference posture feature set representing that the object is taking a preset posture is sampled from a posture codebook based on the first text feature.

[0099] Here, the preset posture may be a posture set in advance and configured to provide a reference to generation of the video, such as a model posture of the object. For example, the preset posture may be a front standing posture of a person, or a side sitting posture of a cat.

[00100] The first reference posture feature set may include first reference posture features, each corresponding to at least one pixel of the image frame in the first video to be generated.

[00101] In some implementation modes, a third index value corresponding to the preset posture in the posture codebook may be predicted based on the first text feature, to obtain a predicted third index value set, and the first reference posture feature set representing that the object is taking the preset posture is determined from the posture codebook according to the third index value set.

- [00102] In some implementation modes, a posture feature corresponding to the preset posture may be predicted based on the first text feature, to obtain a predicted posture feature set, and the posture feature matched with the posture feature in the posture codebook is determined as one first reference posture feature according to each posture feature in the posture feature set, to obtain the first reference posture feature set.
- [00103] At S205, feature extraction is performed on the action description text, to obtain a second text feature.
- [00104] Here, the above S205 corresponds to the above S113. During implementation, the implementation modes of the above S113 may be referred to.
- [00105] At S206, a first posture feature set sequence of the object is sampled from the posture codebook based on the second text feature and the first reference posture feature set; the posture codebook includes a plurality of posture features; the first posture feature set sequence includes first posture feature sets, each corresponding to at least one image frame in the first video to be generated; and each of the first posture feature sets includes first posture features, each corresponding to at least one pixel in the corresponding image frame.
- [00106] Here, the first posture feature set sequence of the object is sampled from the posture codebook based on the second text feature and the first reference posture feature set by using any appropriate sampling network.
- [00107] In some implementation modes, each of the posture features in the posture codebook is provided with one index value; second index values, each corresponding to at least one pixel of each image frame in the first video to be generated may be predicted based on the second text feature and the first reference posture feature set, and the posture features, each corresponding to a respective one of the second index values in the posture codebook are determined as the first posture features, thereby obtaining the first posture feature set sequence.
- [00108] In some implementation modes, the posture features, each corresponding to at least one pixel of each image frame in the first video to be generated may be predicted based on the second text feature and the first reference posture feature set, to obtain a predicted seventh posture feature set sequence, and a plurality of first appearance feature sets are determined from the posture codebook based on each of seventh posture feature sets in the seventh posture feature set sequence, to obtain a first appearance feature set

sequence.

[00109] At S207, the first video containing the object is generated based on the first appearance feature set and the first posture feature set sequence.

[00110] Here, the above S207 corresponds to the above S104. During implementation, the implementation mode of the above S104 may be referred to.

[00111] In the embodiment of the disclosure, sampling may be performed by using the first reference posture feature set representing that the object is taking the preset posture as a reference, to obtain the first posture feature set sequence which is more accurate, thereby improving continuity of an action of the object in the generated first video, and further improving the fidelity of the first video.

[00112] In some embodiments, the above S206 may include S211 to S212 below.

[00113] At S211, a second posture feature set sequence is determined; and at least one second posture feature set in the second posture feature set sequence is masked.

[00114] Here, the second posture feature set sequence may be set in advance or determined randomly, which is not limited by the embodiment of the disclosure.

[00115] During implementation, a plurality of second posture feature sets in the second posture feature set sequence may be masked. For example, part or all of the second posture feature sets in the second posture feature set sequence may be masked. For example, the plurality of second posture feature sets located in the middle, in the beginning or in the end of the second posture feature set sequence may be masked.

[00116] It is to be understood that the number of the second posture feature sets in the second posture feature set sequence is the same as the number of the first reference posture features in the first reference posture feature set and the number of the first posture feature sets in the first posture feature set sequence.

[00117] At S212, at least one feature sampling is performed in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence until each of the second posture feature sets in the second posture feature set sequence is obtained, and the sampled second posture feature set sequence is used as the first posture feature set sequence.

[00118] Here, at least one feature sampling is performed in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence by using any appropriate sampling network, to obtain each of the

masked second posture feature sets in the second posture feature set sequence, thereby obtaining each of the second posture feature sets in the second posture feature set sequence.

[00119] In some implementation modes, all of the second posture feature sets in the second posture feature set sequence are masked; feature sampling may be performed in the posture codebook based on the second text feature and the first reference posture feature set, to obtain each of the second posture feature sets in the second posture feature set sequence.

[00120] In some implementation modes, part of the second posture feature sets in the second posture feature set sequence are masked; feature sampling may be performed in the posture codebook based on the second text feature, the first reference posture feature set and at least one unmasked second posture feature set in the second posture feature set sequence, to obtain each of the second posture feature sets in the second posture feature set sequence.

[00121] In some implementation modes, the posture codebook may be sampled through one feature sampling process, to obtain all of the masked second posture feature sets in the second posture feature set sequence, thereby obtaining the first posture feature set sequence.

[00122] In some implementation modes, at least one masked second posture feature set in the second posture feature set sequence may be sampled in one feature sampling process; and in a next feature sampling process, feature sampling is performed in the posture codebook based on the second text feature, the first reference posture feature set and each unmasked second posture feature set in the second posture feature set sequence to obtain at least one masked second posture feature set in the second posture feature set sequence, the procedure is iterated in this way until each masked second posture feature set in the second posture feature set sequence is obtained. During implementation, gradual sampling may be performed by using any appropriate sampling order to obtain each masked second posture feature set in the second posture feature set sequence, which is not limited by the embodiment of the disclosure. For example, gradual sampling may be performed by using the sampling order from front to back, from back to front or from two ends to middle, to obtain each masked second posture feature set in the second posture feature set sequence.

- [00123] In some implementation modes, each masked second posture feature set in the second posture feature set sequence may be predicted by using a non-causal Transformer. Therefore, sampling may be assisted by using overall information, thereby guaranteeing the consistency between the image frames in the generated video.
- [00124] In the above embodiment, as the postures corresponding to image frames in the video are continuous in time, the posture features in the adjacent or similar image frames are correlated; by masking the plurality of second posture feature sets in the second posture feature set sequence and predicting each masked second posture feature set in the second posture feature set sequence based on the second text feature, the first reference posture feature set and the masked second posture feature set sequence, the correlation between the posture features in the adjacent or similar image frames may be better utilized, such that the action of the object in the generated first video is better in continuity, thereby further improving the fidelity of the first video.
- [00125] In some embodiments, all of the second posture feature sets in the second posture feature set sequence are masked. The above step S212 may include S221 to S222 below.
- [00126] At S221, a first one of the second posture feature sets and an n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook are predicted based on the second text feature, the first reference posture feature set and the second posture feature set sequence, to obtain the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets; the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets are updated to the second posture feature set sequence; and n is the number of the second posture feature sets in the second posture feature set sequence.
- [00127] At S222, each of the second posture feature sets between the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook is predicted based on the second text feature, the first reference posture feature set and the updated second posture feature set sequence, and each of the predicted second posture feature sets is updated to the second posture feature set sequence.
- [00128] In the above embodiment, all of the second posture feature sets in the second posture feature set sequence are masked; the second posture feature sets at two ends to

the second posture feature sets in the middle in the second posture feature set sequence are gradually predicted based on the second text feature and the first reference posture feature set, the second posture feature sets in the middle are predicted by using the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets. Therefore, as the first one of the second posture feature sets and the n^{th} one of the second posture feature sets correspond to a first image frame and a last image frame in the video to be generated respectively, in the process of predicting the second posture feature sets corresponding to the middle frames, the continuity of the action of the object in the generated video may be well kept by using the first one of the second posture feature sets and the n^{th} one of the second posture feature sets as a reference, thereby further improving the fidelity of the first video.

[00129] In some embodiments, the above S222 may include at least one of S231 to S232 below.

[00130] At S231, in a case that a predicted $(i-1)^{\text{th}}$ one of the second posture feature sets and a predicted $(n-i+2)^{\text{th}}$ one of the second posture feature sets are updated to the second posture feature set sequence and $n-i+1$ is greater than i , an i^{th} one of the second posture feature sets and an $(n-i+1)^{\text{th}}$ one of the second posture feature sets in the second posture feature set sequence in the posture codebook are predicted based on the second text feature, the reference posture feature set and the updated second posture feature set sequence, and the predicted i^{th} one of the second posture feature sets and $(n-i+1)^{\text{th}}$ one of the second posture feature sets are updated to the second posture feature set sequence, and i is an integer of greater than 1 and $2i$ is not greater than $n+1$.

[00131] Here, it is to be understood that, in a case that the predicted $(i-1)^{\text{th}}$ one of the second posture feature set and $(n-i+2)^{\text{th}}$ one of the second posture feature sets are updated to the second posture feature set sequence and $n-i+1$ is greater than i , at least the i^{th} one of the second posture feature set and the $(n-i+2)^{\text{th}}$ one of the second posture feature sets in the second posture feature set sequence are not determined. In this case, the i^{th} one of the second posture feature sets and the $(n-i+2)^{\text{th}}$ one of the second posture feature sets in the posture codebook may be predicted based on the second text feature, the reference posture feature set and the updated second posture feature set sequence to obtain the predicted i^{th} one of the second posture feature sets and $(n-i+2)^{\text{th}}$ one of the second posture feature sets.

[00132] At S232, in the case that the predicted (i-1)th one of the second posture feature sets and (n-i+2)th one of the second posture features are updated to the second posture feature set sequence and n-i+1 is greater than i, the ith one of the second posture feature set in the second posture feature set sequence in the posture codebook is predicted based on the second text feature, the first reference posture feature set and the updated second posture feature set sequence, and the predicted ith one of the second posture feature sets is updated to the second posture feature set sequence.

[00133] Here, it is to be understood that, in the case that the predicted (i-1)th one of the second posture feature sets and (n-i+2)th one of the second posture features are updated to the second posture feature set sequence and n-i+1 is equal to i, only the ith one of the second posture feature sets in the second posture feature set sequence (i.e., the (n-i+2)th one of the second posture features) is not determined. In this case, the ith one of the second posture feature sets in the posture codebook may be predicted based on the second text feature, the reference posture feature set and the updated second posture feature set sequence to obtain the predicted ith one of the second posture feature sets.

[00134] In the above embodiment, the second posture feature sets in the middle are predicted in consideration of the second text feature, the first reference posture feature set and the currently determined second posture feature sets in the second posture feature set sequence, thereby further improving the continuity of the action of the object in the generated first video.

[00135] In some embodiments, the above step S221 may include S241 below.

[00136] At S241, at least one prediction is performed on the first one of the second posture feature sets and the nth one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, until each of the second posture features in the first one of the second posture feature sets and each of the second posture features in the nth one of the second posture feature sets are obtained, and the first one of the second posture feature sets and the nth one of the second posture feature sets are updated to the second posture feature set sequence.

[00137] Here, when the first one of the second posture feature sets and the nth one of the second posture feature sets in the second posture feature set sequence are predicted, all of the second posture feature sets in the second posture feature set sequence are

masked, and there is no known second posture feature set as a reference. Therefore, in order to reduce missing of the second posture features in the first one of the second posture feature sets and the n^{th} one of the second posture feature sets, at least one prediction may be performed on the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, to obtain each of the second posture features in the first one of the second posture feature sets and each of the second posture features in the n^{th} one of the second posture feature sets, thereby improving integrity of a structure of the object in the first image frame and the last image frame in the generated video.

[00138] During implementation, one prediction may be performed first on the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, to obtain the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets; if there is/are missing second posture feature(s) in the predicted first one of the second posture feature sets or n^{th} one of the second posture feature sets, another prediction may be performed on the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, and the second posture feature obtained by the latest prediction for the first one of the second posture feature sets and/or the n^{th} one of the second posture feature sets are supplemented to the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets; the process is repeated in this way, until each of the second posture features in the first one of the second posture feature sets and each of the second posture features in the n^{th} one of the second posture feature set are obtained.

[00139] In some embodiments, the above step S221 may include S251 to S253 below.

[00140] At S251, the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence are predicted based on the second text feature, the first reference posture feature set and the second posture feature set sequence to obtain a first candidate posture feature set and a second candidate

posture feature set.

[00141] Here, the first candidate posture feature set and the second candidate posture feature set correspond to the first one of the second posture feature sets and the n^{th} one of the second posture feature sets respectively.

[00142] During implementation, the posture features, each corresponding to at least one pixel in the first image frame of the generated first video are predicted based on the first second text feature, the first reference posture feature set and the second posture feature set sequence by using any appropriate sampler, to obtain the first candidate posture feature set, and the posture features, each corresponding to at least one pixel in an n^{th} image frame of the generated first video are predicted to obtain the second candidate posture feature set.

[00143] At S252, the predicted first one of the second posture feature sets is determined from a posture feature codebook based on the first candidate posture feature set.

[00144] At S253, the predicted n^{th} one of the second posture feature sets is determined from the posture feature codebook based on the second candidate posture feature set.

[00145] In some implementation modes, the plurality of posture features matched with first candidate posture features in the first candidate posture feature set may be determined from the posture codebook by using any appropriate matching mode, and are determined as the predicted first one of the second posture feature sets. Similarly, the plurality of posture features matched with second candidate posture features in the second candidate posture feature set may be determined from the posture codebook by using any appropriate matching mode, and determined as the predicted n^{th} one of the second posture feature sets. For example, the posture features, nearest to or the most similar to the first/second candidate posture features, in the posture codebook may be determined as the posture features matched with the first/second candidate posture features.

[00146] It is to be understood that, in some implementation modes, the process of predicting each of the second posture feature sets between the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the updated second posture feature set sequence in the above step S222 may refer to the implementation modes of the above S251 to S253.

[00147] In the above embodiment, on the one hand, the posture features of the object are directly predicted based on the second text feature, the first reference posture feature set and the second posture feature set sequence to obtain the second posture feature sets, that is, each of the second posture features in the second posture feature sets is directly predicted in the distribution space of the continuous posture features, a one-to-many mapping prediction problem resulted from feature redundancy in the codebook may be reduced, such that the predicted posture features are more meaningful, thereby improving the quality of the generated video; on the other hand, after the second posture feature sets are obtained by prediction in the distribution space of the continuous posture features, the second posture features corresponding to the second posture features in the second posture feature set are determined from the posture codebook, and therefore, as each of the second posture features in the second posture feature sets is determined from the posture codebook, the rich posture features stored in the posture codebook may be fully utilized, thereby improving the accuracy for prediction of the second posture feature sets and further improving the quality of the generated video. In addition, compared with a solution in which each second index value respectively corresponding to at least one pixel of the image frame in the first video to be generated may be directly predicted, a process of predicting the second posture features in the distribution space of the continuous posture features is easier to converge, thus shortening the prediction time and improving the efficiency of video generation.

[00148] An embodiment of the disclosure provides a method for training a model, which may be executed by a processor of a computer device. The model to be trained may include an appearance codebook, a posture codebook and a decoding network. As illustrated in FIG. 3, the method includes the following steps S301 to S305.

[00149] At S301, a video sample containing a sample object is obtained, and the video sample includes a first image frame and a second image frame.

[00150] Here, the model to be trained may refer to a model for generating a video.

[00151] In some implementation modes, a type of the sample object included in the video sample may be the same as the type of an object included in the video to be generated. For example, in a case that the object included in the video to be generated is a person, the sample object included in the video sample may be a person as well; and in the case that the object included in the video to be generated is a cat, the sample object

included in the video sample may be a cat as well.

[00152] The first image frame and the second image frame may be any two different image frames in the video sample. For example, the first image frame may be the first image frame in the video sample, the second image frame may be an image frame after the first image frame, in the video sample, such as the second image frame, a third image frame and the like.

[00153] At S302, feature extraction is performed on the first image frame by using the appearance codebook to obtain a third appearance feature set, and feature extraction is performed on the second image frame after being augmented by using the posture codebook to obtain third posture feature sets; the appearance codebook includes the plurality of appearance features; the posture codebook includes a plurality of posture features; and augmenting is configured to reduce appearance information in the second image frame.

[00154] Here, the third appearance features respectively corresponding to pixels in the first image frame may be sampled from the appearance codebook according to image information included in the first image frame, to obtain the third appearance feature set.

[00155] Before performing feature extraction on the second image frame by using the posture codebook, augmenting may be performed first on the second image frame to reduce the appearance information of the object in the second image frame. Then, the third appearance features respectively corresponding to the pixels in the second image frame may be sampled from the posture codebook according to the image information included in the second image frame after being augmented, to obtain the third posture feature sets. Therefore, the appearance information in the second image frame leaking to the third posture feature sets may be reduced, thereby facilitating improving the capability of the appearance features in the updated appearance codebook for representating the appearance information and reducing the algorithm complexity for feature extraction of the second image frame. During implementation, augmenting for the second image frame may include, but is not limited to, color dithering processing and/or Gaussian blur processing.

[00156] In some implementation modes, first image features respectively corresponding to the pixels in the first image frame may be extracted from the first image frame, and the appearance features respectively corresponding to the first image features

in the appearance codebook are determined as the third appearance features to obtain the third appearance feature set; second image features respectively corresponding to the pixels in the second image frame are extracted from the second image frame after being augmented, and the posture features respectively corresponding to the second image features in the posture codebook are determined as the third posture features to obtain the third posture feature sets. For example, the appearance features, nearest to or the most similar to the first image features in the appearance codebook may be determined as the appearance features corresponding to the first image features, and the posture features, nearest to or the most similar to the second image features in the posture codebook are determined as the posture features corresponding to the second image features.

[00157] In some implementation modes, first index values respectively corresponding to the pixels in the first image frame may be predicted, the appearance features respectively corresponding to the first index values in the appearance codebook are determined as the third appearance features to obtain the third appearance feature set; second index values respectively corresponding to the pixels in the second image frame after being augmented are predicted, and the posture features respectively corresponding to the second index values in the posture codebook are determined as the third posture features to obtain the third posture feature sets.

[00158] It is to be noted that the appearance codebook and the posture codebook may be not trained or updated after initialization, or may be updated at least once, which are not limited by the embodiment of the disclosure.

[00159] At S303, the third image frame is generated based on the third appearance feature set and the third posture feature sets by using the decoding network.

[00160] Here, the above step S303 corresponds to the step S123 in the above embodiment. During implementation, it may refer to the implementation mode of generating the image frame in S123.

[00161] At S304, a first loss value is determined based on the third image frame and the second image frame before being augmented.

[00162] Here, the first loss value may be determined based on the third image frame and the second image frame before being augmented by using any appropriate loss function, which is not limited by the embodiment of the disclosure. For example, the first loss value may be determined according to at least one of an absolute value loss, atleast

square error loss, a cosine loss and a cross entropy loss between the second image frame before being augmented and the third image frame.

[00163] At S305, the appearance codebook, the posture codebook and network parameters of the decoding network are updated based on the first loss value, to obtain the updated appearance codebook, posture codebook and decoding network.

[00164] Here, it is determined whether to update the appearance codebook, the posture codebook and the network parameters of the decoding network based on the first loss value; in a case of determining to update the appearance codebook, the posture codebook and the network parameters of the decoding network, the appearance codebook, the posture codebook and the network parameters of the decoding network are updated by using an appropriate updating algorithm, and the third image frame is redetermined by using the updated appearance codebook, posture codebook and decoding network; a new first loss value is determined based on the redetermined third image frame and the second image frame before being augmented, and it is determined whether to continue to update the appearance codebook, the posture codebook and the network parameters of the decoding network based on the new first loss value. In the case of determining not to continue to update the appearance codebook, the posture codebook and the network parameters of the decoding network, the appearance codebook, the finally updated posture codebook and decoding network are determined as the trained appearance codebook, posture codebook and decoding network respectively.

[00165] In some implementation modes, the appearance codebook, the posture codebook and the network parameters of the decoding network may be updated in a case that the first loss value does not satisfy a first preset condition; in the case that the loss value satisfies the first preset condition or the number of times for updating the appearance codebook, the posture codebook and the network parameters of the decoding network reaches a set threshold value, the appearance codebook, the posture codebook and the network parameters of the decoding network are stopped to update, and the appearance codebook, the posture codebook and the decoding network which are finally updated are determined as the trained appearance codebook, posture codebook and decoding network respectively. The first preset condition may include, but is not limited to, at least one of the first loss value being smaller than the set loss threshold value and variation convergence of the first loss value.

[00166] The way of updating the appearance codebook, the posture codebook and the network parameters of the decoding network may be determined according to an actual situation, including, but not limited to, a gradient descent method, a Newton momentum method and the like, which is not limited here.

[00167] In the embodiment of the disclosure, in a process of training and updating the appearance codebook, the posture codebook and the network parameters of the decoding network in the model, the appearance codebook is configured to perform feature extraction on the first image frame to obtain the third appearance feature set, the posture codebook is configured to perform feature extraction on the second image frame to obtain the third posture feature set, the decoding network is configured to generate the third image frame based on the third appearance feature set and the third posture feature set, and the first loss value is generated based on the second image frame and the third image frame, and therefore, the appearance codebook, the posture codebook and the network parameters of the decoding network are updated based on the first loss value, thereby improving the capability of the appearance feature in the appearance codebook for representing the appearance information, the capability of the posture features in the posture codebook for representating posture information, and the capability of the decoding network for generation of the image frame based on the appearance feature set and the posture feature set. Thus, the distribution space of the appearance features of the object may be better separated from the distribution space of the posture features through the updated appearance codebook and posture codebook, and the video including the object may be better generated based on the appearance feature set and the posture feature set of the object sampled from the appearance codebook and the posture codebook respectively through the updated decoding network, therefore, the same first appearance feature set is favorably adopted in a process of generating the image frames of the video, so as to reduce drift of the appearance features of the object among the image frames, such that the appearance of the object in the generated video is well kept among the image frames. In addition, the third appearance feature set and the third posture feature set sequence are sampled separately. In this way, on one hand, the posture feature of the object does not need to be considered in an appearance feature sampling process, thereby reducing the complexity for sampling of the appearance features and further facilitating improving the capability of the appearance features in the appearance

codebook for representing the appearance information; on the other hand, the appearance feature of the object does not need to be considered in a posture feature sampling process, thereby simplifying modeling for actions of the object, reducing the complexity for sampling of the posture features and facilitating improving the capability of the posture features in the posture codebook for representing the posture information.

[00168] In some embodiments, performing feature extraction on the first image frame by using the appearance codebook to obtain the third appearance feature set at S302 may include the following operations: at S311, feature extraction is performed on the first image frame by using a first encoding network to obtain a fourth appearance feature set; at S312, the third appearance feature set is sampled from the appearance codebook based on the fourth appearance feature set.

[00169] In some embodiments, performing feature extraction on the first image frame by using the appearance codebook to obtain the third appearance feature set at S302 may include the following operations: at S313, feature extraction is performed on the second image frame after being augmented by using a second encoding network to obtain the fourth appearance feature set; at S314, the third posture feature sets is sampled from the posture codebook based on the fourth appearance feature set.

[00170] The above step S305 may include the following operation: at S315, the appearance codebook, the posture codebook, the network parameters of the decoding network, the network parameters of the first encoding network and the network parameters of the second encoding network are updated based on the first loss value, to obtain the updated appearance codebook, posture codebook, decoding network, first encoding network and second encoding network.

[00171] Here, network structures of the first encoding network and the second encoding network may be determined according to the actual situation. In some implementation modes, both the first encoding network and the second encoding network may include at least one encoder; the encoder may include at least one layer of convolutional neural network. In the case that the encoder includes multiple layers of convolutional neural networks, parameters (such as a dimension, a convolution kernel size and an amount) of the convolutional neural network at different layers may be the same or different. In some implementation modes, the encoder may further include a parameter activation function and the like.

- [00172] During implementation, the first encoding network may be configured to extract the appearance features of the object in the first image frame, to obtain fourth appearance features respectively corresponding to pixels in the first image frame, thereby obtaining the fourth appearance feature set. The second encoding network may be configured to extract the posture features of the object in the second image frame after being augmented to obtain fourth posture features respectively corresponding to pixels in the second image frame, thereby obtaining a fourth posture feature set.
- [00173] For each of the fourth appearance features in the fourth appearance feature set, the appearance feature corresponding to the fourth appearance feature may be sampled from the appearance codebook as one third appearance feature, thereby obtaining the third appearance feature set. For example, the appearance features, nearest to or the most similar to the fourth appearance features, in the appearance codebook may be determined as the third appearance features corresponding to the fourth appearance features.
- [00174] For each of the fourth posture features in the fourth posture feature set, the posture feature corresponding to the fourth posture feature may be sampled from the posture codebook as a third posture feature, thereby obtaining the third posture feature set. For example, the posture features, nearest to or the most similar to the fourth appearance features, in the posture codebook may be determined as the third posture features corresponding to the fourth posture features.
- [00175] The way of updating the network parameters of the first encoding network and the network parameters of the second encoding network may be determined according to the actual situation, including, but not limited to, at least one of the gradient descent method, the Newton momentum method and the like, which is not limited here.
- [00176] In the above embodiment, on the one hand, the capability of the first encoding network for learning the appearance information of the object can be improved, such that the determined third appearance feature set may better represent the appearance information of the object in the first image frame; on the other hand, the capability of the second encoding network for learning the posture information of the object may be improved, such that the determined third posture feature set may better represent the posture information of the object in the second image frame.
- [00177] In some embodiments, the above step S304 may include S321 to S324 below.
- [00178] At S321, a first sub-loss value is determined based on the third image frame

and the second image frame before being augmented.

[00179] Here, the first sub-loss value represents loss between the third image frame and the second image frame before being augmented. During implementation, the first sub-loss value may be determined by using any appropriate loss function, which is not limited by the embodiment of the disclosure. For example, the loss function may include, but is not limited to, at least one of an absolute value loss function, a least square error loss function, a cosine loss function, a cross entropy loss function, etc.

[00180] At S322, a second sub-loss value is determined based on the third appearance feature set and the fourth appearance feature set.

[00181] Here, the second sub-loss value represents the loss between the third appearance feature set and the fourth appearance feature set. During implementation, the second sub-loss value may be determined by using any appropriate loss function, which is not limited by the embodiment of the disclosure.

[00182] In some implementation modes, the second sub-loss value may be determined based on the third appearance feature set and the fourth appearance feature set. Alternatively, a stop gradient operation may be performed first on any of the appearance feature sets in the third appearance feature set and the fourth appearance feature set, and then the second sub-loss value is determined based on the loss between the appearance feature set subjected to the stop gradient operation and the other appearance feature set not subjected to the stop gradient operation.

[00183] At S323, a third sub-loss value is determined based on the third posture feature set and the fourth posture feature set.

[00184] Here, the third sub-loss value represents the loss between the third posture feature set and the fourth posture feature set. During implementation, the third sub-loss value between the third posture feature set and the fourth posture feature set may be determined by using any appropriate loss function, which is not limited by the embodiment of the disclosure.

[00185] In some implementation modes, the third sub-loss value may be determined based on the loss between the third posture feature set and the fourth posture feature set. Alternatively, the stop gradient operation may be performed first on any of the posture feature sets in the third posture feature set and the fourth posture feature set, and then the third sub-loss value is determined based on the loss between the posture feature set

subjected to the stop gradient operation and the other posture feature set not subjected to the stop gradient operation.

[00186] At S324, the first loss value is determined based on the first sub-loss value, the second sub-loss value and the third sub-loss value.

[00187] In some implementation modes, a sum of the sub-loss values may be determined as the first loss value, or different weights are set for the sub-loss values respectively and weighted summation is performed to obtain the first loss value.

[00188] In some implementation modes, the first loss value L_1 may be determined by using formula 3-1:

$$L_1 = \left\| I_k - \hat{I}_k \right\|_1 + \left\| sg(\hat{f}_{a2}) - f_{a2} \right\|_2^2 + \left\| sg(f_{a2}) - \hat{f}_{a2} \right\|_2^2 + \left\| sg(\hat{f}_{p2}) - f_{p2} \right\|_2^2 + \left\| sg(f_{p2}) - \hat{f}_{p2} \right\|_2^2 \quad (3-1)$$

[00189] Where I_k is the second image frame before being augmented before being augmented, $\hat{I}_k$ is the third image frame, $\hat{f}_{a2}$ is the third appearance feature set, f_{a2} is the fourth appearance feature set, $\hat{f}_{p2}$ is the third posture feature set, f_{p2} is the fourth posture feature set, and $sg(-)$ refers to the stop gradient operation.

[00190] In the above embodiment, the loss between the third image frame and the second image frame, the loss between the third appearance feature set and the fourth appearance feature set and the loss between the third posture feature set and the fourth posture feature set are comprehensively considered in the process of determining the first loss value, thereby better facilitating learning of the appearance codebook, the posture codebook, the decoding network, the first encoding network and the second encoding network.

[00191] In some embodiments, the first encoding network includes a first encoder and a second encoder; the second encoding network includes the first encoder and a third encoder.

[00192] The step S311 may include the following operation: at S331, feature extraction is performed on the first image frame by using the first encoder to obtain a fifth appearance feature set; feature extraction is performed on the fifth appearance feature set by using the second encoder to obtain the fourth appearance feature set.

[00193] The step S313 may include the following operation: at S332, feature extraction is performed on the second image frame after being augmented by using the first encoder to obtain fifth posture feature set; feature extraction is performed on the fifth posture feature set by using the third encoder to obtain the fourth posture feature set.

[00194] Here, structures of the first encoder, the second encoder and the third encoder may be determined according to the actual situation and are not limited by the embodiment of the disclosure.

[00195] In some implementation modes, the fourth appearance feature set f_{a2} may be determined by using formula 3-2:

$$f_{a2} = E_a(E_s(I_0)) \quad (3-2)$$

[00196] Where E_s is the first encoder, E_a is the second encoder and I_0 is the first image frame.

[00197] In some implementation modes, the fourth posture feature set f_{p2} may be determined by using formula 3-3:

$$f_{p2} = E_p(E_s(I_k^a)) \quad (3-3)$$

[00198] Where E_s is the first encoder, E_p is the third encoder and I_k^a is the second image frame after being augmented.

[00199] In the above embodiment, extraction of the fifth appearance feature set and the fifth posture feature set is performed by the first encoder, that is, the same encoder is shared in the process of extracting the fifth appearance feature set and the fifth posture feature set, and thus the number of parameters to be learned may be reduced, a calculation amount in a model training process may be reduced and the model training efficiency is improved.

[00200] In some embodiments, a dimension of the posture feature in the posture codebook is lower than the dimension of the appearance feature in the appearance codebook. The decoding network includes a first decoder and a second decoder. The above step S303 may include S341 to S342 below.

[00201] At S341, up-sampling is performed on each posture feature in the third posture feature set by using the first decoder to obtain an up-sampled third posture feature set; the dimension of each of the third posture features in the up-sampled third posture feature set

is the same as the dimension of each of the appearance features in the third appearance feature set.

[00202] At S342, the third image frame is generated based on the third appearance feature set and the up-sampled third posture feature set by using the second decoder.

[00203] Here, as the dimension of the posture features in the posture codebook is lower than the dimension of the appearance features in the appearance codebook, the dimension of each of the third posture features in the third posture feature set sampled from the posture codebook is lower than the dimension of each of the third appearance features in the third appearance feature set sampled from the appearance codebook as well. Therefore, up-sampling may be performed on each posture feature in the third posture feature set by using the first decoder to improve the dimension of each posture feature in the third posture feature set, such that the dimension of each of the third posture features in the up-sampled third posture feature set is the same as the dimension of each of the appearance features in the third appearance feature set, thereby facilitating decoding the third appearance feature set and the up-sampled third posture feature set by the second decoder to generate the third image frame.

[00204] In some implementation modes, the third image frame $\hat{I}_k$ may be generated based on the third appearance feature set $\hat{f}_{a2}$ and the third posture feature set $\hat{f}_{p2}$ by using formula 3-4:

$$\hat{I}_k = D([\hat{f}_{a2}, D_p(\hat{f}_{p2})]) \quad (3-4)$$

[00205] Where D_p is the first decoder, D is the second decoder and $[-]$ refers to a connection operation.

[00206] An embodiment of the disclosure provides a method for training a model, which may be executed by a processor of a computer device. The model to be trained may include the appearance codebook, a posture codebook and a decoding network. The method includes the following steps S401 to S408.

[00207] At S401, a video sample containing a sample object is obtained; the video sample includes a first image frame and a second image frame.

[00208] At S402, feature extraction is performed on the first image frame by using a first encoding network to obtain a fourth appearance feature set.

- [00209] At S403, a third appearance feature set is sampled from the appearance codebook based on the fourth appearance feature set; the appearance codebook includes a plurality of appearance features.
- [00210] At S404, feature extraction is performed on the second image frame after being augmented by using a second encoding network to obtain a fourth posture feature set; and the augmenting is configured to reduce appearance information in the second image frame.
- [00211] At S405, a third posture feature set is sampled from the posture codebook based on the fourth posture feature set; and the posture codebook includes a plurality of posture features.
- [00212] At S406, a third image frame is generated based on the third appearance feature set and the third posture feature set by using the decoding network.
- [00213] At S407, a first loss value is determined based on the third image frame and the second image frame before being augmented.
- [00214] At S408, the appearance codebook, the posture codebook, network parameters of the decoding network, the network parameters of the first encoding network and the network parameters of the second encoding network are updated based on the first loss value to obtain the updated appearance codebook, posture codebook, decoding network, first encoding network and second encoding network.
- [00215] Here, S401, S406 and S407 correspond to S301, S303 and S304 in the above embodiment respectively; S402 to S405 correspond to S311 to S314 in the above embodiment respectively; S408 corresponds to S315 in the above embodiment; during implementation, it may refer to the implementation modes of the above corresponding steps.
- [00216] In some embodiments, the video sample is provided with an appearance description text label of the sample object. The model further includes a text feature extraction network and a first sampling network. After the above S408, the method further includes S411 to S415 below.
- [00217] At S411, feature extraction is performed on the appearance description text label by using the text feature extraction network to obtain a third text feature.
- [00218] Here, the appearance description text label may be a text label automatically labeled for the video sample in advance through a person or a system and configured to

describe appearance of an object included in the video sample.

[00219] At S412, a sixth appearance feature set of the sample object is sampled from the updated appearance codebook based on the third text feature by using the first sampling network, and a second reference posture feature set representing that the object is taking a preset posture is sampled from the updated posture codebook.

[00220] Here, during implementation, S412 may refer to an implementation mode of obtaining a first appearance feature at S121 and an implementation mode of obtaining a first reference posture feature set in S204.

[00221] At S413, a fourth image frame corresponding to the preset posture is generated based on the sixth appearance feature set and the second reference posture feature set by using the updated decoding network.

[00222] Here, S413 corresponds to S123. During implementation, it may refer to the implementation mode of generating one image frame based on a first appearance feature set and a first posture feature set in S123.

[00223] At S414, a second loss value is determined based on the fourth image frame and a fifth image frame corresponding to the preset posture in the video sample.

[00224] Here, the second loss value may be determined based on the fourth image frame and the fifth image frame by using any appropriate loss function, which is not limited by the embodiment of the disclosure. For example, the second loss value may be determined according to at least one of an absolute value loss, a least square error loss, a cosine loss and a cross entropy loss between the fourth image frame and the fifth image frame.

[00225] At S415, the network parameters of the first sampling network are updated based on the second loss value to obtain the updated first sampling network.

[00226] Here, it may be determined whether to update the network parameters of the first sampling network based on the second loss value. In a case of determining to update the network parameters of the first sampling network, the network parameters of the first sampling network are updated by using an appropriate updating algorithm; the fourth image frame is redetermined by using the updated first sampling network; a new second loss value is determined based on the redetermined fourth image frame and the fifth image frame; and it is determined whether to continue to update the network parameters of the first sampling network based on the new second loss value. In the case of

determining not to continue to update the network parameters of the first sampling network, the finally updated first sampling network is determined as the trained first sampling network.

[00227] In some implementation modes, the network parameters of the first sampling network may be updated in a case that the second loss value does not satisfy a second preset condition; in the case that the loss value satisfies the second preset condition or the number of times for updating the network parameters of the first sampling network reaches a set threshold value, the network parameters of the first sampling network are stopped to update, and the finally updated first sampling network is determined as the trained first sampling network. The second preset condition may include, but is not limited to, at least one of the second loss value being smaller than the set loss threshold value, variation convergence of the first loss value, etc.

[00228] The way of updating the network parameters of the first sampling network may be determined according to the actual situation, including, but not limited to, at least one of a gradient descent method, a Newton momentum method and the like, which is not limited here.

[00229] In the above embodiment, after the appearance codebook, the posture codebook and the decoding network are trained and updated, the network parameters of the first sampling network may be trained and updated by using the updated appearance codebook, posture codebook and decoding network, thereby reducing the training workload; further, the first sampling network is trained by taking the updated appearance codebook, posture codebook and decoding network as a condition, thereby facilitating the first sampling network in better learning the capability for representing the appearance features of the object.

[00230] In some embodiments, the video sample is further provided with an action description text label of the sample object. The model further includes a second sampling network. The method further includes the following steps S421 to S426.

[00231] At S421, feature extraction is performed on the action description text label by using the text feature extraction network to obtain a fourth text feature.

[00232] Here, the action description text label may be a text label labeled by the person or the system automatically for the video sample in advance and configured to describe an action of the object included in the video sample.

- [00233] At S422, a third posture feature set sequence of the sample object is sampled from the updated posture codebook based on the fourth text feature by using the second sampling network.
- [00234] At S423, a seventh appearance feature set of the sample object is sampled from the updated appearance codebook based on the third text feature by using the updated first sampling network.
- [00235] At S424, a second video is generated based on the seventh appearance feature set and the third posture feature set sequence by using the updated decoding network.
- [00236] Here, steps S422 to S424 respectively correspond to steps S122, S121 to S123 in the above embodiment. During implementation, it may refer to the implementation modes of steps S122, S121 and S123.
- [00237] At S425, a third loss value is determined based on the second video and the sample video.
- [00238] Here, the third loss value may be determined based on the second video and the sample video by using any appropriate loss function, which is not limited by the embodiment of the disclosure. For example, the third loss value may be determined according to at least one of the absolute value loss, the least square error loss, the cosine loss and the cross entropy loss between the second video and the sample video.
- [00239] At S426, the network parameters of the second sampling network are updated based on the third loss value to obtain the updated second sampling network.
- [00240] Here, it may be determined whether to update the network parameters of the second sampling network based on the third loss value; in the case of determining to update the network parameters of the second sampling network, the network parameters of the second sampling network are updated by using an appropriate updating algorithm; the second video is redetermined by using the updated second sampling network; a new third loss value is determined based on the redetermined second video and the sample video; and it is determined whether to continue to update the network parameters of the second sampling network based on the new third loss value. In the case of determining not to continue to update the network parameters of the second sampling network, the finally updated second sampling network is determined as the trained second sampling network.
- [00241] In some implementation modes, the network parameters of the second

sampling network may be updated in a case that the third loss value does not satisfy a third preset condition; in the case that the loss value satisfies the third preset condition or the number of times for updating the network parameters of the second sampling network reaches the set threshold value, the network parameters of the second sampling network are stopped to update, and the finally updated second sampling network is determined as the trained second sampling network. The third preset condition may include, but is not limited to, at least one of the third loss value being smaller than the set loss threshold value, variation convergence of the third loss value, etc. During implementation, the third preset condition may be determined according to the actual situation and is not limited by the embodiment of the disclosure.

[00242] The way of updating the network parameters of the second sampling network may be determined according to the actual situation, including, but not limited to, at least one of the gradient descent method, the Newton momentum method and the like, which is not limited here.

[00243] In the above embodiments, after the appearance codebook, the posture codebook, the decoding network and the first sampling network are trained and updated, the network parameters of the second sampling network may be trained and updated by using the updated appearance codebook, posture codebook, decoding network and first sampling network, thereby reducing the training workload; further, the second sampling network is trained by taking the updated appearance codebook, posture codebook, decoding network and first sampling network as the condition, thereby facilitating the second sampling network in better learning the capability for representing the posture features of the object.

[00244] In some embodiments, the method further includes the following step S431.

[00245] At S431, the third reference posture feature set representing that the object is taking the preset posture is sampled from the updated posture codebook based on the third text feature by using the updated first sampling network.

[00246] Here, step S431 corresponds to the above S204. During implementation, it may refer to the implementation mode of obtaining the first reference posture feature set by sampling in the above S204.

[00247] The above step S422 may include the following step S432.

[00248] At S432, the third posture feature set sequence of the sample object is sampled

from the updated posture codebook based on the fourth text feature and the third reference posture feature set by using the second sampling network.

[00249] Here, step S432 corresponds to the above step S206. During implementation, it may refer to the implementation mode of obtaining the first reference posture feature set by sampling in the above S206.

[00250] In some embodiments, step S432 may include steps S441 to S442 below.

[00251] At S441, a fourth posture feature set sequence is determined; and at least one fourth posture feature set in the fourth posture feature set sequence is masked.

[00252] Here, the fourth posture feature set sequence may be set in advance or determined randomly, which is not limited by the embodiment of the disclosure.

[00253] During implementation, a plurality of fourth posture feature sets in the fourth posture feature set sequence may be masked. For example, part or all of the fourth posture feature sets in the fourth posture feature set sequence may be masked. For another example, the plurality of fourth posture feature sets in the middle, in the beginning or in the end of the fourth posture feature set sequence may be masked.

[00254] It is to be understood that the number of the fourth posture feature sets in the fourth posture feature set sequence is the same as the number of the third reference posture features in the third reference posture feature set and the number of the third posture feature sets in the third posture feature set sequence.

[00255] In some implementation modes, in the case that a first one of the fourth posture feature sets and/or a last one of the fourth posture feature sets in the fourth posture feature set sequence are/is unmasked, before the fourth posture feature set sequence is input to the second sampling network, part of the fourth posture features in the first one of the fourth posture feature sets and/or the last one of the fourth posture feature sets are masked, so that the masked fourth posture features may be predicted by using the second sampling network, thereby improving the capability of the second sampling network for predicting posture features.

[00256] At S442, the third posture feature set sequence of the sample object is sampled from the updated posture codebook based on the fourth text feature, the third reference posture feature set and the fourth posture feature set sequence by using the second sampling network.

[00257] Here, step S442 corresponds to step S212 in the above embodiment. During

implementation, it may refer to the implementation mode of the above step S212.

[00258] In the above embodiment, the third posture feature set sequence of the sample object is sampled from the updated posture codebook based on the fourth text feature, the third reference posture feature set and the fourth posture feature set sequence by using the second sampling network. As at least one fourth posture feature set in the initial fourth posture feature set sequence is masked, the capability of the second sampling network for predicting the masked fourth posture feature set in the fourth posture feature set sequence is beneficially improved.

[00259] In some embodiments, the above step S442 may include the following steps S451 to S452.

[00260] At S451, each of the masked fourth posture feature sets in the fourth posture feature set sequence is predicted based on the fourth text feature, the third reference posture feature set and the fourth posture feature set sequence by using the second sampling network to obtain a predicted fourth posture feature set sequence.

[00261] At S452, the third posture feature set sequence is sampled from the updated posture codebook based on the predicted fourth posture feature set sequence.

[00262] Here, the above steps S451 to S452 may refer to the implementation mode of determining a first one of the second posture feature sets in the above steps S251 to S252.

[00263] In some implementation modes, the predicted fourth posture feature set sequence $\bar{F}_p$ may be obtained by using formula 4-1:

$$\bar{F}_p = S_{\theta} \left(\left[T(t_m), \hat{f}_{p3}^0, F_{p3}^n \right] \right) \quad (4-1)$$

[00264] Where t_m is the action description text label, $T(t_m)$ is the fourth text feature, $\hat{f}_{p3}^0$ is the third reference posture feature set, F_{p3}^n is an unpredicted fourth posture feature set sequence, $[-]$ refers to a connection operation and S_{θ} is the second sampling network.

[00265] In some embodiments, the above step S425 may include the following steps S461 to S463.

[00266] At S461, a fourth sub-loss value is determined based on the second video and the sample video.

- [00267] At S462, a fifth sub-loss value is determined based on the third posture feature set sequence and the fourth posture feature set sequence.
- [00268] Here, the fourth sub-loss value represents the loss between the second video and the sample video. The fifth sub-loss value represents the loss between the third posture feature set sequence and the fourth posture feature set sequence.
- [00269] During implementation, the fourth sub-loss value and the fifth sub-loss value may be determined by using any appropriate loss function, for example, the loss function may include, but is not limited to, at least one of an absolute value loss function, a least square error loss function, a cosine loss function, a cross entropy loss function, a reconstruction loss function, etc.
- [00270] In some implementation modes, the fifth sub-loss value may be determined according to the loss between the third posture feature set sequence and the fourth posture feature set sequence. Alternatively, a stop gradient operation may be performed first on one of the third posture feature set sequence and the fourth posture feature set sequence, and then the fifth sub-loss value is determined based on the loss between the posture feature set sequence subjected to the stop gradient operation and another posture feature set sequence not subjected to the stop gradient operation.
- [00271] At S463, the third loss value is determined based on the fourth sub-loss value and the fifth sub-loss value.
- [00272] In some implementation modes, the sum of the fourth sub-loss value and the fifth sub-loss value may be determined as the third loss value. Alternatively, different weights are set for the fourth sub-loss value and the fifth sub-loss value respectively, and weighted summation is performed to obtain the third loss value.
- [00273] In the above embodiment, the loss between the second video and the sample video and the loss between the third posture feature set sequence and the fourth posture feature set sequence are comprehensively considered in the process of determining the third loss value, thereby better facilitating learning of the second sampling network.
- [00274] In some embodiments, the method further includes steps S471 to S472 below.
- [00275] At S471, feature extraction is performed on each of the image frames in the sample video by using the second sampling network to obtain a fifth posture feature set sequence.
- [00276] Here, the fifth posture feature set sequence may be obtained with reference to

the way of performing feature extraction on the second image frame after being augmented by using the posture codebook in the above S302 to obtain the third posture feature sets.

[00277] At S472, a sixth sub-loss value is determined based on the fifth posture feature set sequence and the fourth posture feature set sequence.

[00278] Here, the sixth sub-loss value represents the loss between the fifth posture feature set sequence and the fourth posture feature set sequence.

[00279] The above step S463 may include: at S473, the third loss value is determined based on the fourth sub-loss value, the fifth sub-loss value and the sixth sub-loss value.

[00280] In some implementation modes, the sum of the fourth sub-loss value, the fifth sub-loss value and the sixth sub-loss value may be determined as the third loss value. Alternatively, different weights may be set for the fourth sub-loss value, the fifth sub-loss value and the sixth sub-loss value respectively, and weighted summation is performed to obtain the third loss value.

[00281] In some implementation modes, the third loss value L_3 may be determined by using formula 4-2:

$$L_3 = \|\bar{F}_p - F_p\|_1 + \|sg(\hat{F}_p) - \bar{F}_p\|_2^2 + \|sg(\bar{F}_p) - \hat{F}_p\|_2^2 + L_{rec}(\hat{V}_1, V_1) \quad (4-2)$$

[00282] Where $\hat{V}_1$ is the second video, V_1 is the sample video, $\hat{F}_p$ is the third posture feature set sequence, $\bar{F}_p$ is the predicted fourth posture feature set sequence, F_p is the fifth posture feature set sequence, $sg(-)$ refers to the gradient stop operation, $L_{rec}(-)$ refers to the reconstruction loss function which may be composed of an L1 loss function and a perception loss function.

[00283] In the above embodiment, the loss between the second video and the sample video, the loss between the third posture feature set sequence and the fourth posture feature set sequence and the loss between the fifth posture feature set sequence and the fourth posture feature set sequence are comprehensively considered in the process of determining the third loss value, thereby better facilitating learning of the second sampling network.

[00284] Applications of the method for generating a video and the method for training a model provided by the embodiments of the disclosure in practical scenes are described

below.

[00285] Taking a text-driven human body video generation task as an example, a human body video (corresponding to the first video) is generated according to the appearance description text and the action description text of a synthesized person (that is, the objected to be included in the first video to be generated). Compared with the general text-driven video generation, generation of the human body video requires maintaining appearance of a human body when the synthesized person is executing a complex action. Therefore, generation of a text-driven human video brings multiple unique challenges: (1) a human body structure is articulated, movements of joints of different body parts form many out-of-plane movements, such as, rotation, which are complex; and (2) when executing the complex action, the appearance of the synthesized person among the image frames is to be kept consistent, for example, the appearance of the synthesized person after back turning is to be kept consistent with the appearance of the synthesized person before back turning.

[00286] In view of this, an embodiment of the disclosure provides a text-driven method for generating a video and method for training a model, aiming to generate a human body video with high fidelity and improve consistency of appearance of a synthesized person among image frames in a generated video. In the method for generating a video, the human body video with high quality may be synthesized according to an appearance description text and an action description text input by a user by using a deep learning algorithm.

[00287] The embodiment of the disclosure at least includes the following advantages.

[00288] 1) Image features with a unified code in the related technology are decomposed into an appearance feature for representing human body appearance and a posture feature for representing a human body posture in an unsupervised way, the appearance feature and the posture feature are stored in an appearance codebook and a posture codebook respectively, and the distribution space of the appearance feature of the object may be separated from the distribution space of the posture feature of the object, and therefore, the consistency of the human body appearance among the image frames in the generated video may be improved and motion modeling may be simplified.

[00289] 2) The continuous posture features are predicted by using a non-causal transformer, the posture features nearest to the continuous posture features are

determined from a codebook with postures and configured to synthesize the image frames of the video, thereby better performing motion modeling of a human body and facilitating improving continuity of the image frames in the generated video in time sequence.

[00290] 3) by using a posture feature mask solution with a non-casual transformer in combination with motion perception, the video with high quality and sequential continuity is favorably generated.

[00291] FIG. 4A illustrates an implementation flowchart of a method for generating a video according to an embodiment of the disclosure. As illustrated in FIG. 4A, the method for generating a video includes S501 to S507 below.

[00292] At S501, an appearance description text and an action description text of the synthesized person are obtained.

[00293] At S502, feature extraction is performed on the appearance description text to obtain a first text feature, and feature extraction is performed on the action description text by using a text feature extraction network to obtain a second text feature.

[00294] At S503, the appearance features of the synthesized person and the posture features corresponding to a preset posture are predicted based on the first text feature to obtain a second appearance feature set and a fourth reference posture feature set.

[00295] At S504, a first appearance feature set is determined from an appearance codebook based on the second appearance feature set, and a first reference posture feature set is determined from a posture codebook based on the fourth reference posture feature set.

[00296] At S505, a second posture feature set sequence is determined, and at least one second posture feature set in the second posture feature set sequence is masked.

[00297] At S506, feature sampling is performed gradually, in the posture codebook, on the masked second posture feature sets in the second posture feature set sequence from two ends to the middle based on the second text feature, the first reference posture feature set and the second posture feature set sequence, until each of the second posture feature sets in the second posture feature set sequence is obtained, and the sampled second posture feature set sequence is used as a first posture feature set sequence.

[00298] At S507, the video containing the synthesized person is generated based on the first appearance feature set and the first posture feature set sequence.

[00299] On this basis, an embodiment of the disclosure provides a video generating model, which may be configured to implement the above method for generating a video. FIG. 4B illustrates a schematic diagram of a composition structure of a video generating model according to an embodiment of the disclosure. As illustrated in FIG. 4B, the video generating model includes a text feature extraction network 51, a first sampling network 52, a second sampling network 53, a decoding network 54, an appearance codebook C_a and a posture codebook C_p .

[00300] The text feature extraction network 51 is configured to: perform feature extraction on an appearance description text t_a by using a first text feature extractor T1 to obtain a first text feature e_a , and perform feature extraction on an action description text t_m by using a second text feature extractor T2 to obtain a second text feature.

[00301] The first sampling network 52 is configured to: predict the appearance features of a synthesized person and the posture features corresponding to a preset posture based on the first text feature e_a by using a first sampler S1 to obtain a second appearance feature set f_{a1} and a fourth reference posture feature set f_p^0 , determine a first appearance feature set $\hat{f}_{a1}$ from the appearance codebook C_a based on the second appearance feature set f_{a1} , and determine a first reference posture feature set $\hat{f}_p^0$ from the posture codebook C_p based on the fourth reference posture feature set f_p^0 .

[00302] The second sampling network 53 is configured to: predict a first one of the second posture feature sets and a last one of the second posture feature sets in a second posture feature set sequence F_p^n in which all of the second posture feature sets are masked according to the second text feature e_m and the first reference posture feature set $\hat{f}_p^0$ by using a non-casual transformer (NCT), to obtain a first candidate posture feature set $\bar{f}_p^1$ and a second candidate posture feature set $\bar{f}_p^n$ respectively; retrieve the posture features corresponding to each of first candidate posture features in the first candidate posture feature set $\bar{f}_p^1$ in the posture codebook C_p to obtain a predicted first one of the second posture feature sets $\hat{f}_p^1$, and retrieve the posture features corresponding to each of

second candidate posture features $\bar{f}_p^n$ in the second candidate posture feature set in the posture codebook C_p to obtain a predicted last one of the second posture feature sets $\hat{f}_p^n$; update the predicted first one $\hat{f}_p^1$ of the second posture feature sets and the predicted last one $\hat{f}_p^n$ of the second posture feature sets to the second posture feature set sequence $\mathbf{F}_p^n$; respectively predict a second one of the second posture feature sets and a last but one of the second posture feature sets in the second posture feature set sequence $\mathbf{F}_p^n$ according to the second text feature e_m , the first reference posture feature set $\hat{f}_p^0$ and the updated second posture feature set sequence $\mathbf{F}_p^n$ by using a Non-Casual Transformer (NCT), to obtain a third candidate posture feature set $\bar{f}_p^2$ and a fourth candidate posture feature set $\bar{f}_p^{n-1}$; retrieve the posture features corresponding to each of third candidate posture features in the third candidate posture feature set $\bar{f}_p^2$ in the posture codebook C_p , to obtain a predicted second one $\hat{f}_p^2$ of the second posture feature sets, and retrieve the posture features corresponding to each of fourth candidate posture features in the fourth candidate posture feature set $\bar{f}_p^{n-1}$ in the posture codebook C_p to obtain the predicted last but one $\hat{f}_p^{n-1}$ of the second posture feature sets; and update the predicted second one $\hat{f}_p^2$ of the second posture feature sets and last but one $\hat{f}_p^{n-1}$ of the second posture feature sets to the second posture feature set sequence; the process is repeated in this way until each of the second posture feature sets in the second posture feature set sequence is obtained by prediction, and the second posture feature set sequence which is finally obtained is used as the first posture feature set sequence $\hat{\mathbf{F}}_p^n$.

[00303] The decoding network 54 is configured to: perform up-sampling on first posture features in each of the first posture feature sets in the first posture feature set sequence $\hat{\mathbf{F}}_p^n$ by using a first decoder D_p to obtain an up-sampled first posture feature set sequence, and generate a video $\hat{\mathbf{V}}$ containing the synthesized person based on the first appearance feature set $\hat{f}_{a1}$ and the up-sampled first posture feature set sequence by using

a second decoder D .

[00304] It is to be noted that, during implementation, the first text feature extractor T1 and the second text feature extractor T2 may be the same text feature extractor, or different text feature extractors.

[00305] An embodiment of the disclosure provides a method for training a model, which may be configured to train the above video generating model. The method for training a model includes the following three training stages.

[00306] 1) An appearance codebook, a posture codebook and a decoding network are trained.

[00307] FIG. 4C illustrates a flowchart of training an appearance codebook, a posture codebook and a decoding network provided by the embodiment of the disclosure. As illustrated in FIG. 4C, a first image frame I_0 and an second image frame after being augmented I_k^a in a sample video may be respectively used, as an identity frame and a posture frame, to provide appearance information and posture information; the first image frame I_0 and the second image frame after being augmented I_k^a are respectively input to a first encoding network 61 and a second encoding network 62 to obtain a fourth appearance feature set f_{a2} and a fourth posture feature set f_{p2} ; the first encoding network 61 includes a first encoder E_s and a second encoder E_a ; the second encoding network 62 includes the first encoder E_s and a third encoder E_p ; that is, the first encoding network 61 and the second encoding network 62 share the first encoder E_s ; a third appearance feature set $\hat{f}_{a2}$ is sampled from the appearance codebook C_a based on the fourth appearance feature set f_{a2} ; a third posture feature set $\hat{f}_{p2}$ is sampled from the posture codebook C_p based on the fourth posture feature set f_{p2} ; up-sampling is performed on each third posture features in the third posture feature set $\hat{f}_{p2}$ by using a first decoder D_p in a decoding network 54 to obtain the up-sampled third posture feature set; the up-sampled third posture feature set and the third appearance feature set $\hat{f}_{a2}$ are input to a second decoder D to obtain a third image frame $\hat{I}_k$; after the third image

frame $\hat{I}_k$ is generated, a first loss value may be determined based on the third image frame $\hat{I}_k$, an second image frame before being augmented, the third appearance feature set $\hat{f}_{a2}$, the fourth appearance feature set f_{a2} , the third posture feature set $\hat{f}_{p2}$ and the fourth posture feature set f_{p2} by using the above formula 3-1; and the appearance codebook C_a , the posture codebook C_p , network parameters of the first decoder D_p and the second decoder D in the decoding network 54, the network parameters of the first encoding network 61 and the network parameters of the second encoding network 62 are updated according to the first loss value.

[00308] 2) A first sampling network is trained by using the trained and updated appearance codebook, posture codebook and decoding network.

[00309] FIG. 4D illustrates a flowchart of training a first sampling network provided by the embodiment of the disclosure. As illustrated in FIG. 4D, feature extraction may be performed on an appearance description text label t'_a by using a first text feature extractor T1 in a text feature extraction network 51 to obtain a third text feature e'_a ; by using a first sampler S1 in a first sampling network 52, a sixth appearance feature set $\hat{f}_{a4}$ of a sample human body is sampled from the updated appearance codebook C_a based on the third text feature e'_a , and a second reference posture feature set $\hat{f}_{p4}^0$ representing that the sample human body is taking a preset posture is sampled from the updated posture codebook C_p ; a fourth image frame $\hat{I}_0$ corresponding to the preset posture is generated based on the sixth appearance feature set $\hat{f}_{a4}$ and the second reference posture feature set $\hat{f}_{p4}^0$ by using an updated decoding network 54; a second loss value is determined based on the fourth image frame $\hat{I}_0$ and a fifth image frame I_0 corresponding to the preset posture in a video sample; and the network parameters of the first sampler S1 in the first sampling network 52 are updated based on the second loss value to obtain the updated first sampling network 52.

[00310] 3) A second sampling network is trained by using the trained and updated appearance codebook, posture codebook, decoding network and first sampling network.

[00311] FIG. 4E illustrates a flowchart of training a second sampling network according to an embodiment of the disclosure. As illustrated in FIG. 4E, feature extraction may be performed on the appearance description text label t'_a by using the first text feature extractor T1 in the text feature extraction network 51 to obtain the third text feature e'_a ; a seventh appearance feature set $\hat{f}_{a5}$ of the sample human body may be sampled from the updated appearance codebook C_a based on the third text feature e'_a by using the first sampler S1 in the first sampling network 52; a third reference posture feature set $\hat{f}_{p3}^0$ representing that the sample human body is taking the preset posture is sampled from the updated posture codebook C_p based on the third text feature e'_a by using the first sampler S1 in the first sampling network 52; feature extraction is performed on an action description text label t'_m by using a second text feature extractor T2 in the text feature extraction network 51 to obtain a fourth text feature e'_m ; each of the masked fourth posture feature sets in a fourth posture feature set sequence F_{p3}^n is predicted based on the fourth text feature e'_m , the third reference posture feature set $\hat{f}_{p3}^0$ and the fourth posture feature set sequence F_{p3}^n with at least one masked fourth posture feature set by using a Non-Causal Transformer (NCT) in the second sampling network 53 to obtain the predicted fourth posture feature set sequence $\bar{F}_p$; a third posture feature set sequence $\hat{F}_p$ is sampled from the updated posture codebook C_p based on the predicted fourth posture feature set sequence $\bar{F}_p$; a second video $\hat{V}_1$ is generated based on the seventh appearance feature set $\hat{f}_{a5}$ and the third posture feature set sequence $\hat{F}_p$ by using the updated decoding network 54; feature extraction is performed on each of the image frames in the sample video V_1 by using the updated second encoding network 62 to obtain a fifth posture feature set sequence F_p (including a fifth posture feature set $f_p^1, \dots, f_p^n$); a third loss value is determined based on the second video $\hat{V}_1$, the sample video V_1 , the third posture feature set sequence $\hat{F}_p$, the predicted fourth posture feature

set sequence $\bar{F}_p$ and the fifth posture feature set sequence F_p by using the above formula 4-2; and the network parameters of the NCT in the second sampling network 53 are updated based on the third loss value to obtain the updated second sampling network 53.

[00312] In order to improve fidelity of the generated video, the sampled third posture feature set sequence should be reasonable in time and space. In order to enable that the continuous NCT learns an action description text correctly and generates the reasonable human body posture for each image frame, the embodiment of the disclosure provides a novel motion-perception-based masking strategy, which carefully masks the fourth posture feature sets in the fourth posture feature set sequence in time and space. In a time dimension, if a first one of the fourth posture feature sets and a last one of the fourth posture feature sets in the fourth posture feature set sequence are shielded by a mask, all of the fourth posture feature sets in the fourth posture feature set sequence are shielded to prevent the fourth posture feature sets in the middle from providing information to help predict the fourth posture feature sets at two ends. In addition, a probability of shielding all of the fourth posture feature sets in the fourth posture feature set sequence is high, thereby assisting the NCT in better learning prediction of the fourth posture feature sets at the two ends with greater challenges. In a space dimension, at least one prediction may be performed on the first one of the fourth posture feature sets and the last one of the fourth posture feature sets in a process of predicting the fourth posture feature set sequence to obtain all of fourth posture features in the first one of the fourth posture feature sets and all of the fourth posture features in the last one of the fourth posture feature sets, thereby guiding the non-casual transformer to perform prediction according to contextual information in an image frame (that is, part of the sampled fourth posture features corresponding to part of pixels in the image frame), thereby reducing a situation of incompleteness of a human body structure in the generated video resulted from incompleteness of the sampled fourth posture features in the fourth posture feature sets, and better maintaining consistency of the generated image frames in space. Correspondingly, in a training process, part of the fourth posture features in the first one of the fourth posture feature sets and/or part of the fourth posture features in the last one of the fourth posture feature sets in the fourth posture feature set sequence may be masked in the space dimension. Thus, through the motion-perception-based masking

strategy, correctness of motion of the human body in the generated video in the time dimension and completeness of the structure of the human body in the space dimension may be improved.

[00313] FIG. 4F illustrates a flowchart of an image frame of a human body video generated by a method for generating a video according to an embodiment of the disclosure. As illustrated in FIG. 4F, input text descriptions include an appearance description text t_a and an action description text t_m , a first video may be generated according to the appearance description text t_a and the action description text t_m , the appearance of a person in the image frames from $\hat{I}_0$ to $\hat{I}_4$ is consistent, the human body posture in each image frame has good continuity in time sequence, and the human body structure in each image frame is complete.

[00314] The method for generating a video provided by the embodiment of the disclosure may be applied to any appropriate task scene for a presentation object (such as a person, a cat, a dog and/or a tree), for example, a commodity display scene and/or a commodity customization scene in an e-commerce platform, a character presentation scene in a movie, etc. According to the method for generating a video provided by the embodiment of the disclosure, a user describes appearance and an action of the object by inputting a passage to generate the video containing the object executing the action.

[00315] In addition, the method for generating a video provided by the embodiment of the disclosure may be further applied to video software oriented to a non-professional user and is configured to support the non-professional user to perform video creation. For example, the non-professional user may design a virtual image of himself or herself in a metaverse, and then drive the appearance of the virtual image to update and/or move through the text, for example, the virtual image is driven to dance with music through the text, the virtual image is driven to re-dress through the text and the like.

[00316] FIG. 5 illustrates a schematic diagram of a composition structure of an apparatus for generating a video according to an embodiment of the disclosure. As illustrated in FIG. 5, the apparatus for generating a video includes a first obtaining module 610, a first sampling module 620, a second sampling module 630 and a first generation module 640.

[00317] The first obtaining module 610 is configured to obtain an appearance

description text and an action description text of an object.

[00318] The first sampling module 620 is configured to sample a first appearance feature set of the object from an appearance codebook based on the appearance description text; the appearance codebook includes a plurality of appearance features; the first appearance feature set includes first appearance features, each corresponding to at least one pixel of an image frame in a first video to be generated.

[00319] The second sampling module 630 is configured to sample a first posture feature set sequence of the object from a posture codebook based on the action description text; the posture codebook includes a plurality of posture features; the first posture feature set sequence includes first posture feature sets, each corresponding to at least one image frame in the first video to be generated; and each of the first posture feature sets includes first posture features, each corresponding to at least one pixel in the corresponding image frame.

[00320] The first generation module 640 is configured to generate the first video containing the object based on the first appearance feature set and the first posture feature set sequence.

[00321] In some embodiments, the first sampling module is further configured to: perform feature extraction on the appearance description text to obtain a first text feature, and sample the first appearance feature set from the appearance codebook based on the first text feature; the second sampling module is configured to perform feature extraction on the action description text to obtain a second text feature, and sample the first posture feature set sequence from the posture codebook based on the second text feature.

[00322] In some embodiments, the first sampling module is further configured to: predict the appearance feature of the object based on the first text feature to obtain the second appearance feature set, and determine the first appearance feature set from the appearance codebook based on the second appearance feature set.

[00323] In some embodiments, the first sampling module is further configured to: for each of second appearance features in the second appearance feature set, select first appearance features matched with the second appearance features from the appearance codebook, and obtain the first appearance feature set based on each of the first appearance features.

[00324] In some embodiments, the first sampling module is further configured to

sample a first reference posture feature set representing that the object is taking a preset posture from the posture codebook based on the first text feature; the second sampling module is further configured to sample a first posture feature set sequence from the posture codebook based on the second text feature and the first reference posture feature set.

[00325] In some embodiments, the second sampling module is further configured to determine a second posture feature set sequence; at least one second posture feature set in the second posture feature set sequence is masked; perform at least one feature sampling in the posture codebook based on the second text feature, the first reference posture feature sets and the second posture feature set sequence until each of the second posture feature sets in the second posture feature set sequence is obtained, and use the sampled second posture feature set sequence as the first posture feature set sequence.

[00326] In some embodiments, all of the second posture feature sets in the second posture feature set sequence are masked. The second sampling module is further configured to: predict a first one of the second posture feature sets and an n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature sets and the second posture feature set sequence, to obtain the predicted first one of the second posture feature sets and the predicted n^{th} one of the second posture feature sets, and update the second posture feature set sequence with the predicted first one of the second posture feature sets and the predicted n^{th} one of the second posture feature sets, where n is a number of the second posture feature sets in the second posture feature set sequence; and predict each of the second posture feature sets between the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature sets and the updated second posture feature set sequence, and update the second posture feature set sequence with each of the predicted second posture feature sets.

[00327] In some embodiments, the second sampling module is further configured to perform at least one of the following operations: in a case that a predicted $(i-1)^{\text{th}}$ one of the second posture feature sets and a predicted $(n-i+2)^{\text{th}}$ one of the second posture features are updated to the second posture feature set sequence and $n-i+1$ is greater than i ,

predicting an i th one of the second posture feature sets and an $(n-i+1)$ th one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the reference posture feature set and the updated second posture feature set sequence, and updating the second posture feature set sequence with the predicted i th one of the second posture feature sets and the predicted $(n-i+1)$ th one of the second posture feature sets; i is an integer of greater than 1 and $2i$ is no more than $n+1$; in the case that the predicted $(i-1)$ th one of the second posture feature sets and the $(n-i+2)$ th one of the second posture feature sets are updated to the second posture feature set sequence and $n-i+1$ is equal to i , predicting the i th one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature sets and the updated second posture feature set sequence, and updating the second posture feature set sequence with the predicted i th one of the second posture feature sets.

[00328] In some embodiments, the second sampling module is further configured to: perform at least one prediction on the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature sets and the second posture feature set sequence until each of the second posture features in the first one of the second posture feature sets and each of the second posture features in the n^{th} one of the second posture feature sets are obtained; and update the second posture feature set sequence with the first one of the second posture feature sets and the n^{th} one of the second posture feature sets.

[00329] In some embodiments, the second sampling module is further configured to: predict the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence based on the second text feature, the first reference posture feature sets and the second posture feature set sequence to obtain a first candidate posture feature set and a second candidate posture feature set; determine the predicted first one of the second posture feature sets from a posture feature codebook based on the first candidate posture feature set; and determine the predicted n^{th} one of the second posture feature sets from the posture feature codebook based on the second candidate posture feature set.

[00330] In some embodiments, the first video includes a plurality of image frames

respectively corresponding to the first posture feature sets. The first sampling module is further configured to sample the first appearance feature set of the object from the appearance codebook based on the first text feature by using a first sampling network. The second sampling module is further configured to sample the first posture feature set sequence of the object from the posture codebook based on the second text feature by using a second sampling network. The first generation module is further configured to, for each of the first posture feature sets in the first posture feature set sequence, generate one image frame corresponding to the first posture feature set in the first video based on the first appearance feature set and the first posture feature set by using a decoding network.

[00331] In some embodiments, the dimension of the posture features in the posture codebook is lower than the dimension of the appearance features in the appearance codebook.

[00332] FIG. 6 illustrates a schematic diagram of a composition structure of an apparatus for training a model according to an embodiment of the disclosure. The model includes an appearance codebook, a posture codebook and a decoding network. As illustrated in FIG. 6, the apparatus 700 for training a model includes a second obtaining module 710, a first extraction module 720, a second generation module 730, a first determination module 740 and a first updating module 750.

[00333] The second obtaining module 710 is configured to obtain a video sample containing a sample object, and the video sample includes a first image frame and a second image frame.

[00334] The first extraction module 720 is configured to perform feature extraction on the first image frame by using the appearance codebook to obtain a third appearance feature set, and perform feature extraction on the second image frame after being augmented, by using the posture codebook to obtain a third posture feature set; the appearance codebook includes a plurality of appearance features; the posture codebook includes a plurality of posture features; and the augmenting is configured to reduce appearance information in the second image frame.

[00335] The second generation module 730 is configured to generate a third image frame based on the third appearance feature set and the third posture feature set by using the decoding network.

- [00336] The first determination module 740 is configured to determine a first loss value based on the third image frame and the second image frame before being augmented.
- [00337] The first updating module 750 is configured to update the appearance codebook, the posture codebook and network parameters of the decoding network based on the first loss value to obtain the updated appearance codebook, posture codebook and decoding network.
- [00338] In some embodiments, the first extraction module is further configured to: perform feature extraction on the first image frame by using a first encoding network to obtain a fourth appearance feature set, sample the third appearance feature set from the appearance codebook based on the fourth appearance feature set, perform feature extraction on the second image frame after being augmented by using a second encoding network to obtain a fourth posture feature set, and sample a third posture feature set from the posture codebook based on the fourth posture feature set. The first updating module is further configured to update the appearance codebook, the posture codebook, network parameters of the decoding network, network parameters of the first encoding network and network parameters of the second encoding network based on the first loss value, to obtain the updated appearance codebook, posture codebook, decoding network, first encoding network and second encoding network.
- [00339] In some embodiments, the first determination module is configured to: determine a first sub-loss value based on the third image frame and the second image frame; determine a second sub-loss value based on the third appearance feature set and the fourth appearance feature set; determine a third sub-loss value based on the third posture feature set and the fourth posture feature set; and determine the first loss value based on the first sub-loss value, the second sub-loss value and the third third sub-loss value.
- [00340] In some embodiments, the first encoding network includes a first encoder and a second encoder; the second encoding network includes the first encoder and a third encoder. The first extraction module is further configured to: perform feature extraction on the first image frame by using the first encoder to obtain a fifth appearance feature set, perform feature extraction on the fifth appearance feature set by using the second encoder to obtain the fourth appearance feature set, perform feature extraction on the second image frame after being augmented by using the first encoder to obtain a fifth posture

feature set, and perform feature extraction on the fifth posture feature set by using a third encoder to obtain the fourth posture feature set.

[00341] In some embodiments, the video sample is provided with an appearance description text label of the sample object. The model may further include a text feature extraction network and a first sampling network. The apparatus may further include a second extraction module, a third sampling module, a third generation module, a second determination module and a second updating module. The second extraction module is configured to perform feature extraction on the appearance description text label by using the text feature extraction network to obtain a third text feature. The third sampling module is configured to sample a sixth appearance feature set of the sample object from the updated appearance codebook based on the third text feature by using the first sampling network, and sample a second reference posture feature set representing that the sample object is taking a preset posture from the updated posture codebook. The third generation module is configured to generate a fourth image frame corresponding to the preset posture based on the sixth appearance feature set and the second reference posture feature set by using the updated decoding network. The second determination module is configured to determine a second loss value based on the fourth image frame and a fifth image frame corresponding to the preset posture in the video sample. The second updating module is configured to update network parameters of the first sampling network based on the second loss value to obtain the updated first sampling network.

[00342] In some embodiments, the video sample is further provided with an action description text label of the sample object. The model may further include a second sampling network. The apparatus may further include a third extraction module, a fourth sampling module, a fifth sampling module, a fourth generation module, a third determination module and a third updating module. The third extraction module is configured to perform feature extraction on the action description text label by using the text feature extraction network to obtain a fourth text feature. The fourth sampling module is configured to sample a third posture feature set sequence of the sample object from the updated posture codebook based on the fourth text feature by using the second sampling network. The fifth sampling module sampling is configured to sample a seventh appearance feature set of the sample object from the updated appearance codebook based on the third text feature by using the updated first sampling network. The fourth

generation module is configured to generate a second video based on the seventh appearance feature set and the third posture feature set sequence by using the updated decoding network. The third determination module is configured to determine a third loss value based on the second video and the sample video. The third updating module is configured to update network parameters of the second sampling network based on the third loss value to obtain the updated second sampling network.

[00343] In some embodiments, the apparatus further includes a sixth sampling module configured to sample a third reference posture feature set representing that the sample object is taking the preset posture from the updated posture codebook based on the third text feature by using the updated first sampling network. The fourth sampling module is further configured to sample the third posture feature set sequence of the sample object from the updated posture codebook based on the fourth text feature and the third reference posture feature set by using the second sampling network.

[00344] In some embodiments, the fourth sampling module is further configured to determine a fourth posture feature set sequence, in which at least one fourth posture feature set is masked; and sample the third posture feature set sequence of the sample object from the updated posture codebook based on the fourth text feature, the third reference posture feature set and the fourth posture feature set sequence by using the second sampling network.

[00345] In some embodiments, the fourth sampling module is further configured to: predict each of the masked fourth posture feature sets in the fourth posture feature set sequence based on the fourth text feature, the third reference posture feature set and the fourth reference posture feature set sequence by using the second sampling network, to obtain the predicted fourth posture feature set sequence; and sample the third posture feature set sequence from the updated posture codebook based on the predicted fourth posture feature set sequence.

[00346] In some embodiments, the third determination module is further configured to: determine a fourth sub-loss value based on the second video and the sample video, determine a fifth sub-loss value based on the third posture feature set sequence and the fourth posture feature set sequence, and determine a third loss value based on the fourth sub-loss value and the fifth sub-loss value.

[00347] In some embodiments, the apparatus may further include a fourth extraction

module and a fourth determination module. The fourth extraction module is configured to perform feature extraction on each of the image frames in the sample video by using the second encoding network to obtain a fifth posture feature set sequence. The fourth determination module is configured to determine a sixth sub-loss value based on the fifth posture feature set sequence and the fourth posture feature set sequence. The third determination module is further configured to determine the third loss value based on the fourth sub-loss value, the fifth sub-loss value and the sixth sub-loss value.

[00348] In some embodiments, a dimension of the posture features in the posture codebook is lower than that of the appearance features in the appearance codebook. The decoding network includes a first decoder and a second decoder. The second generation module is further configured to: perform up-sampling on each posture feature in the third posture feature set by using the first decoder to obtain an up-sampled third posture feature set; and the dimension of each of third posture features in the up-sampled third posture feature set is the same as that of each of appearance features in the third appearance feature set; and generate the third image frame based on the third appearance feature set and the up-sampled third posture feature set by using the second decoder.

[00349] The above descriptions about the apparatus embodiments are similar to descriptions about the method embodiments and beneficial effects similar to those of the method embodiments are achieved.

[00350] Technical details undisclosed in the apparatus embodiments of the disclosure may be understood with reference to the descriptions about the method embodiments of the disclosure.

[00351] It is to be noted that in the embodiments of the disclosure, when being implemented in form of software function module and sold or used as an independent product, the method for generating a video or the method for training a model may also be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the embodiments of the disclosure substantially or parts making contributions to the related art may be embodied in form of software product, and the software product is stored in a storage medium, including a plurality of instructions configured to enable a computer device (which may be a personal computer, a server, a network device and the like) to execute all or part of the method in each embodiment of the disclosure. The storage medium includes: various media capable of storing program

codes such as a U disk, a mobile hard disk, a Read Only Memory (ROM), a magnetic disk or an optical disk. Therefore, the embodiments of the disclosure are not limited to any specific combination of hardware and software.

[00352] An embodiment of the disclosure provides a computer device, which may include: a memory and a processor. The memory stores a computer program capable of running on the processor. The processor implements the steps in the above method when executing the program.

[00353] An embodiment of the disclosure provides a computer readable storage medium, in which a computer program is stored. When executed by the processor, the computer program implements the steps in the above method.

[00354] An embodiment of the disclosure provides a computer program product, which may include a non-transitory computer-readable storage medium, in which a computer program is stored. When being read and executed by a computer, the computer program implements part or all steps in the above method. The computer program product may be specifically realized by means of hardware, software or a combination thereof. In an optional embodiment, the computer program product is specifically embodied as a computer storage medium, and in another optional embodiment, the computer program product is specifically embodied as software products, such as a Software Development Kit (SDK).

[00355] It is to be pointed out here that the above descriptions about the storage medium, computer program product and device embodiments are similar to descriptions about the method embodiments and beneficial effects similar to those of the method embodiments are achieved. Technical details undisclosed in the storage medium, computer program product and device embodiments of the disclosure may be understood with reference to the descriptions about the method embodiments of the disclosure.

[00356] It is to be noted that FIG. 7 illustrates a schematic diagram of a hardware entity of a computer device according to an embodiment of the disclosure. As illustrated in FIG. 7, the hardware entity of the computer device 800 includes a processor 801, a communication interface 802 and a memory 803.

[00357] The processor 801 usually controls an overall operation of the computer device 800.

[00358] The communication interface 802 may enable the computer device to

communicate with other terminals or servers through the network.

[00359] The memory 803 is configured to store an instruction and application executable for the processor 801, may also cache data (for example, image data, video data, voice communication data and video communication data) to be processed or having been processed by the processor 801 and each module in the computer device 800 and may be implemented through a flash or a Random Access Memory (RAM).

[00360] It is to be understood that "one embodiment" and "an embodiment" mentioned in the whole specification mean that specific features, structures or characteristics related to the embodiment is included in at least one embodiment of the disclosure. Therefore, "in one embodiment" or "in an embodiment" appearing at any place of the whole specification does not always refer to the same embodiment. In addition, these specific features, structures or characteristics may be combined in one or more embodiments in any proper manner. It should be understood that, in various embodiments of the present disclosure, a magnitude of a sequence number of each process does not mean an execution sequence and the execution sequence of each process should be determined by its function and an internal logic and should not form any limit to an implementation process of the embodiments of the disclosure. The sequence numbers of the embodiments of the disclosure are just for describing, instead of representing superiority-inferiority of the embodiments.

[00361] It is to be noted that terms "include" and "contain" or any other variant thereof is intended to cover nonexclusive inclusions herein, so that a process, method, object or device including a series of components not only includes those components but also includes other components which are not clearly listed or further includes components intrinsic to the process, the method, the object or the device. Under the condition of no more limitations, a component defined by the statement "including a/an....." does not exclude existence of the same other components in a process, method, object or device including the component.

[00362] In several embodiments provided by the disclosure, it should be understood that the disclosed device and method may be implemented in another manner. The device embodiment described above is only schematic, and for example, division of the units is only logic function division, and other division manners may be adopted during practical implementation. For example, multiple units or components may be combined or

integrated into another system, or some characteristics may be neglected or not executed. In addition, coupling or direct coupling or communication connection between each displayed or discussed component may be indirect coupling or communication connection, implemented through some interfaces, of the device or the units, and may be electrical and mechanical or adopt other forms.

[00363] The units described as separate parts may or may not be physically separated, and parts displayed as units may or may not be physical units, and namely may be located in the same place, or may also be distributed to multiple network units. Part of all of the units may be selected according to a practical requirement to achieve the purposes of the solutions of the embodiments.

[00364] In addition, each functional unit in each embodiment of the disclosure may be integrated into a processing unit, each unit may also serve as an independent unit and two or more than two units may also be integrated into a unit. The integrated unit may be implemented in a hardware form and may also be implemented in form of hardware and software functional unit.

[00365] Those of ordinary skill in the art should know that all or part of the steps of the method embodiment may be implemented by related hardware instructed through a program, the program may be stored in a computer-readable storage medium, and the program is executed to execute the steps of the method embodiment. The storage medium which may be a volatile storage medium includes: various media capable of storing program codes such as a mobile storage device, a ROM, a magnetic disk or a compact disc. Or, when being implemented in form of software function module and sold or used as an independent product, the integrated unit of the disclosure may also be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the disclosure substantially or parts making contributions to the related art may be embodied in form of software product, and the computer software product is stored in a storage medium, including a plurality of instructions configured to enable a computer device (which may be a personal computer, a server, a network device and the like) to execute all or part of the method in each embodiment of the disclosure. The storage medium includes: various media capable of storing program codes such as a mobile hard disk, a ROM, a magnetic disk or a compact disc.

[00366] The above is only the specific implementation manner of the disclosure and

not intended to limit the scope of protection of the disclosure. Any variations or replacements apparent to those skilled in the art within the technical scope disclosed by the disclosure shall fall within the scope of protection of the disclosure.

CLAIMS

1. A video generating method, comprising:

- obtaining an appearance description text and an action description text of an object;
- sampling a first appearance feature set of the object from an appearance codebook based on the appearance description text, the appearance codebook comprising a plurality of appearance features, and the first appearance feature set comprising first appearance features, each corresponding to at least one pixel of an image frame in a first video to be generated;
- sampling a first posture feature set sequence of the object from a posture codebook based on the action description text, the posture codebook comprising a plurality of posture features, the first posture feature set sequence comprising first posture feature sets, each corresponding to at least one image frame in the first video to be generated, and each of the first posture feature sets comprising first posture features, each corresponding to at least one pixel in the corresponding image frame; and
- generating the first video containing the object based on the first appearance feature set and the first posture feature set sequence.

2. The method according to claim 1, wherein sampling the first appearance feature set of the object from the appearance codebook based on the appearance description text comprises:

- performing feature extraction on the appearance description text to obtain a first text feature; and
- sampling the first appearance feature set from the appearance codebook based on the first text feature;
- wherein sampling the first posture feature set sequence of the object from the posture codebook based on the action description text comprises:
 - performing feature extraction on the action description text to obtain a second text feature; and
 - sampling the first posture feature set sequence from the posture codebook based on the second text feature.

3. The method according to claim 2, wherein sampling the first appearance feature set from the appearance codebook based on the first text feature comprises:

predicting the appearance features of the object based on the first text feature to obtain a second appearance feature set;

for each of second appearance features in the second appearance feature set, selecting a first appearance feature matched with the second appearance feature from the appearance codebook, to obtain first appearance features; and

obtaining the first appearance feature set based on each of the first appearance features.

4. The method according to any one of claims 2 to 3, further comprising:

sampling a first reference posture feature set representing that the object is taking a preset posture from the posture codebook based on the first text feature;

wherein sampling the first posture feature set sequence from the posture codebook based on the second text feature comprises:

sampling the first posture feature set sequence from the posture codebook based on the second text feature and the first reference posture feature set.

5. The method according to claim 4, wherein sampling the first posture feature set sequence from the posture codebook based on the second text feature and the first reference posture feature set comprises:

determining a second posture feature set sequence, at least one second posture feature set in the second posture feature set sequence being masked; and

performing at least one feature sampling in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, until each of the second posture feature sets in the second posture feature set sequence is obtained, and taking the sampled second posture feature set sequence as the first posture feature set sequence.

6. The method according to claim 5, wherein all of the second posture feature sets in the second posture feature set sequence are masked; and performing at least one feature sampling in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, until each of the second posture feature sets in the second posture feature set sequence is obtained, comprises:

predicting a first one of the second posture feature sets and an n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, to obtain the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets, and updating the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets to the second posture feature set sequence, and where n is a number of the second posture feature sets in the second posture feature set sequence; and

predicting each of the second posture feature sets between the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text features, the first reference posture feature set and the updated second posture feature set sequence, and updating each of the predicted second posture feature sets to the second posture feature set sequence.

7. The method according to claim 6, wherein predicting the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence to obtain the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets comprises:

performing at least one prediction on the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first reference posture feature set and the second posture feature set sequence, until each of second posture features in the first one of the second posture feature sets and each of the second posture features in the n^{th} one of the second posture feature sets are obtained, and updating the first one of the second posture feature sets and the n^{th} one of the second posture feature sets to the second posture feature set sequence.

8. The method according to claim 6 or 7, wherein predicting the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence in the posture codebook based on the second text feature, the first

reference posture feature set and the second posture feature set sequence to obtain the predicted first one of the second posture feature sets and n^{th} one of the second posture feature sets comprises:

predicting the first one of the second posture feature sets and the n^{th} one of the second posture feature sets in the second posture feature set sequence based on the second text features, the first reference posture feature set and the second posture feature set sequence to obtain a first candidate posture feature set and a second candidate posture feature set;

determining the predicted first one of the second posture feature sets from a posture feature codebook based on the first candidate posture feature set; and

determining the predicted n^{th} one of the second posture feature sets from the posture feature codebook based on the second candidate posture feature set.

9. A method for training a model, wherein the model comprises an appearance codebook, a posture codebook and a decoding network; and the method comprises:

obtaining a video sample containing a sample object, the video sample comprising a first image frame and a second image frame;

performing, by using the appearance codebook, feature extraction on the first image frame, to obtain a third appearance feature set; and performing, by using the posture codebook, feature extraction on the second image frame after being augmented, to obtain a third posture feature set; the appearance codebook comprising a plurality of appearance features, the posture codebook comprising a plurality of posture features, and the augmenting being configured to reduce appearance information in the second image frame;

generating a third image frame based on the third appearance feature set and the third posture feature set by using the decoding network;

determining a first loss value based on the third image frame and the second image frame before being augmented; and

updating the appearance codebook, the posture codebook and network parameters of the decoding network based on the first loss value, to obtain the updated appearance codebook, posture codebook and decoding network.

10. The method according to claim 9, wherein,

performing, by using the appearance codebook, feature extraction on the first image

frame to obtain the third appearance feature set comprises: performing, by using a first encoding network, feature extraction on the first image frame, to obtain a fourth appearance feature set, and sampling the third appearance feature set from the appearance codebook based on the fourth appearance feature set;

performing, by using the posture codebook, feature extraction on the second image frame after being augmented to obtain the third appearance feature set comprises: performing, by using a second encoding network, feature extraction on the second image frame after being augmented to obtain a fourth posture feature set, and sampling the third posture feature set from the posture codebook based on the fourth posture feature set;

updating the appearance codebook, the posture codebook and the network parameters of the decoding network based on the first loss value to obtain the updated appearance codebook, posture codebook and decoding network comprises: updating the appearance codebook, the posture codebook, the network parameters of the decoding network, the network parameters of the first encoding network and the network parameters of the second encoding network based on the first loss value, to obtain the updated appearance codebook, posture codebook, decoding network, first encoding network and second encoding network.

11. The method according to claim 10, wherein the video sample is provided with an appearance description text label of the sample object; the model further comprises a text feature extraction network and a first sampling network; and the method further comprises:

performing feature extraction on the appearance description text label by using the text feature extraction network, to obtain a third text feature; and

sampling a sixth appearance feature set of the sample object from the updated appearance codebook based on the third text feature by using the first sampling network, and sampling a second reference posture feature set representing that the sample object is taking a preset posture from the updated posture codebook by using the first sampling network;

generating a fourth image frame corresponding to the preset posture based on the sixth appearance feature set and the second reference posture feature set by using the updated decoding network;

determining a second loss value based on the fourth image frame and a fifth image frame corresponding to the preset posture in the video sample; and

updating the network parameters of the first sampling network based on the second

loss value to obtain the updated first sampling network.

12. The method according to claim 11, wherein the video sample is further provided with an action description text label of the sample object; the model further comprises a second sampling network; and the method further comprises:

- performing feature extraction on the action description text label by using the text feature extraction network, to obtain a fourth text feature;

- sampling a third posture feature set sequence of the sample object from the updated posture codebook based on the fourth text feature by using the second sampling network;

- sampling a seventh appearance feature set of the sample object from the updated appearance codebook based on the third text feature by using the updated first sampling network;

- generating a second video based on the seventh appearance feature set and the third posture feature set sequence by using the updated decoding network;

- determining a third loss value based on the second video and a sample video; and

- updating the network parameters of the second sampling network based on the third loss value to obtain the updated second sampling network.

13. The method according to claim 12, further comprising:

- sampling a third reference posture feature set representing that the sample object is taking the preset posture from the updated posture codebook based on the third text feature by using the updated first sampling network;

- wherein sampling the third posture feature set sequence of the sample object from the updated posture codebook based on the fourth text feature by using the second sampling network comprises:

- sampling the third posture feature set sequence of the sample object from the updated posture codebook based on the fourth text feature and the third reference posture feature set by using the second sampling network.

14. The method according to claim 13, wherein sampling the third posture feature set sequence of the sample object from the updated posture codebook based on the fourth text feature and the third reference posture feature set by using the second sampling network

comprises:

- determining a fourth posture feature set sequence, at least one fourth posture feature set in the fourth posture feature set sequence being masked;

- predicting each of the masked fourth posture feature sets in the fourth posture feature set sequence based on the fourth text feature, the third reference posture feature set and the fourth reference posture feature set sequence by using the second sampling network, to obtain the predicted fourth posture feature set sequence; and

- sampling the third posture feature set sequence from the updated posture codebook based on the predicted fourth posture feature set sequence.

15. The method according to claim 14, further comprising:

- performing feature extraction on each of image frames in the sample video by using the second sampling network, to obtain a fifth posture feature set sequence;

- determining a sixth sub-loss value based on the fifth posture feature set sequence and the fourth posture feature set sequence;

- wherein determining the third loss value based on the second video and the sample video comprises:

 - determining a fourth sub-loss value based on the second video and the sample video;

 - determining a fifth sub-loss value based on the third posture feature set sequence and the fourth posture feature set sequence; and

 - determining the third loss value based on the fourth sub-loss value, the fifth sub-loss value and the sixth sub-loss value.

16. An apparatus for generating a video, comprising:

- a first obtaining module, configured to obtain an appearance description text and an action description text of an object;

- a first sampling module, configured to sample a first appearance feature set of the object from an appearance codebook based on the appearance description text, wherein the appearance codebook comprises a plurality of appearance features, and the first appearance feature set comprises first appearance features, each corresponding to at least one pixel of an image frame in a first video to be generated;

- a second sampling module, configured to sample a first posture feature set sequence of

the object from a posture codebook based on the action description text, wherein the posture codebook comprises a plurality of posture features, the first posture feature set sequence comprises first posture feature sets, each corresponding to at least one image frame in the first video to be generated, and each of the first posture feature sets comprises first posture features, each corresponding to at least one pixel in the corresponding image frame; and

a first generation module, configured to generate the first video containing the object based on the first appearance feature set and the first posture feature set sequence.

17. An apparatus for training a model, wherein the model comprises an appearance codebook, a posture codebook and a decoding network; and the apparatus comprises:

a second obtaining module, configured to obtain a video sample containing a sample object, wherein the video sample comprises a first image frame and a second image frame;

a first extraction module, configured to perform feature extraction on the first image frame by using the appearance codebook to obtain a third appearance feature set, and perform feature extraction on the second image frame after being augmented by using the posture codebook to obtain a third posture feature set, wherein the appearance codebook comprises a plurality of appearance features, the posture codebook comprises a plurality of posture features, and the augmenting is configured to reduce appearance information in the second image frame;

a second generation module, configured to generate a third image frame based on the third appearance feature set and the third posture feature set by using the decoding network;

a first determination module, configured to determine a first loss value based on the third image frame and the second image frame before the augmentation; and

a first updating module, configured to update the appearance codebook, the posture codebook and network parameters of the decoding network based on the first loss value to obtain the updated appearance codebook, posture codebook and decoding network.

18. A computer device, comprising a memory and a processor, the memory storing a computer program capable of running in the processor, wherein the processor, when executing the computer program, is configured to implement the steps in the method of any one of claims 1-15.

19. A computer-readable storage medium, in which a computer program is stored, wherein the computer program, when executed by a processor, implements the steps in the method of any one of claims 1-15.

20. A computer program product, comprising a non-transitory computer-readable storage medium, in which a computer program is stored, wherein the computer program, when read and executed by a computer, implements the steps in the method of any one of claims 1-15.