

METHOD AND APPARATUS FOR GENERATING VIRTUAL OBJECT, DEVICE, AND STORAGE MEDIUM

TECHNICAL FIELD

- [0001] The present disclosure relates, but is not limited, to the technical field of computers, and in particular, to a method and apparatus for generating a virtual object, a device, and a storage medium.

BACKGROUND

- [0002] Digital avatars have become an essential component of scenes such as movies, games, and virtual reality. The process of producing the digital avatars includes creating the shapes of characters, painting textures, assembling skeletons, and driving the avatars with captured actions. However, this process is usually implemented by professional personnel with the help of professional software and professional devices, or may be implemented by a neural network trained based on a large number of training samples.

SUMMARY

- [0003] Embodiments of the present disclosure provide a method and apparatus for generating a virtual object, a device, and a storage medium.
- [0004] The technical solutions of the embodiments of the present disclosure are implemented as follows.
- [0005] An embodiment of the present disclosure provides a method for generating a virtual object. The method includes the following operations.
- [0006] First prompt information and second prompt information are acquired. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.
- [0007] A first virtual model is determined based on the first prompt information.
- [0008] The virtual object is generated based on the second prompt information and the first virtual model.
- [0009] An embodiment of the present disclosure provides an apparatus for generating a virtual object. The apparatus includes a first determination module, a second determination module, and a first generation module.
- [0010] The first determination module is configured to acquire first prompt information and second prompt information. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.
- [0011] The second determination module is configured to determine a first virtual model based on the first prompt information.
- [0012] The first generation module is configured to generate the virtual object based on the second prompt information and the first virtual model.
- [0013] An embodiment of the present disclosure provides electronic device. The electronic device includes a processor and a memory. The memory stores a computer program executable on the processor. The processor, when executing the computer program, implements the above method for generating a virtual object.
- [0014] An embodiment of the present disclosure provides a computer-readable storage medium. The computer-readable storage medium stores a computer program. The above method for generating a virtual object is implemented when the computer program is

performed by a processor.

[0015] In the embodiments of the present disclosure, the first prompt information and the second prompt information are acquired, the first prompt information represents the body shape information of the virtual object, the second prompt information represents the appearance information of the virtual object; the first virtual model is determined based on the first prompt information; and the virtual object is generated based on the second prompt information and the first virtual model. In this way, based on the body shape information and the appearance information, the virtual object matching the body shape information and the appearance information can be automatically and accurately generated, so that the threshold for making a digital avatar can be reduced. In addition, the virtual object in the related art is usually made manually by those skilled in the art based on prior information such as experience and actual application scenarios, or is generated by using a neural network trained with a large number of training samples, while in the method for generating a virtual object provided in the embodiments of the present disclosure, there is no need to use the prior information and/or additional training samples to train the neural network, so that the cost of generating the virtual object can be reduced.

[0016] It should be understood that, the above general description and the following detailed description are merely exemplary and explanatory, and do not limit the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] Drawings are incorporated into the specification and constitute part of the specification. The drawings illustrate embodiments in accordance with the present disclosure and serve to describe the technical solutions of the present disclosure together with the specification.

[0018] FIG. 1 is a schematic diagram of an implementation process of a method for generating a virtual object according to an embodiment of the present disclosure.

[0019] FIG. 2 is a schematic diagram of an implementation process of a method for generating a virtual object according to an embodiment of the present disclosure.

[0020] FIG. 3 is a schematic diagram of an implementation process of a method for generating a virtual object according to an embodiment of the present disclosure.

[0021] FIG. 4 is a schematic diagram of an implementation process of generating a target animation of a virtual object according to an embodiment of the present disclosure.

[0022] FIG. 5A is a schematic diagram of a composition structure of a system for generating a virtual object according to an embodiment of the present disclosure.

[0023] FIG. 5B is a schematic diagram of a system for generating a virtual object according to an embodiment of the present disclosure.

[0024] FIG. 5C is a schematic diagram of determination of a target body shape image according to an embodiment of the present disclosure.

[0025] FIG. 5D is a schematic diagram of a process for generating a virtual object according to an embodiment of the present disclosure.

[0026] FIG. 5E is a schematic diagram of generation of a target animation of a virtual object according to an embodiment of the present disclosure.

[0027] FIG. 5F is a schematic diagram of determination of a first candidate pose image according to an embodiment of the present disclosure.

[0028] FIG. 5G is a schematic diagram of determination of a second target loss value according to an embodiment of the present disclosure.

- [0029] FIG. 5H is a schematic diagram of a pre-trained VAE network structure according to an embodiment of the present disclosure.
- [0030] FIG. 5I is a schematic diagram of results of generating a virtual object based on a plurality of methods according to an embodiment of the present disclosure.
- [0031] FIG. 5J is a schematic diagram of results of generating a first virtual model based on a plurality of methods according to an embodiment of the present disclosure.
- [0032] FIG. 5K is a schematic diagram of results of generating a virtual object based on a plurality of methods according to an embodiment of the present disclosure.
- [0033] FIG. 5L is a schematic diagram of results of generating a target animation based on a plurality of methods according to an embodiment of the present disclosure.
- [0034] FIG. 6 is a schematic diagram of a composition structure of an apparatus for generating a virtual object according to an embodiment of the present disclosure.
- [0035] FIG. 7 is a schematic diagram of a hardware entity of an electronic device according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

- [0036] For making the purposes, technical solutions and advantages of the present disclosure clearer, the technical solutions of the present 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 present 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 present disclosure.
- [0037] "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.
- [0038] 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 present disclosure described herein in sequences except the illustrated or described ones.
- [0039] Unless otherwise defined, all technical and scientific terms in the specification have the same meaning as those skilled in the art, belonging to the present disclosure, usually understand. Terms used in the specification are only used for describing the purpose of the present disclosure, but not intended to limit the present disclosure.
- [0040] In the related art, digital avatars have become an essential component of scenes such as movies, games, and virtual reality. The process of producing the digital avatars includes creating the shapes of characters, painting textures, assembling skeletons, and driving the avatars with captured actions. Modeling and motion capture for the digital avatars are critical technologies for the film and game industry. However, during the modeling and motion capture, it often requires a large number of professional personnel to use professional software and expensive devices to achieve the purpose, which is time-consuming and expensive. In addition, conventional neural networks are trained based on a large amount of training data, and cannot be automatically generated based on a natural language.
- [0041] An embodiment of the present disclosure provides a method for generating a virtual object. Based on body shape information and appearance information, a virtual object matching the body shape information and the appearance information can be

automatically and accurately generated, so that the threshold for making a digital avatar can be reduced. In addition, there is no need to use the prior information and/or additional training samples to train the neural network, thereby reducing the cost of generating the virtual object. The method for generating a virtual object provided in the embodiments of the disclosure may be performed by an electronic device. The electronic device may be various types of terminals such as a notebook computer, a tablet computer, a desktop computer, a set-top box, or a mobile device (for example, a mobile phone, a portable music player, a personal digital assistant, a special message device, and a portable game device), or may be implemented as a server. The server may be an independent physical server, may be a server cluster formed by a plurality of physical servers or a distributed system, or may be a cloud server configured to provide basic cloud computing services such as a cloud service, a cloud database, cloud computing, a cloud function, cloud storage, a network service, cloud communication, a middleware service, a domain name service, a security service, a Content Delivery Network (CDN), and a big data and artificial intelligence platform.

[0042] The technical solutions in the embodiments of the disclosure will be clearly and completely described below with reference to the accompanying drawings in the embodiments of the disclosure.

[0043] FIG. 1 is a schematic diagram of an implementation process of a method for generating a virtual object according to an embodiment of the present disclosure. As shown in FIG. 1, the method includes S11 to S13.

[0044] At S11, first prompt information and second prompt information are acquired. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.

[0045] Here, the first prompt information and the second prompt information may be text prompt information, voice prompt information, and the like. For example, the first prompt information may be text prompt information or voice prompt information that indicates body shape, such as tall, short, fat, thin, cube. For another example, the second prompt information may be text prompt information or voice prompt information that indicates appearance, such as "Avatar", "Iron Man", "Spider Man", "rainbow", "tree", "girl in skirt", "boy wearing helmet", "student with schoolbag". The virtual object may be a two-dimensional model, or may be a three-dimensional model. For example, the virtual object may be an Iron Man model.

[0046] In some implementation modes, the manner of acquiring the first prompt information and the second prompt information may be determined according to a practical application scenario. For example, the first prompt information and the second prompt information may be acquired based on information inputted by a user through an input device of an electronic device. For another example, the first prompt information and the second prompt information may be acquired based on information transmitted by other electronic devices. For example, the user inputs "tall and strong Iron Man" by using a keyboard, the electronic device performs parsing to obtain the first prompt information of "tall and strong man" and the second prompt information of "Iron Man". During implementation, those skilled in the art may determine the manner of acquiring the first prompt information and the second prompt information according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0047] At S12, a first virtual model is determined based on the first prompt information.

[0048] Here, the first virtual model may be a two-dimensional model, a three-

dimensional model, or a multi-dimensional model. The first virtual model is a model matching the first prompt information. The first virtual model is a model which has a shape. For example, the first virtual model may be a three-dimensional human body model. For another example, the first virtual model may be a two-dimensional rectangular model.

[0049] In some implementation modes, a body shape image matching the first prompt information may be determined, and three-dimensional modeling may be performed on the body shape image to obtain the first virtual model.

[0050] In some implementation modes, the first virtual model may be determined through a preset second neural network based on the first prompt information. The preset second neural network may be a neural network pre-trained based on a large number of training samples. The training samples may be the first prompt information. During implementation, those skilled in the art may select an appropriate second neural network according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0051] At S13, the virtual object is generated based on the second prompt information and the first virtual model.

[0052] Here, the virtual object is a model matching the second prompt information. For example, the virtual object may be an Iron Man model. For another example, the virtual object may be a colored three-dimensional house model.

[0053] In some implementation modes, the virtual object may be generated through a preset third neural network based on the second prompt information. The preset third neural network is configured to perform geometry and texture rendering on the first virtual model. The preset third neural network may be a neural network pre-trained based on a large number of training samples. The training samples may be body shape images, or body shape models. For example, the preset third neural network may be a neural network obtained by optimizing a volume rendering model NeuS. NeuS is a neural network which combines a Signed Distance Function (SDF) and Neural Radiance Fields (NeRF). The optimization may include, but not limited to, adding a random color network. Through the random color network, the texture and style of the first virtual model may be rendered more vividly. During implementation, those skilled in the art may select the appropriate third neural network according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0054] In this embodiment of the present disclosure, the first prompt information and the second prompt information are acquired, the first prompt information represents the body shape information of the virtual object, the second prompt information represents the appearance information of the virtual object; the first virtual model is determined based on the first prompt information; and the virtual object is generated based on the second prompt information and the first virtual model. In this way, based on the body shape information and the appearance information, the virtual object matching the body shape information and the appearance information can be automatically and accurately generated, so that the threshold for making a digital avatar can be reduced. In addition, the virtual object in the related art is usually made manually by those skilled in the art based on prior information such as experience and actual application scenarios, or is generated by using a neural network trained with a large number of training samples, while in the method for generating a virtual object provided in the embodiments of the present disclosure, there is no need to use the prior information and/or additional training samples to train the neural network, so that the cost of generating the virtual object can

be reduced.

[0055] FIG. 2 is a schematic diagram of an implementation process of a method for generating a virtual object according to an embodiment of the present disclosure. As shown in FIG. 2, the method includes S21 to S24.

[0056] At S21, first prompt information and second prompt information are acquired. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.

[0057] S21 corresponds to S11, and may refer to the specific implementation mode of S11 during implementation.

[0058] At S22, a target body shape image matching the first prompt information is determined based on a preset first correspondence.

[0059] Here, the first correspondence represents a correspondence between the first prompt information and the body shape image.

[0060] In some implementation modes, a mapping table of the correspondence between the first prompt information and the body shape image may be established in advance, and the mapping table is stored in the electronic device or other terminals. In the case that the mapping table is stored in the electronic device, the electronic device determines the body shape image matching the first prompt information in the mapping table according to the first prompt information. In the case that the mapping table is stored in other terminals, the electronic device transmits the first prompt information to other terminals, so that other terminals determine the body shape image matching the first prompt information in the mapping table according to the first prompt information and returns the corresponding body shape image to the electronic device.

[0061] In some implementation modes, the first correspondence may be established by using a preset fourth neural network.

[0062] The preset fourth neural network is a neural network pre-trained based on a large number of training samples. The training samples are text-image pairs. For example, the preset fourth neural network may be a Contrastive Language-Image Pre-Training (CLIP) model. During implementation, those skilled in the art may select the appropriate fourth neural network according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0063] In some implementation modes, the preset fourth neural network may include one image encoder E_I and one text encoder E_T . The manner of training the encoders is to establish correspondences based on pairs of image and text. The network structure of a first model is optimized based on a loss value. The loss value may be calculated by using the following formula (2-1):

$$\mathcal{L}_{clip}(I, T) = 1 - \text{norm}(E_I(I)) \cdot \text{norm}(E_T(T)) \quad (2-1)$$

where $(\cdot)$ represents a cosine distance; I represents an image; T represents a text prompt; and norm is a function and is used for regularizing an inputted vector. Through minimization of $\mathcal{L}_{clip}(I, T)$, the generated image I matches the description of the text prompt T .

[0064] In some implementation modes, S22 includes S221 to S223.

[0065] At S221, a first distance between each body shape image in the first correspondence and a set standard body shape image is determined respectively.

[0066] Here, the preset standard body shape image may be inputted by the user, or may be pre-stored in the electronic device. The preset standard body shape image corresponds to standard prompt information.

[0067] In some implementation modes, the first distance may be a similarity between

an image feature of the body shape image and an image feature of the standard body shape image, may be a similarity between a weighted image feature of the body shape image and a weighted image feature of the standard body shape image, or may be determined based on the image encoder of a preset neural network. During implementation, those skilled in the art may select the manner of determining the first distance according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0068] In some implementation modes, the first distance may be calculated by using the following formula (2-2):

$$\Delta f_I = E_I(\mathcal{R}(M_B[i])) - E_I(\mathcal{R}(M_n)) \quad (2-2)$$

where $M_B[i]$ represents the i-th body shape image, M_n represents the standard body shape image, $\mathcal{R}(\cdot)$ represents a mesh renderer, and E_I represents the image encoder.

[0069] At S222, a second distance between the first prompt information and standard prompt information corresponding to the standard body shape image is determined.

[0070] In some implementation modes, the second distance may be a similarity between a text feature of the first prompt information and a text feature of the standard prompt information, may be a similarity between a weighted text feature of the first prompt information and a weighted text feature of the standard prompt information, or may be determined based on the text encoder of the preset neural network. During implementation, those skilled in the art may select the manner of determining the second distance according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0071] In some implementation modes, the second distance may be calculated by using the following formula (2-3):

$$\Delta f_T = E_T(t_{shape}) - E_T(t_n) \quad (2-3)$$

where t_{shape} represents the first prompt information, t_n represents the standard prompt information, and E_T represents the text encoder.

[0072] At S223, the target body shape image is determined based on each first distance and each second distance.

[0073] In some implementation modes, each first distance may be respectively scored based on the second distance to obtain at least one score. The target body shape image is determined based on the at least one score. For example, when the score indicates a difference between the first distance and the second distance, the body shape image corresponding to the first distance with the minimum score is used as the target body shape image. For another example, when the score indicates a similarity between the first distance and the second distance, the body shape image corresponding to the first distance with the maximum score is used as the target body shape image. During implementation, those skilled in the art may use an appropriate scoring rule to determine the target body shape image according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0074] During implementation, the score may be implemented by using the following formula (2-4):

$$s_i = 1 - \text{norm}(\Delta f_I) \cdot \text{norm}(\Delta f_T) \quad (2-4)$$

where $(\cdot)$ represents a cosine distance; and norm is a function and is used for regularizing an inputted vector.

[0075] At S23, three-dimensional modeling is performed based on the target body shape image, to generate the first virtual model.

- [0076] Here, the first virtual model may be a three-dimensional model of the target body shape image, or may be a three-dimensional model generated obtained from transforming the target body shape image.
- [0077] In some implementation modes, three-dimensional modeling is performed on the target body shape image based on a preset fifth neural network, to generate the first virtual model. The preset fifth neural network may include, but not limited to, a Skinned Multi-Person Linear (SMPL) model, SMPL-X, Generative Adversarial Networks (GAN), and the like. During implementation, those skilled in the art may use an appropriate fifth neural network to perform three-dimensional modeling on the target body shape image according to actual requirements, which is not limited in the embodiments of the present disclosure.
- [0078] At S24, the virtual object is generated based on the second prompt information and the first virtual model.
- [0079] S24 corresponds to S13, and may refer to the specific implementation mode of S13 during implementation.
- [0080] In the embodiment of the disclosure, the target body shape image matching the body shape information is determined based on the first correspondence, and three-dimensional modeling is performed based on the target body shape image, to generate the first virtual model; the virtual object is generated based on the first virtual model and the appearance information. In this way, the target body shape image matching the body shape information may be obtained based on the first correspondence, so as to further enhance the accuracy of the virtual object. In addition, three-dimensional modeling is directly performed based on the target body shape image by using a pre-trained neural network, so that zero sample generation capability can be achieved. Compared with a conventional method, there is no need to use the prior information and/or additional training samples to train the neural network, so that overfitting to the prior information and/or the learning during neural network training can be reduced. Therefore, recognition capabilities on body shape prompt information and appearance prompt information that are not considered in the prior information and/or body shape prompt information and appearance prompt information that are not involved during neural network training can be enhanced, and a generalization capability for recognizing the body shape prompt information and the appearance prompt information can be improved.
- [0081] FIG. 3 is a schematic diagram of an implementation process of a method for generating a virtual object according to an embodiment of the present disclosure. As shown in FIG. 3, the method includes S31 to S36.
- [0082] At S31, first prompt information and second prompt information are acquired. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.
- [0083] At S32, a first virtual model is determined based on the first prompt information.
- [0084] S31 to S32 correspond to S11 to S12, and may refer to specific implementation modes of S11 to S12 during implementation.
- [0085] At S33, for each first observation view angle of at least one first observation view angle, a first observation image of the first virtual model at the first observation view angle is determined.
- [0086] Here, at least one first observation view angle may be an observation view angle determined based on at least one camera. Each first observation image is an image obtained after the first virtual model is observed based on each observation view angle.

The first observation image is a two-dimensional image, and includes at least one pixel point.

[0087] At S34, for each first observation image, the first observation image is updated for at least one time by using a preset first neural network, to obtain a second observation image.

[0088] Here, the first neural network may be a pre-trained neural network. For example, the preset first neural network may be a neural network NeuS N' obtained by optimizing the NeuS. In some implementation modes, the first neural network includes a distance neural network $f(p)$, a first color network $c(p)$, and a second color network $c_c(p)$. $c_c(p)$ is a random color network. $f(p)$ and $c(p)$ are responsible for reconstructing the first virtual model. $f(p)$ and $c_c(p)$ are responsible for performing texture and pattern rendering on the reconstructed model.

[0089] At S35, a first target loss value is determined based on each second observation image and the second prompt information.

[0090] In some implementation modes, the first target loss value may be determined based on the first observation image, the second observation image, and the second prompt information.

[0091] In some implementation modes, the second prompt information may include face prompt information and back prompt information.

[0092] At S36, in the case that the first target loss value satisfies a first preset condition, the first virtual model is rendered based on each second observation image, to generate the virtual object.

[0093] Here, the first preset condition may include, but is not limited to, the first target loss value being less than a threshold, first target loss value change convergence, and the like. During implementation, those skilled in the art may independently determine the first preset condition according to actual requirements, which is not limited in the embodiments of the present disclosure.

[0094] In the embodiment of the disclosure, the first prompt information and the second prompt information are acquired, the first prompt information represents body shape information of the virtual object, the second prompt information represents appearance information of the virtual object; the first virtual model is determined based on the first prompt information; for each first observation view angle of the at least one first observation view angle, the first observation image of the first virtual model at the first observation view angle is determined; for each first observation image, the first observation image is updated for at least one time by using the preset first neural network, to obtain the second observation image; the first target loss value is determined based on each second observation image and the second prompt information; in the case that the first target loss value satisfies the first preset condition, the first virtual model is rendered based on each second observation image, to generate the virtual object. In this way, modeling is directly performed based on the first virtual model by using the pre-trained first neural network, so that zero sample generation capability can be achieved. Compared with a conventional method, there is no need to use the prior information and/or additional training samples to train the model, so that overfitting to the prior information and/or the learning during neural network training can be reduced. Therefore, recognition capabilities on prompt information that is not considered in the prior information and/or prompt information that is not involved during neural network training can be enhanced, and a generalization capability for recognizing the prompt information can be improved. At the same time, the first virtual model is optimized by

using the optimized neural network based on the appearance information. Therefore, the uniformity of texture rendering can be improved, and the accuracy of generating the virtual object can be further enhanced.

- [0095] In some implementation modes, the method further includes S37 to S39.
- [0096] At S37, in the case that the first target loss value does not satisfy the first preset condition, a network parameter of the first neural network is updated based on the first target loss value.
- [0097] Here, the network parameter of the first neural network may include, but is not limited to, network parameters in the distance neural network, the first color network, and the second color network. In some implementation modes, the manner of updating the network parameter of the first neural network may be determined according to actual situations, and may include, but is not limited to, at least one of a gradient descent method or a Newton momentum method, which is not limited here.
- [0098] At S38, each first observation image is updated for at least one time by using the updated first neural network, to obtain the updated second observation image.
- [0099] At S39, an updated first target loss value is determined based on each updated second observation image and the second prompt information.
- [00100] S38 to S39 correspond to S34 to S35, and may refer to specific implementation modes of S34 to S35 during implementation.
- [00101] In the implementation mode of the disclosure, next updating may be performed on the network parameter of the first neural network in a case that the first target loss value does not meet the first preset condition. Based on the first neural network after the next update, each first observation image is updated for at least one time, so that the virtual object can be further accurately obtained through continuous iterative update.
- [00102] In some implementation modes, the first observation view angle includes a first observation point and at least one first observation direction corresponding to the first observation point. The first observation image includes at least one first pixel point. Each first pixel point corresponds to one first intersection point on the first virtual model. Each first intersection point represents the first intersection point between the first virtual model and a line coming from the first observation point and extending along one first observation direction. S34 includes S341 to S343.
- [00103] At S341, at least one first spatial point corresponding to each first observation direction is determined based on the first observation point and each first observation direction.
- [00104] Here, the first spatial point is a spatial point on a ray in the first observation direction based on the first observation point.
- [00105] At S342, for each first spatial point of each first observation direction, a first distance value between the first intersection point corresponding to the first observation direction where the first spatial point is located and the first spatial point is determined by using the preset first neural network, and a first color value obtained by observing the first virtual model according to the first observation direction is determined at the first spatial point.
- [00106] Here, each first distance value may be determined based on the distance neural network of the first neural network. Each first color value is determined based on the first color network of the first neural network. For example, each first distance value $f(p(t))$ is determined based on $f(p)$ in N' . Each first color value $c_c(p(t), v)$ is determined based on $c_c(p)$ in N' .
- [00107] At S343, the first observation image is updated for at least one time based on

the first distance value and the first color value corresponding to each first spatial point, to obtain the second observation image.

[00108] In some implementation modes, for each first observation direction, a color value of the first pixel point corresponding to the observation direction may be determined based on a first weight value of each first spatial point in the first observation direction and the first color value. The first pixel point is updated based on the color value of the first pixel point, to achieve at least one updating on the first observation image.

[00109] In some implementation modes, the color value of the first pixel point may include, but is not limited to, a cumulative value between each first weight value and each first color value, a cumulative value between each weighted first weight value and each weighted first color value, a mean/variance of each first color value, and a mean/variance of each weighted first color value. During implementation, those skilled in the art may determine the manner of the color value of the first pixel point according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00110] In some implementation modes, the color value of the first pixel point may be calculated by using the following formula (3-1):

$$c_c(o, v) = \int_0^\infty w(t) c_c(p(t), v) dt \quad (3-1)$$

where o is the first observation point, v is the first observation direction, $p(t)$ refers to the first spatial point, $w(t)$ denotes the first weight value of the first spatial point, and $c_c(p(t), v)$ denotes the first color value of the first spatial point.

[00111] In some implementation modes, the first target loss value may be calculated based on the first distance value and the first color value. The first observation image is updated for at least one time based on the first target loss value and the first preset condition. The first observation image is updated for at least one time in the case that the first target loss value does not satisfy the first preset condition. The first virtual model is rendered based on the second observation image in the case that the first target loss value satisfies the first preset condition.

[00112] In some implementation modes, S343 includes S3431 to S3432.

[00113] At S3431, for each first observation direction, the first weight value of each first spatial point is determined based on the first distance value corresponding to each first spatial point in the first observation direction.

[00114] In some implementation modes, the first weight value may be a ratio of the first distance value of the first spatial point to a sum of the first distance values of all of the first spatial points in the first observation direction, or may be a ratio of the weighted first distance value of the first spatial point to a weighted sum of the first distance values of all of the first spatial points in the first observation direction, or may be a mean of the first distance values of all of the first spatial points in the first observation direction. During implementation, those skilled in the art may determine the manner of the first weight value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00115] In some implementation modes, the first weight value may be calculated by using the following formula (3-2):

$$w(t) = \frac{\phi_s(f(p(t)))}{\int_0^\infty \phi_s(f(p(u))) du} \quad (3-2)$$

where ϕ_s is logical density distribution, f is the distance neural network, and $p(t)$ and $p(u)$ refer to first spatial points.

[00116] At S3432, at least one first pixel point in the first observation image is updated

based on the first weight value and the first color value of each first spatial point, to obtain the second observation image.

[00117] In some implementation modes, the updating of the first pixel point may include, but is not limited to, updating the color value of the first pixel point.

[00118] In some implementation modes, the color value of the first pixel point may include, but is not limited to, a cumulative value of each first weight value and each first color value, a cumulative value of each weighted first weight value and each weighted first color value, a mean/variance of each first color value, a mean/variance of each weighted first color value, a product between a cumulative value of each first weight value and a distance gradient value of each first distance value and the first color value, or a product between a weighted cumulative value of each first weight value and a distance gradient value of each first distance value and the weighted first color value. During implementation, those skilled in the art may determine the manner of the color value of the first pixel point according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00119] In some implementation modes, the color value of each first pixel point may be calculated by using the following formula (3-3):

$$C_{shade}(o, v) = A + D \times n(o, v) \cdot l \cdot C_c(o, v) \quad (3-3)$$

where $A \sim \mathcal{U}(0, 0.2)$ is randomly sampled from uniform distribution; D is $1-A$, which is a diffuse reflection value; l denotes a sampling light direction; $C_c(o, v)$ denotes the first color value; $n(o, v)$ denotes the distance gradient value between each first spatial point $p(t)$ and a first intersection point corresponding to the first observation direction v .

[00120] In some implementation modes, the distance gradient value may be calculated by using the following formula (3-4):

$$n(o, v) = \int_0^\infty w(t) \nabla f(p(t)) dt \quad (3-4)$$

where ∇f denotes a gradient of a function f , and $w(t)$ denotes the first weight value.

[00121] In some implementation modes, S35 includes S351 to S353.

[00122] At S351, for each first observation image, a first loss sub-value is determined based on the second prompt information and the first observation image.

[00123] Here, the first loss sub-value may be determined based on parameters such as the first weight value and the distance gradient value of each first spatial point, as well as the second prompt information.

[00124] In some implementation modes, S351 includes S3511 to S3512.

[00125] At S3511, a third observation image is determined based on the first observation image.

[00126] In some implementation modes, each first pixel point in the first observation image may be updated based on the first weight value and the distance gradient value of each first spatial point, and the updated first observation image is used as the third observation image.

[00127] In some implementation modes, S3511 includes S361 to S362.

[00128] At S361, for each first observation direction, the distance gradient value of each first spatial point is determined based on the first distance value corresponding to each first spatial point in the first observation direction.

[00129] Here, the distance gradient value may be calculated by using the above formula (3-4).

[00130] At S362, at least one first pixel point in the first observation image is updated based on the first weight value of each first spatial point and the distance gradient value,

to obtain the third observation image.

[00131] In some implementation modes, the updating of the first pixel point may include, but is not limited to, updating the color value of the first pixel point.

[00132] In some implementation modes, the color value of the first pixel point may include, but is not limited to, a cumulative value of each first weight value and each distance gradient value, and a cumulative value of each weighted first weight value and each weighted distance gradient value. During implementation, those skilled in the art may determine the manner of the color value of the first pixel point according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00133] In some implementation modes, the color value of each first pixel point may be calculated by using the following formula (3-5):

$$C_{gray}(o, v) = A + D \times n(o, v) \cdot l \quad (3-5)$$

where $A \sim \mathcal{U}(0, 0.2)$ is randomly sampled from uniform distribution; D is $1-A$, which is a diffuse reflection value; l denotes the sampling light direction; $n(o, v)$ denotes the distance gradient value between each first spatial point $p(t)$ and the first intersection point corresponding to the first observation direction v .

[00134] At S3512, the first loss sub-value is determined based on the second prompt information and the third observation image.

[00135] In some implementation modes, the first loss sub-value may be a similarity between the text feature of the second prompt information and the image feature of the third observation image, or may be a similarity between the weighted text feature of the second prompt information and the weighted image feature of the third observation image are weighted. During implementation, those skilled in the art may determine the manner of the first loss sub-value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00136] In some implementation modes, the first loss sub-value may be calculated by using the following formula (3-6):

$$\mathcal{L}_{clip}^g = \mathcal{L}_{clip}(I_g, t_{app}) \quad (3-6)$$

where I_g denotes the third observation image; t_{app} denotes the second prompt information; and $\mathcal{L}_{clip}$ denotes a similarity between the third observation image and the second prompt information, which may be calculated by using the above formula (2-1) during implementation.

[00137] At S352, for each second observation image, a second loss sub-value is determined based on the second prompt information and the second observation image.

[00138] In some implementation modes, the second loss sub-value may be a similarity between the text feature of the second prompt information and the image feature of the second observation image, or may be a similarity between the weighted text feature of the second prompt information and the weighted image feature of the second observation image are weighted. During implementation, those skilled in the art may determine the manner of the second loss sub-value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00139] In some implementation modes, the second loss sub-value may be calculated by using the following formula (3-7):

$$\mathcal{L}_{clip}^c = \mathcal{L}_{clip}(I_c, t_{app}) \quad (3-7)$$

where I_c denotes the second observation image; t_{app} denotes the second prompt information; and $\mathcal{L}_{clip}$ denotes a similarity between the second observation image and the second prompt information, which may be calculated by using the above formula (2-1)

during implementation.

[00140] At S353, the first target loss value is determined based on each first loss sub-value and each second loss sub-value.

[00141] Here, the first target loss value may include, but is not limited to, a sum of each first loss sub-value and each second loss sub-value, and a sum of each weighted first loss sub-value and each weighted second loss sub-value. During implementation, those skilled in the art may determine the manner of the first target loss value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00142] In some implementation modes, the first target loss value may be calculated by using the following formula (3-8):

$$\mathcal{L}_2 = \lambda_3 \mathcal{L}_{clip}^c + \lambda_4 \mathcal{L}_{clip}^g \quad (3-8)$$

where $\mathcal{L}_{clip}^g$ denotes the first loss sub-value, and $\mathcal{L}_{clip}^c$ denotes the second loss sub-value.

[00143] In the implementation mode of the present disclosure, the first target loss value is determined based on the first loss sub-value and the second loss sub-value. In this way, the accuracy of the first target loss value can be enhanced, so that the virtual object matching the appearance prompt information can be accurately obtained.

[00144] In some implementation modes, the first target loss value further includes a first reconstruction loss value in reconstruction of the first virtual model. The first reconstruction loss value may include, but is not limited to, a color reconstruction loss value, a regularization loss value, a mask loss value and the like.

[00145] In some implementation modes, the first reconstruction loss value may include, but is not limited to, a sum of the color reconstruction loss value, the regularization loss value and the mask loss value, or a sum of the weighted color reconstruction loss value, the weighted regularization loss value and the weighted mask loss value. During implementation, those skilled in the art may determine the manner of the first reconstruction loss value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00146] In some implementation modes, the first reconstruction loss value may be calculated by using the following formula (3-9):

$$\mathcal{L}_1 = \mathcal{L}_{color} + \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{mask} \quad (3-9)$$

where $\mathcal{L}_{color}$ is a color reconstruction loss value between a rendered image and multi-view rendering $\{I_{M_t}^i\}$. $\mathcal{L}_{reg}$ is a regular loss value of regularization $f(p)$, and $\mathcal{L}_{mask}$ is a mask loss.

[00147] In some implementation modes, the color reconstruction loss value may be a mean between a predicted color value and a color value, which is actually measured, of a ray corresponding to each observation direction, or a mean between a weighted predicted color value and a weighted, which is actually measured, of the ray corresponding to each observation direction. During implementation, those skilled in the art may determine the manner of the color reconstruction loss value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00148] In some implementation modes, $\mathcal{L}_{color}$ may be calculated according to the following formula (3-10):

$$\mathcal{L}_{color} = \frac{1}{m} \sum_k \mathcal{R}(\hat{C}_k, C_k) \quad (3-10)$$

where $\mathcal{R}$ is a function, $\hat{C}_k$ denotes the predicted color value for the k-th ray, and C_k is the color value which is actually measured.

[00149] In some implementation modes, the predicted color value may include, but is

not limited to, a cumulative value between each first weight value and each second color value, a cumulative value between each weighted first weight value and each weighted second color value, a mean or variance of each second color value, and a mean or variance of each weighted second color value. The second color value may be determined based on a second color network of the first neural network. For example, each second color value $c(p(t), v)$ is determined based on $c(p)$ in N' . During implementation, those skilled in the art may determine the manner of the predicted color value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00150] In some implementation modes, $\hat{C}_k$ may be calculated according to the following formula (3-11):

$$C(o, v) = \int_0^\infty w(t)c(p(t), v)dt \quad (3-11)$$

where o is the first observation point; v is the first observation direction; $p(t)$ refers to the first spatial point; $c(p(t), v)$ denotes the second color value at the point $p(t)$ outputted at an observation view angle of v ; and $w(t)$ denotes the first weight value of the first spatial point, which may be calculated by using the above formula (3-2).

[00151] In some implementation modes, the regularization loss value may be a mean of a gradient value at each sampling point along the ray corresponding to each observation direction, or may be a mean of a weighted gradient value at each sampling point along the ray corresponding to each observation direction. During implementation, those skilled in the art may determine the manner of the regularization loss value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00152] In some implementation modes, $\mathcal{L}_{reg}$ may be calculated according to the following formula (3-12):

$$\mathcal{L}_{reg} = \frac{1}{nm} \sum_{k,i} (|\nabla f(\hat{P}_{k,i})| - 1)^2 \quad (3-12)$$

where ∇f denotes a gradient of the function f ; $\nabla f(\hat{P}_{k,i})$ denotes the gradient value of the function f at the i -th sampling point along the k -th ray; k denotes the k -th ray; i denotes the i -th sampling point on the k -th ray; n denotes a sample size; and m denotes a sampling batch size.

[00153] In some implementation modes, the mask loss value may be a difference between weights of all of the sampling points along the ray corresponding to each observation direction and a preset mask value, a difference between weighted weights of all of the sampling points along the ray corresponding to each observation direction and a weighted preset mask value, or the square of the difference between the weights of all of the sampling points along the ray corresponding to each observation direction and the preset mask value. During implementation, those skilled in the art may determine the manner of the mask loss value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00154] In some implementation modes, $\mathcal{L}_{mask}$ may be calculated according to the following formula (3-13):

$$\mathcal{L}_{mask} = BCE(M_k, \hat{O}_k) \quad (3-13)$$

where $\hat{O}_k$ represents the sum of the weights of all of the sampling points along the k -th ray; $M_k \in \{0, 1\}$, which is an optional mask value.

[00155] In some implementation modes, $\hat{O}_k$ may be calculated according to the following formula (3-14):

$$\begin{cases} \alpha_{k,i} = \max\left(\frac{\phi_s(f(p(t_i))) - \phi_s(f(p(t_{i+1})))}{\phi_s(f(p(t_i)))}, 0\right); \\ T_{k,i} = \phi_s(f(p(t_i))); \\ \hat{O}_k = \sum_{i=1}^n T_{k,i} \alpha_{k,i}; \end{cases} \quad (3-14)$$

where ϕ_s is logical density distribution, f is the distance neural network, and $p(t_i)$ denotes the i -th sampling point.

- [00156] In the implementation mode of the present disclosure, the first target loss value is determined based on the first reconstruction loss value, each first loss sub-value, and each second loss sub-value. In this way, the accuracy of the first target loss value can be enhanced, so that the virtual object matching the appearance prompt information can be obtained more accurately.
- [00157] In some implementation modes, S36 includes S371 to S372.
- [00158] At S371, augmentation is performed on each second observation image.
- [00159] Here, the augmentation may include, but is not limited to, performing at least one of occlusion, scaling, cropping, resizing, filling, flipping, color dithering, grayscale processing, Gaussian blur, or random erasure on the background of the second observation image. During implementation, those skilled in the art may implement proper augmentation on a first sub-image according to actual situations to obtain a second sub-image, which is not limited in the embodiments of the present disclosure.
- [00160] At S372, the first virtual model is rendered based on each augmented second observation image, to generate the virtual object.
- [00161] In the embodiment of the present disclosure, augmentation is performed on the second observation image, which helps the fourth neural network to focus more on a foreground object and prevents the generation of a randomly floating volume, so that robustness brought during optimization can be further enhanced.
- [00162] FIG. 4 is a schematic diagram of an implementation process of a method for generating a target animation of a virtual object according to an embodiment of the present disclosure. As shown in FIG. 4, the method includes S41 to S45.
- [00163] At S41, first prompt information and second prompt information are acquired. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.
- [00164] At S42, a first virtual model is determined based on the first prompt information.
- [00165] At S43, the virtual object is generated based on the second prompt information and the first virtual model.
- [00166] S41 to S43 correspond to S11 to S13, and may refer to specific implementation modes of S11 to S13 during implementation.
- [00167] At S44, third prompt information is acquired, the third prompt information representing animation information of the virtual object.
- [00168] Here, the third prompt information may be text prompt information, voice prompt information, and the like. For example, the third prompt information may be the text prompt information that represents animations such as running, sitting, and anger. For another example, the third prompt information may be the voice prompt information that represents animations such as rising arms, playing balls, and drinking water. The

virtual object may be a two-dimensional model, or may be a three-dimensional model. For example, the virtual object may be an Iron Man model.

- [00169] In some implementation modes, the manner of acquiring the third prompt information may be determined according to a practical application scenario. For example, the third prompt information may be acquired based on information inputted by a user through an input device of electronic device. For another example, the third prompt information may be acquired based on information transmitted by other electronic devices.
- [00170] In some implementation modes, the first prompt information, the second prompt information, and the third prompt information may be acquired simultaneously or separately.
- [00171] For example, the user inputs "a tall and strong iron man is running" by using a keyboard, the electronic device parses to obtain the first prompt information of "tall and strong man", the second prompt information of "iron man", and the third prompt information of "running".
- [00172] For another example, the user first input voice information of "fat spider man". In this case, the electronic device parses to obtain the first prompt information of "fat man", the second prompt information of "spider man"; then the user inputs "drinking water" through the keyboard, and the third prompt information is "drinking water". During implementation, those skilled in the art may determine the manner of acquiring the third prompt information according to actual requirements, which is not limited in the embodiments of the present disclosure.
- [00173] At S45, a target animation of the virtual object is generated based on the third prompt information.
- [00174] Here, the target animation is an animation matching the third prompt information. For example, the target animation may be an animation showing that the iron man is running. For another example, the target animation may be an animation of building a three-dimensional house.
- [00175] In some implementation modes, the target animation of the virtual object may be generated through a preset sixth neural network based on the third prompt information. The preset sixth neural network may be a neural network pre-trained based on a large number of training samples. The training samples may be animations. For example, the preset sixth neural network may be a Variational Auto-Encoder (VAE). During implementation, those skilled in the art may select the appropriate sixth neural network according to actual requirements, which is not limited in the embodiments of the present disclosure.
- [00176] In some implementation modes, the sixth neural network includes a motion encoder (E_{motion}), a reparameterization module, and a motion encoder (D_{motion}). Based on the trained sixth neural network, the target animation of the virtual object is generated.
- [00177] In the embodiment of the present disclosure, the first prompt information and the second prompt information are acquired, the first prompt information represents the body shape information of the virtual object, the second prompt information represents the appearance information of the virtual object; the first virtual model is determined based on the first prompt information; the virtual object is generated based on the second prompt information and the first virtual model; the third prompt information is acquired, the third prompt information represents the animation information of the virtual object; the target animation of the virtual object is generated based on the third prompt information. In this way, the target animation matching the animation prompt information

can be automatically and accurately generated based on animation prompt information, so that the threshold of making a digital animation can be reduced. In addition, there is no need to use the prior information and/or additional training data to train a model, so that overfitting to the prior information and/or the learning during model training can be reduced. Therefore, recognition capabilities on the prompt information that is not considered in the prior information and/or the animation prompt information that is not involved during model training can be enhanced, and a generalization capability for recognizing the animation prompt information can be improved.

[00178] In some implementation modes, a training process of the sixth neural network includes S471 to S474.

[00179] At S471, an animation sample is coded by using an encoder of the sixth neural network, to obtain a coded animation.

[00180] In some implementation modes, the encoder E_{motion} includes a projection layer (E_p), a position embedding layer (ϕ), several transformer encoder layers ($TransEncoder$), and an output layer (E_o). At this time, E_{motion} may be defined by using the following formula (4-1):

$$\mathcal{Y} = E_{motion}(\theta^+) = E_o(TransEncoder(\phi(E_p(\theta^+)))) \quad (4-1)$$

where θ^+ is the animation sample, E_p is a fully connected layer that enhances an original input to higher dimensions, and ϕ represents an position embedding operation. $TransEncoder$ includes a plurality of transformer encoder layers. The output layer E_o projects a feature to a low-dimensional space.

[00181] At S472, a random latent code is generated based on the coded animation by using the reparameterization module of the sixth neural network.

[00182] In some implementation modes, the reparameterization module is configured to produce Gaussian distribution, calculate a mean u and a standard deviation σ of Gaussian distribution respectively, and sample the random latent code $z_{motion} \sim \mathcal{N}(u, \sigma)$ under the assumption of Gaussian distribution. During implementation, u and σ may be generated by using the following formula (4-2):

$$u = R_u(\mathcal{Y}), \quad \sigma = R_\sigma(\mathcal{Y}) \quad (4-2)$$

where R_u and R_σ are fully connected layers.

[00183] At S473, the target animation based on the random latent code by using the decoder of the sixth neural network.

[00184] In some implementation modes, the reparameterization module is used to generate the random latent code z_{motion} . z_{motion} is decoded by D_{motion} to obtain the target animation θ^* . At this time, D_{motion} may be defined by using the following formula (4-3):

$$\theta^* = D_{motion}(z_{motion}) = D_o(TransDecoder(D_p(z_{motion}), \phi(0))) \quad (4-3)$$

where D_p and D_o are fully connected layers, and 0 is a zero vector. $TransDecoder$ includes several transformer decoder layers.

[00185] At S474, a network parameter of the sixth neural network is updated for at least one time based on the animation sample and the target animation, to obtain the trained sixth neural network.

[00186] In some implementation modes, a third loss value may be calculated based on the animation sample and the target animation. The network parameter of the sixth neural network is updated for at least one time based on the third loss value and a third preset condition. In the case that the third loss value does not satisfy the third preset condition, the network parameter of the sixth neural network is updated, and the updated sixth

neural network is trained by using the animation sample. In the case that the third loss value satisfies the third preset condition, the sixth neural network is used as the trained sixth neural network.

[00187] Here, the third preset condition may include, but is not limited to, the loss value being less than a threshold, loss value change convergence, and the like. During implementation, those skilled in the art may independently determine the third preset condition according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00188] In some implementation modes, the manner of updating the network parameter of the sixth neural network may be determined according to actual situations, and may include, but is not limited to, at least one of a gradient descent method or a Newton momentum method, which is not limited here.

[00189] In some implementation modes, the third loss value may include, but is not limited to, a random latent code loss value, and a second reconstruction loss value.

[00190] In some implementation modes, the third loss value may be a sum of the random latent code loss value and the second reconstruction loss value, or a sum of the weighted random latent code loss value and the weighted second reconstruction loss value. During implementation, those skilled in the art may determine the manner of the third loss value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00191] In some implementation modes, the third loss value $\mathcal{L}_{mVAE}$ may be calculated according to the following formula (4-4):

$$\mathcal{L}_{mVAE} = \lambda_5 \cdot \mathcal{L}_{KL} + \mathcal{L}_{recon} \quad (4-4)$$

where λ_5 is a hyper-parameter. $\mathcal{L}_{KL}$ is the random latent code loss value of the random latent code randomly generated in Gaussian distribution. $\mathcal{L}_{recon}$ uses an error of mean square between θ^+ and θ^* to represent the second reconstruction loss value.

[00192] In some implementation modes, the random latent code loss value $\mathcal{L}_{KL}$ may be calculated according to the following formula (4-5):

$$\mathcal{L}_{KL} = \log(\sigma) + \frac{1+u^2}{2\sigma^2} - \frac{1}{2} \quad (4-5)$$

where u is a mean of the above Gaussian distribution, and σ is a standard deviation of the above Gaussian distribution.

[00193] In some implementation modes, the second reconstruction loss value $\mathcal{L}_{recon}$ may be calculated according to the following formula (4-6):

$$\mathcal{L}_{recon} = \text{mean}((\theta^+ - \theta^*)^2) \quad (4-6)$$

where θ^+ is the animation sample, and θ^* is the target animation.

[00194] In the implementation mode of the present disclosure, the entire network structure of the sixth neural network is trained based on a large number of the animation samples, so that the robustness of the sixth neural network and the performance of the sixth neural network can both be enhanced. Therefore, the target animation matching the animation prompt information can be obtained based on the trained sixth neural network.

[00195] In some implementation modes, S45 includes S451 to S453.

[00196] At S451, a first candidate pose image set is determined based on the third prompt information.

[00197] Here, the first candidate pose image set includes at least one pose image. Each pose image matches the third prompt information.

[00198] In some implementation modes, S451 includes S4511 to S4512.

[00199] At S4511, a preset second correspondence is queried by using the third prompt

information, to obtain at least one candidate pose image matching the third prompt information.

[00200] Here, the second correspondence represents a correspondence between the third prompt information and the pose image.

[00201] In some implementation modes, a mapping table of the correspondence between the third prompt information and the pose image may be established in advance, and the mapping table is stored in the electronic device or other terminals. In the case that the mapping table is stored in the electronic device, the electronic device determines at least one pose image matching the third prompt information in the mapping table according to the third prompt information. In the case that the mapping table is stored in other terminals, the electronic device transmits the third prompt information to the other terminals, so that the other terminals determine at least one pose image matching the third prompt information in the mapping table according to the third prompt information and return the at least one pose image to the electronic device.

[00202] For example, fuzzy matching may be performed from the second correspondence based on the third prompt information, to obtain the at least one pose image. For another example, exact matching may be performed from the second correspondence based on the third prompt information, to obtain the at least one pose image.

[00203] In some implementation modes, the second correspondence may be established by using the fourth neural network. For example, the second correspondence is established through CLIP.

[00204] At S4512, the first candidate pose image set is determined based on each candidate pose image.

[00205] Here, each candidate pose image is used as a candidate pose image in the first candidate pose image set.

[00206] At S452, a first animation is generated based on a preset first parameter.

[00207] Here, the first parameter may be preset by the user, or may be randomly generated. In some implementation modes, the first parameter may be initialized to zero, or may be obtained based on random sampling from standard Gaussian distribution.

[00208] In some implementation modes, the first animation may be generated based on the first parameter by using the decoder of the trained sixth neural network.

[00209] At S453, the first parameter is updated based on the first candidate pose image set and the first animation, to generate the target animation of the virtual object.

[00210] In some implementation modes, a second target loss value may be determined based on the first candidate pose image set and the first animation, and whether the first parameter needs to be updated is determined based on the second target loss value. For example, the second target loss value is compared with a threshold. In the case that the second target loss value is greater than the threshold, the first parameter is updated. In the case that the second target loss value is not greater than the threshold, the first animation is determined as the target animation. For another example, the second target loss value is compared with the last second target loss value. In the case that the second target loss value is greater than the last second target loss value, the first parameter is updated. In the case that the second target loss value basically equals the last second target loss value, the first animation is determined as the target animation.

[00211] In some implementation modes, S453 includes S4531 to S4532.

[00212] At S4531, the second target loss value is determined based on the first candidate pose image set and the first animation.

- [00213] In some implementation modes, the second target loss value may be determined based on each pose image in the first candidate pose image set, each frame image in the first animation, two adjacent frame images in the first animation, and the third prompt information.
- [00214] At S4532, the first animation is used as the target animation in the case that the second target loss value satisfies a second preset condition.
- [00215] Here, the second preset condition may include, but is not limited to, the second target loss value being less than a threshold, second target loss value change convergence, and the like. During implementation, those skilled in the art may independently determine the second preset condition according to actual requirements, which is not limited in the embodiments of the present disclosure.
- [00216] In some implementation modes, S453 includes S4533 to S4535.
- [00217] At S4533, in the case that the second target loss value does not satisfy the second preset condition, the first parameter is updated to obtain the updated first parameter.
- [00218] In some implementation modes, the first parameter may be updated based on the reparameterization module of the sixth neural network. In some implementation modes, the first parameter may be updated based on the first animation.
- [00219] At S4534, the updated first animation is generated based on the updated first parameter.
- [00220] At S4535, the second target loss value is determined based on the first candidate pose image set and the updated first animation.
- [00221] S4534 to S4535 correspond to S4531 to S4532, and may refer to specific implementation modes of S4531 to S4532 during implementation.
- [00222] In the implementation of the disclosure, in the case that the second target loss value does not satisfy the second preset condition, next updating may be performed on the first parameter, and a new first animation is generated based on the first parameter after the next updating, so that the virtual target animation can be further accurately obtained through continuous iterative updating.
- [00223] In some implementation modes, S4531 includes S461 to S464.
- [00224] At S461, a third loss sub-value is determined based on the first candidate pose image set and each frame image in the first animation.
- [00225] In some implementation modes, the third loss sub-value may be a distance between each pose image and each frame image in the first animation, a similarity between each pose image and each frame image in the first animation, a distance between each weighted pose image and each weighted frame image in the first animation, or a similarity between each weighted pose image and weighted each frame image in the first animation. During implementation, those skilled in the art may determine the manner of the third loss sub-value according to actual requirements, which is not limited in the embodiments of the present disclosure.
- [00226] In some implementation modes, the third loss sub-value may be calculated by using the following formula (4-7):

$$\mathcal{L}_{pose} = \sum_{i=1}^k \lambda_{pose}(i) \min_j \{\|\theta_i - \theta_j\|\} \quad (4-7)$$

where λ_{pose} denotes a hyper-parameter coefficient, θ_i denotes the i -th pose image in the first candidate pose image set, $i \in \{1, 2, \dots, k\}$, k denotes the total number of pose images in the first candidate pose image set, θ_j denotes the j -th frame image in the first animation, $j \in \{1, 2, \dots, L\}$, and L represents the total number of frames of the first

animation.

[00227] In some implementation modes, the hyper-parameter coefficient may be calculated by using the following formula (4-8):

$$\lambda_{pose}(i) = 1 - \frac{i-1}{k} \quad (4-8)$$

where $i \in \{1, 2, \dots, k\}$, k denotes the total number of the pose images in the first candidate pose image set.

[00228] At S462, a fourth loss sub-value is determined based on each two adjacent frame images in the first animation.

[00229] In some implementation modes, the fourth loss sub-value may be a distance between each two adjacent frame images in the first animation, a similarity between each two adjacent frame images, a distance between each two adjacent weighted frame images, or a similarity between each two adjacent weighted frame images. During implementation, those skilled in the art may determine the manner of the fourth loss sub-value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00230] In some implementation modes, the fourth loss sub-value may be calculated by using the following formula (4-9):

$$\mathcal{L}_{delta} = -\sum_{i=1}^{L-1} \|\theta_i - \theta_{i+1}\| \quad (4-9)$$

where θ_i denotes the i -th frame image in the first animation, θ_{i+1} denotes the $(i+1)$ -th frame image, $i \in \{1, 2, \dots, L\}$, and L denotes the total number of frames of the first animation.

[00231] At S463, a fifth loss sub-value is determined based on the third prompt information and each frame image in the first animation.

[00232] In some implementation modes, the fifth loss sub-value may be a distance between the third prompt information and each frame image in the first animation, a similarity between the third prompt information and each frame image in the first animation, a distance between the weighted third prompt information and each weighted frame image in the first animation, or a similarity between the weighted third prompt information and each weighted frame image in the first animation. During implementation, those skilled in the art may determine the manner of the fifth loss sub-value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00233] In some implementation modes, the fifth loss sub-value may be calculated by using the following formula (4-10):

$$\mathcal{L}_{clip}^m = \sum_{i=1}^L \lambda_{clip}(i) \cdot s_i \quad (4-10)$$

where $\lambda_{clip}(\cdot)$ denotes a hyper-parameter function, L denotes the total number of frames of the first animation, and s_i denotes a similarity score between θ_i and the third prompt information.

[00234] In some implementation modes, $\lambda_{clip}(i)$ may be calculated according to the following formula (4-11):

$$\lambda_{clip}(i) = \frac{i}{L} \quad (4-11)$$

where L denotes the total number of frames of the first animation.

[00235] In some implementation modes, s_i may be calculated according to the following formula (4-12):

$$s_i = 1 - \text{norm}(E_I(\mathcal{R}(\theta_i))) \cdot \text{norm}(E_T(t_{motion})) \quad (4-12)$$

where $(\cdot)$ denotes a cosine distance; and norm is a function used for regularizing an

inputted vector, t_{motion} denotes the third prompt information, and θ_i denotes the i-th frame image in the first animation.

[00236] At S464, the second target loss value is determined based on the third loss sub-value, the fourth loss sub-value, and the fifth loss sub-value.

[00237] Here, the second target loss value may include, but is not limited to, a sum of the third loss sub-value, the fourth loss sub-value and the fifth loss sub-value, or a sum of the weighted third loss sub-value, the weighted fourth loss sub-value, and the weighted fifth loss sub-value. During implementation, those skilled in the art may determine the manner of the second target loss value according to actual requirements, which is not limited in the embodiments of the present disclosure.

[00238] In some implementation modes, the second target loss value may be calculated by using the following formula (4-13):

$$\mathcal{L}_{motion} = \mathcal{L}_{pose} + \lambda_6 \mathcal{L}_{delta} + \lambda_7 \mathcal{L}_{clip}^m \quad (4-13)$$

where $\mathcal{L}_{pose}$ denotes the third loss sub-value, $\mathcal{L}_{delta}$ denotes the fourth loss sub-value, and $\mathcal{L}_{clip}^m$ denotes the fifth loss sub-value.

[00239] In the implementation mode of the present disclosure, the second target loss value is determined based on the third loss sub-value, the fourth loss sub-value and the fifth loss sub-value. In this way, the accuracy of the second target loss value can be enhanced, so that the target animation matching the animation prompt information can be accurately obtained.

[00240] FIG. 5A is a schematic diagram of a composition structure of a system 50 for generating a virtual object according to an embodiment of the present disclosure. As shown in FIG. 5A, the system 50 for generating a virtual object includes a first acquisition module 51, a body shape model generation module 52, and a virtual object generation module 53.

[00241] The first acquisition module 51 is configured to acquire first prompt information and second prompt information. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.

[00242] The body shape model generation module 52 is configured to determine a first virtual model based on the first prompt information.

[00243] The virtual object generation module 53 is configured to generate the virtual object based on the second prompt information and the first virtual model.

[00244] FIG. 5B is a schematic diagram of a system for generating a virtual object according to an embodiment of the present disclosure. As shown in FIG. 5B, the system for generating a virtual object determines the first virtual model M_t based on first prompt information 501, reconstructs the first virtual model M_t based on second prompt information 502 and the first neural network N' , to generate the virtual object 503.

[00245] In some implementation modes, the body shape model generation module 52 is further configured to: determine a target body shape image matching the first prompt information based on a preset first correspondence, the first correspondence representing a correspondence between the first prompt information and a body shape image; and perform three-dimensional modeling based on the target body shape image, to generate the first virtual model.

[00246] In some implementation modes, the body shape model generation module 52 is further configured to: respectively determine a first distance between each body shape image in the first correspondence and a set standard body shape image; determine a

second distance between the first prompt information and standard prompt information corresponding to the standard body shape image; and determine the target body shape image based on the first distance and the second distance.

[00247] FIG. 5C is a schematic diagram of determination of a target body shape image according to an embodiment of the present disclosure. As shown in FIG. 5C, the system first trains a shape VAE, establishes a first correspondence 504 based on the shape VAE, and determines a target body shape image from each body shape image 507 of the first correspondence 504 based on the first prompt information 501, standard prompt information 505, and a standard body shape image 506.

[00248] In some implementation modes, the virtual object generation module 53 is further configured to: determine a first observation image of the first virtual model at a first observation view angle for each of the at least one first observation view angle; for each first observation image, update the first observation image for at least one time by using a preset first neural network so as to obtain a second observation image; determine a first target loss value based on each second observation image and the second prompt information; and in the case that the first target loss value satisfies a first preset condition, render the first virtual model based on each second observation image, to generate the virtual object.

[00249] In some implementation modes, the virtual object generation module 53 is further configured to: in the case that the first target loss value does not satisfy the first preset condition, update a network parameter of the first neural network based on the first target loss value; update each first observation image for at least one time by using the updated first neural network, to obtain the updated second observation image; and determine an updated first target loss value based on each updated second observation image and the second prompt information.

[00250] In some implementation modes, the first observation view angle includes a first observation point and at least one first observation direction corresponding to the first observation point. The first observation image includes at least one first pixel point. Each first pixel point corresponds to one first intersection point on the first virtual model. Each first intersection point represents the first intersection point between the first virtual model and a line coming from the first observation point and extending along one first observation direction. The virtual object generation module 53 is further configured to: determine at least one first spatial point corresponding to each first observation direction based on the first observation point and each first observation direction; for each first spatial point of each first observation direction, determine, by using the preset first neural network, a first distance value between the first intersection point corresponding to the first observation direction where the first spatial point is located and the first spatial point, and determine a first color value obtained by observing the first virtual model according to the first observation direction at the first spatial point; and update the first observation image for at least one time based on the first distance value and the first color value corresponding to each first spatial point, to obtain the second observation image.

[00251] In some implementation modes, the virtual object generation module 53 is further configured to: for each first observation direction, determine a first weight value of each first spatial point based on the first distance value corresponding to each first spatial point in the first observation direction; and update the at least one first pixel point in the first observation image based on the first weight value and the first color value of each first spatial point, to obtain the second observation image.

[00252] In some implementation modes, the virtual object generation module 53 is

further configured to: for each first observation image, determine a first loss sub-value based on the second prompt information and the first observation image; for each second observation image, determine a second loss sub-value based on the second prompt information and the second observation image; and determine the first target loss value based on each first loss sub-value and each second loss sub-value.

[00253] In some implementation modes, the virtual object generation module 53 is further configured to: determine a third observation image based on the first observation image; and determine the first loss sub-value based on the second prompt information and the third observation image.

[00254] In some implementation modes, the virtual object generation module 53 is further configured to: for each first observation direction, determine a distance gradient value of each first spatial point based on the first distance value corresponding to each first spatial point in the first observation direction; and update the at least one first pixel point in the first observation image based on the first weight value of each first spatial point and the distance gradient value, to obtain the third observation image.

[00255] FIG. 5D is a schematic diagram of a process for generating a virtual object according to an embodiment of the present disclosure. As shown in FIG. 5D, a process for generating the virtual object includes: the first virtual model is reconstructed by using $f(p)$ and $c(p)$ in the first neural network N' , and texture and pattern rendering is performed on the reconstructed model by using $f(p)$ and $c_c(p)$; the network parameter in N' is optimized based on the second prompt information; the virtual object is generated based on the optimized N' ; at least one observation view angle is determined based on at least one camera 508, each observation view angle including a first observation point o and at least one first observation direction v ; a first observation image of the first virtual model at each observation view angle and at least one first spatial point corresponding to each observation direction v are determined; a first distance value $f(p(t))$ of each first spatial point is obtained based on the network $f(p)$; obtaining a first color value $c_c(p(t))$ of each first spatial point is obtained based on the network $c_c(p)$; a second color value $c(p(t))$ of each first spatial point is obtained based on the network $c(p)$; a first weight value $w(t)$ of each first spatial point and a distance gradient value $\nabla f(p(t))$ are obtained based on each first distance value; each first pixel point of the first observation image is updated based on $c(p(t))$ and $w(t)$, to generate an image $I_t^{i'}$; each first pixel point of the first observation image is updated based on $w(t)$ and $\nabla f(p(t))$, to generate an image I_n ; each first pixel point of the first observation image is updated based on I_n and a ray of light l , to generate an image I_g ; each first pixel point of the first observation image is updated based on I_n , the ray of light l , and $c_c(p(t))$, to generate an image I_c ; a first reconstruction loss value $\mathcal{L}_1$ is determined based on the image $I_t^{i'}$; a first loss sub-value $\mathcal{L}_{\text{clip}}^g$ is determined based on the image I_g and the second prompt information 502; a second loss sub-value $\mathcal{L}_{\text{clip}}^c$ is determined based on the image I_c and the second prompt information 502; and a first target loss value $\mathcal{L}_2$ is determined based on the first reconstruction loss value, the first loss sub-value, and the second loss sub-value.

[00256] In some implementation modes, the system for generating a virtual object further includes a second acquisition module and a target animation generation module.

[00257] The second acquisition module is configured to acquire third prompt information. The third prompt information represents animation information of the virtual object.

[00258] The target animation generation module is configured to generate a target

animation of the virtual object based on the third prompt information.

[00259] FIG. 5E is a schematic diagram of generation of a target animation of a virtual object according to an embodiment of the present disclosure. As shown in FIG. 5E, the system performs animation initialization on the virtual object generated based on N' in advance, to generate $M(\theta)$. A first candidate pose image set 509 is generated based on the third prompt information 508. A first animation 510 is generated based on the first candidate pose image set 509. The target animation 511 of the virtual object is generated based on $M(\theta)$ and the first animation.

[00260] In some implementation modes, the target animation generation module is further configured to: determine a first candidate pose image set based on the third prompt information; generate a first animation based on a preset first parameter; and determine the target animation of the virtual object based on the first candidate pose image set and the first animation.

[00261] In some implementation modes, the target animation generation module is further configured to: a preset second correspondence is queried by using the third prompt information, to obtain at least one candidate pose image matching the third prompt information, the second correspondence representing a correspondence between the third prompt information and the pose image; and determine the first candidate pose image set based on each candidate pose image.

[00262] FIG. 5F is a schematic diagram of determination of a first candidate pose image according to an embodiment of the present disclosure. As shown in FIG. 5F, the system first trains a pose VAE, constructs a second correspondence 512 based on the pose VAE, and determines a first candidate pose image from each pose image $R(B\theta[i])$ of the second correspondence 512 based on the third prompt information 508.

[00263] In some implementation modes, the target animation generation module is further configured to: determine a second target loss value based on the first candidate pose image set and the first animation; and use the first animation as the target animation in the case that the second target loss value satisfies a second preset condition.

[00264] In some implementation modes, the target animation generation module is further configured to: in the case that the second target loss value does not satisfy the second preset condition, update the first parameter, to obtain the updated first parameter; generate the updated first animation based on the updated first parameter; and determine the second target loss value based on the first candidate pose image set and the updated first animation.

[00265] In some implementation modes, the target animation generation module is further configured to: determine a third loss sub-value based on the first candidate pose image set and each frame image in the first animation; determine a fourth loss sub-value based on each two adjacent frame images in the first animation; determine a fifth loss sub-value based on the third prompt information and each frame image in the first animation; and determine the second target loss value based on the third loss sub-value, the fourth loss sub-value, and the fifth loss sub-value.

[00266] FIG. 5G is a schematic diagram of determination of a second target loss value according to an embodiment of the present disclosure. As shown in FIG. 5G, the system uses a decoder D_{motion} of the pre-trained VAE to decode the first parameter, to generate the first animation 510; determines the third loss sub-value $\mathcal{L}_{pose}$ based on each pose image θ_i in the first candidate pose image set and each frame image θ_j in the first animation 510; determines the fourth loss sub-value $\mathcal{L}_{delta}$ based on each two adjacent

frame images θ_i and θ_{i+1} in the first animation 510; determines the fifth loss sub-value $\mathcal{L}_{clip}^m$ based on the third prompt information 508 and each frame image θ_i in the first animation 510; determines the second target loss value $\mathcal{L}_{motion}$ based on the third loss sub-value $\mathcal{L}_{pose}$, the fourth loss sub-value $\mathcal{L}_{delta}$, and the fifth loss sub-value $\mathcal{L}_{clip}^m$.

[00267] FIG. 5H is a schematic diagram of a pre-trained VAE network structure according to an embodiment of the present disclosure. As shown in FIG. 5H, the VAE includes an encoder E_{motion} , a decoder D_{motion} , and a reparameterization module 512. The encoder E_{motion} includes a projection layer E_p , a position embedding layer ϕ , several transformer encoder layers $TransEncoder$, and an output layer E_o . The decoder D_{motion} includes a position embedding layer ϕ , a fully connected layer D_p , a fully connected layer D_o , and several transformer decoder layers $TransDecoder$. An action sequence sample 513 is inputted, and is transformed into an animation sample θ^+ . θ^+ is coded by using the encoder E_{motion} , to obtain a coded animation $\mathcal{Y}$. A mean u and a standard deviation σ of Gaussian distribution are generated after $\mathcal{Y}$ passes through the fully connected layers R_u and R_σ . The random latent code $z_{motion} \sim \mathcal{N}(u, \sigma)$ is sampled from the Gaussian distribution, and the random latent code z_{motion} is further decoded by the decoder D_{motion} to generate an animation θ^* . A loss value $\mathcal{L}_{KL}$ is determined based on the mean u and the standard deviation σ . A loss value $\mathcal{L}_{recon}$ is determined based on θ^+ and θ^* . $\mathcal{L}_{mVAE}$ is determined based on $\mathcal{L}_{KL}$ and $\mathcal{L}_{recon}$. The network parameter of VAE is updated for at least one time based on $\mathcal{L}_{mVAE}$. In the case that $\mathcal{L}_{mVAE}$ satisfies a preset condition, the VAE is used as a trained VAE.

[00268] Compared with methods in the related art, the method provided in the embodiments of the present disclosure has at least the following improvement.

[00269] 1) In the related art, a preset model is used to generate a virtual object based on the text prompt information. The preset model is trained based on a large amount of training data and cannot be automatically generated based on natural language. Alternatively, professional personnel use professional software and expensive devices to generate the virtual object based on the text prompt information. In the embodiments of the present disclosure, the power of a large-scale pre-trained model is combined, so that zero sample generation capability can be achieved. In addition, based on the body shape prompt information and the appearance prompt information, the virtual object that matches the body shape prompt information and the appearance prompt information can be automatically and accurately generated.

[00270] 2) In the related art, the preset model is used to generate a target animation based on the animation prompt information. The preset model is trained based on a large amount of training data and cannot be automatically generated based on natural language. Alternatively, professional personnel use professional software and expensive devices to generate the target animation based on the animation prompt information. In the embodiments of the present disclosure, the power of a large-scale pre-trained model is combined, so that zero sample generation capability can be achieved. In addition, based on the animation prompt information, the target animation matching the animation prompt information can be automatically and accurately generated.

[00271] The method provided in the embodiments of the present disclosure has at least the following beneficial effects: (1) based on the body shape prompt information, the appearance prompt information, and motion prompt information, the target animation of the virtual object that conforms to the three prompt information can be automatically generated; (2) by combining the power of the large-scale pre-trained model, zero sample

generation capability can be achieved, so that the virtual object with high-quality textures and geometries and highly consistent with the corresponding input natural language can be stably generated; (3) zero-shot learning is realized based on the VAE, animation synthesis is guided based on texts, which is highly consistent with the corresponding input natural language.

[00272] In order to better describe the beneficial effects of the embodiments of the present disclosure, experimental data of the method provided in the embodiments of the present disclosure and the method in the related art are compared and described below.

[00273] (1) Initialization of some parameters

[00274] In the shape VAE, both the encoder and decoder use a two-layer Multilayer Perceptron (MLP) with a 16-dimensional latent space. Then, 40,960 latent codes are sampled from the Gaussian distribution and clustered into 2,048 centers by K-Means of a synthesized codebook.

[00275] For NeuS, an SDF network uses a 6-layer MLP, and a color network uses a 4-layer MLP. For each ray, uniform sampling and importance sampling are respectively performed for 32 times. 30,000 optimization iterations are performed by using an Adam algorithm with a learning rate of 5×10^{-4} .

[00276] With regard to the generation of the first candidate pose image set, K-Means is used to acquire 4,096 cluster centroids from an AMASS dataset, and the first 5 poses are selected.

[00277] For the preset model VAE, the transformer encoder layer and the transformer decoder layer respectively include 4 transformer encoder layers and 4 transformer decoder layers. A frame length is 60. An Adam optimizer with a learning rate of 5×10^{-4} is used, and λ_5 is 10^{-5} . During the optimization of the first parameter, Adam performs 5,000 optimization iterations at a learning rate of 1×10^{-2} . λ_6 and λ_7 are respectively set to 10^{-3} and 5×10^{-4} .

[00278] (2) Baseline

[00279] Two baselines are designed: (i) a direct interpolation between candidate poses; and (ii) a VAE with unweighted $\mathcal{L}_{\text{pose}}$.

[00280] (3) Experimental result

[00281] 1) A comparison result of the virtual object generated by the method AvatarCLIP provided in the embodiments of the present disclosure and a baseline method

[00282] FIG. 5I is a schematic diagram of results of generating a virtual object based on a plurality of methods according to an embodiment of the present disclosure. As shown in FIG. 5I, the first column is the virtual object generated based on the baseline; the second column is the virtual object generated based on the baseline + background augmentation; the third column is the virtual object generated based on the baseline + texture-less rendering; the fourth column is based on the baseline + shading on textured rendering; the fifth column is the virtual object generated based on the AvatarCLIP. By comparing the first column with the second column in FIG. 5I, it can be learned that background augmentation has a great influence on texture generation; and without background augmentation, the textures tend to be very dark. By comparing the second column with the third column in FIG. 5I, geometry quality can be greatly improved by adding supervision over texture-less rendering. The geometry of "Ablation 2" has many random bumps which causes a surface to be noisy, while the geometry of "Ablation 3" is smooth and may produce detailed clothing folds. By comparing the third column with the fourth column in FIG. 5I, more uniform texture colors are produced by adding random shadows on texture rendering. Without human semantic awareness, the first four ablation

settings fail to generate correct faces for avatars. The last column further uses semantic aware prompt autmentation and has an optimal effect on generated faces, which is one of the most critical parts in generating the virtual object.

[00283] 2) A comparison result of M_t generated by the method AvatarCLIP provided in the embodiments of the present disclosure and a baseline method

[00284] FIG. 5J is a schematic diagram of results of generating a first virtual model based on a plurality of methods according to an embodiment of the present disclosure. As shown in FIG. 5J, a baseline (i) is to directly perform optimization on an SMPL β space, and a baseline (ii) is to directly optimize the first parameter in VAE. As shown in FIG. 5J, neither of the two optimization methods can generate M_t that is consistent with the described text, even opposite textual guidances (such as "thin" and "overweight") may generate a same M_t . In contrast, AvatarCLIP can robustly generate a reasonable body shape consistent with an input text.

[00285] 3) A comparison result of the virtual object generated by the method AvatarCLIP provided in the embodiments of the present disclosure and a baseline method Text2Mesh

[00286] FIG. 5K is a schematic diagram of results of generating a virtual object based on a plurality of methods according to an embodiment of the present disclosure. As shown in FIG. 5K, the first row is a multi-operation result with five different random seeds by using Text2Mesh; the second row is a multi-operation result with five different random seeds by using AvatarCLIP. It can be learned from FIG. 5K, Text2Mesh fails most operations, and cannot generate the virtual object consistent with the input text, while AvatarCLIP outputs a result with high quality and consistent with the input text. In indicates that M_t is unstable in the optimization of the neural network, especially in the case that supervision is relatively weak. M_t is optimized based on N' , which facilitates to improve the robustness of an algorithm.

[00287] 4) A comparison result of the target animation generated by the method AvatarCLIP provided in the embodiments of the present disclosure and the baseline method

[00288] FIG. 5L is a schematic diagram of results of generating a target animation based on a plurality of methods according to an embodiment of the present disclosure. As shown in FIG. 5L, the first column is a result generated based on a baseline (i). The baseline (i) is to first sort all of the candidate poses from the candidate pose S by using a similarity score s_i corresponding to the baseline, and then perform direct interpolation on the latent code between each pair of adjacent poses, so as to generate the target animation. The second column is a result generated based on a baseline (ii). The baseline (ii) is to directly calculate a reconstruction loss without re-weighting $\mathcal{L}_{\text{pose}}$. The third column is a result generated based on AvatarCLIP. As shown in FIG. 5L, with regard to "tooth brushing", the baseline (i) incorrectly organizes a sequence of the candidate poses to generate irrelevant action segments. The baseline (ii) cannot generate reasonable motions due to its blind focus on reconstructing disordered S. By introducing a re-weighting mechanism and two other losses, AvatarCLIP not only focuses on reconstruction, but also considers the rationality of animations. With regard to "kicking the ball", animations of the baseline (i) and the baseline (ii) do not show drastic changes when the legs are kicked out. $\mathcal{L}_{\text{delta}}$ plays an important role here to control the strength of the motions. With regard to "rising the arms", a motion sequence from a neutral pose to a pose of rising the arms should be generated. However, an action sequence of a motion sequence from the

baseline (i) is opposite to that expected. The baseline (ii) introduces a plurality of unrelated actions. With the assistance of $\mathcal{L}_{\text{clip}}^m$, AvatarCLIP can generate the motions with a correct sequence, and meanwhile, the overall sequence is more consistent with a given description.

[00289] Based on the above embodiments, an embodiment of the present disclosure provides an apparatus for generating a virtual object. FIG. 6 is a schematic diagram of a composition structure of an apparatus for generating a virtual object according to an embodiment of the present disclosure. As shown in FIG. 6, the apparatus 60 for generating a virtual object includes a first acquisition module 61, a first determination module 62, and a first generation module 63.

[00290] The first acquisition module 61 is configured to acquire first prompt information and second prompt information. The first prompt information represents body shape information of the virtual object. The second prompt information represents appearance information of the virtual object.

[00291] The first determination module 62 is configured to determine a first virtual model based on the first prompt information.

[00292] The first generation module 63 is configured to generate the virtual object based on the second prompt information and the first virtual model.

[00293] In some implementation modes, the first determination module 62 is further configured to: determine a target body shape image matching the first prompt information based on a preset first correspondence, the first correspondence representing a correspondence between the first prompt information and a body shape image; and perform three-dimensional modeling based on the target body shape image, to generate the first virtual model.

[00294] In some implementation modes, the first determination module 62 is further configured to: respectively determine a first distance between each body shape image in the first correspondence and a set standard body shape image; determine a second distance between the first prompt information and standard prompt information corresponding to the standard body shape image; and determine the target body shape image based on each first distance and the second distance.

[00295] In some implementation modes, the first generation module 63 is further configured to: for each first observation view angle of at least one first observation view angle, determine a first observation image of the first virtual model at the first observation view angle; for each first observation image, the first observation image for at least one time is updated by using a preset first neural network, to obtain a second observation image; determine a first target loss value based on each second observation image and the second prompt information; and in the case that the first target loss value satisfies a first preset condition, render the first virtual model based on each second observation image, to generate the virtual object.

[00296] In some implementation modes, the first generation module 63 is further configured to: in the case that the first target loss value does not satisfy the first preset condition, update a network parameter of the first neural network based on the first target loss value; update each first observation image for at least one time by using the updated first neural network, to obtain the updated second observation image; and determine an updated first target loss value based on each updated second observation image and the second prompt information.

[00297] In some implementation modes, the first observation view angle includes a first observation point and at least one first observation direction corresponding to the

first observation point. The first observation image includes at least one first pixel point. Each first pixel point corresponds to one first intersection point on the first virtual model. Each first intersection point represents the first intersection point of the first observation point and the first virtual model in one first observation direction. The first generation module 63 is further configured to update the first observation image for at least one time by using a preset first neural network to obtain a second observation image in the manner of: determining at least one first spatial point corresponding to each first observation direction based on the first observation point and each first observation direction; for each first spatial point of each first observation direction, determine, by using the preset first neural network, a first distance value between the first intersection point corresponding to the first observation direction where the first spatial point is located and the first spatial point, and determining a first color value obtained by observing the first virtual model according to the first observation direction at the first spatial point; and updating the first observation image for at least one time based on the first distance value and the first color value corresponding to each first spatial point, to obtain the second observation image.

[00298] In some implementation modes, the first generation module 63 is further configured to: for each first observation direction, determine a first weight value of each first spatial point based on the first distance value corresponding to each first spatial point in the first observation direction; and update the at least one first pixel point in the first observation image based on the first weight value and the first color value of each first spatial point, to obtain the second observation image.

[00299] In some implementation modes, the first generation module 63 is further configured to: for each first observation image, determine a first loss sub-value based on the second prompt information and the first observation image; for each second observation image, determine a second loss sub-value based on the second prompt information and the second observation image; and determine the first target loss value based on each first loss sub-value and each second loss sub-value.

[00300] In some implementation modes, the first generation module 63 is further configured to: determine a third observation image based on the first observation image; and determine the first loss sub-value based on the second prompt information and the third observation image.

[00301] In some implementation modes, the first generation module 63 is further configured to: for each first observation direction, determine a distance gradient value of each first spatial point based on the first distance value corresponding to each first spatial point in the first observation direction; and update the at least one first pixel point in the first observation image based on the first weight value of each first spatial point and the distance gradient value, to obtain the third observation image.

[00302] In some implementation modes, the apparatus further includes a second acquisition module and a second generation module. The second acquisition module is configured to acquire third prompt information. The third prompt information represents animation information of the virtual object. The second generation module is configured to generate a target animation of the virtual object based on the third prompt information.

[00303] In some implementation modes, the second generation module is further configured to: determine a first candidate pose image set based on the third prompt information; generate a first animation based on a preset first parameter; and determine the target animation of the virtual object based on the first candidate pose image set and the first animation.

- [00304] In some implementation modes, the second generation module is further configured to: query a preset second correspondence by using the third prompt information, to obtain at least one candidate pose image matching the third prompt information, the second correspondence representing a correspondence between the third prompt information and the pose image; and determine the first candidate pose image set based on each candidate pose image.
- [00305] In some implementation modes, the second generation module is further configured to: determine a second target loss value based on the first candidate pose image set and the first animation; and use the first animation as the target animation in the case that the second target loss value satisfies a second preset condition.
- [00306] In some implementation modes, the second generation module is further configured to: in the case that the second target loss value does not satisfy the second preset condition, update the first parameter to obtain the updated first parameter; generate an updated first animation based on the updated first parameter; and determine the second target loss value based on the first candidate pose image set and the updated first animation.
- [00307] In some implementation modes, the second generation module is further configured to: determine a third loss sub-value based on the first candidate pose image set and each frame image in the first animation; determine a fourth loss sub-value based on each two adjacent frame images in the first animation; determine a fifth loss sub-value based on the third prompt information and each frame image in the first animation; and determine the second target loss value based on the third loss sub-value, the fourth loss sub-value, and the fifth loss sub-value.
- [00308] The descriptions of the above apparatus embodiments are similar to the descriptions of the above method embodiments, and have similar beneficial effects to the method embodiments. For technical details that are not disclosed in the apparatus embodiments of the present disclosure, refer to the descriptions of the method embodiments of the present disclosure for understanding.
- [00309] It is to be noted that, in the embodiments of the present disclosure, if the above method is implemented in the form of a software functional module and sold or used as an independent product, it can be stored in the computer-readable storage medium. Based on such an understanding, the technical solutions of the present disclosure substantially or parts making contributions to the related art may be embodied in form of a software product, and the software product is stored in a storage medium, including a plurality of instructions for causing an electronic device (which may be a personal computer, a server, a network device or the like) to execute all or part of the method in each embodiment of the present disclosure. The foregoing storage medium includes: a USB flash disk, a mobile Hard Disk Drive (HDD), a Read-Only Memory (ROM), and various media that can store program codes, such as a magnetic disk, or an optical disk. In this way, the embodiments of the present disclosure are not limited to any specific combination of hardware and software.
- [00310] An embodiment of the present disclosure provides an electronic device. The electronic device includes a memory and a processor. The memory stores a computer program executable on the processor. The processor, when executing the computer program, implements the above method.
- [00311] An embodiment of the present disclosure provides a computer-readable storage medium. The computer-readable storage medium stores a computer program. The above method is implemented when the computer program is performed by a processor. The

computer-readable storage medium may be transitory or non-transitory.

[00312] An embodiment of the present disclosure provides a computer program product. The computer program product includes a non-transitory computer-readable storage medium storing a computer program. When the computer program is read and executed by a computer, part or all of steps in the above method are implemented. The computer program product can be specifically implemented in a manner of hardware, software or a combination thereof. In an optional embodiment, the computer program product is embodied as a computer storage medium. In another optional embodiment, the computer program product is embodied as a software product, such as a Software Development Kit (SDK) and the like.

[00313] It is to be noted that, FIG. 7 is a schematic diagram of a hardware entity of an electronic device according to an embodiment of the present disclosure. As shown in FIG. 7, a hardware entity of the electronic device 700 includes a processor 701, a communication interface 702, and a memory 703.

[00314] The processor 701 usually controls an overall operation of the electronic device 700.

[00315] The communication interface 702 may cause the electronic device to achieve communication with other terminals or servers through a network.

[00316] The memory 703 is configured to store an instruction and an application executable by the processor 701, or may cache data to be processed or processed by the processor 701 and each module in the electronic device 700 (such as, image data, audio data, voice communication data, and video communication data), which may be implemented by using FLASH or a Random Access Memory (RAM). Data transmission among the processor 701, the communication interface 702, and the memory 703 may be achieved by using a bus 704.

[00317] It is to be pointed out here that the above descriptions about the storage medium 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 and device embodiments of the present disclosure may be understood with reference to the descriptions about the method embodiments of the present disclosure.

[00318] 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 present 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 is to be understood that, in various embodiments of the present disclosure, the sequence number of each process does not mean the sequence of execution. The execution sequence of each process should be determined by its functions and internal logic, which should not constitute any limitation on the implementation process of the embodiments of the present disclosure. The serial numbers of the foregoing embodiments of the present disclosure are merely for description, and do not represent the superiority or inferiority of the embodiments.

[00319] It is to be noted that terms "include" and "comprise" or any other variant thereof is intended to cover nonexclusive inclusions herein, so that a process, method, object or apparatus 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 apparatus. 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 apparatus including the component.

[00320] In several embodiments provided by the present disclosure, it is to be understood that the disclosed device and method may be implemented in other ways. 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, a plurality of 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.

[00321] 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.

[00322] In addition, each functional unit in each embodiment of the present 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.

[00323] 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: a mobile storage device, an ROM, and various media that can store program codes, such as a magnetic disk, or an optical disk.

[00324] Or, if the integrated unit is implemented in the form of the software functional unit and sold or used as an independent product, it can be stored in the computer readable storage medium. Based on such an understanding, the technical solutions of the present disclosure substantially or parts making contributions to the related art may be embodied in form of a software product, and the computer software product is stored in a storage medium, including a plurality of instructions for causing an electronic device (which may be a personal computer, a server, a network device or the like) to execute all or part of the method in each embodiment of the present disclosure. The foregoing storage medium includes a portable storage device, an ROM, and various media that can store program codes, such as a magnetic disk, or an optical disk.

[00325] The above is only the implementation modes of the present disclosure and not intended to limit the scope of protection of the present disclosure. Any variations or replacements apparent to those skilled in the art within the technical scope disclosed by the present disclosure shall fall within the scope of protection of the present disclosure.

CLAIMS

1. A method for generating a virtual object, comprising:
 - acquiring first prompt information and second prompt information, the first prompt information representing body shape information of the virtual object, and the second prompt information representing appearance information of the virtual object;
 - determining a first virtual model based on the first prompt information; and
 - generating the virtual object based on the second prompt information and the first virtual model.
2. The method of claim 1, wherein the determining a first virtual model based on the first prompt information comprises:
 - determining a target body shape image matching the first prompt information based on a preset first correspondence, the first correspondence representing a correspondence between the first prompt information and a body shape image; and
 - performing three-dimensional modeling based on the target body shape image, to generate the first virtual model.
3. The method of claim 2, wherein the determining a target body shape image matching the first prompt information based on a preset first correspondence comprises:
 - respectively determining a first distance between each body shape image in the first correspondence and a set standard body shape image;
 - determining a second distance between the first prompt information and standard prompt information corresponding to the standard body shape image; and
 - determining the target body shape image based on each the first distance and the second distance.
4. The method of any of claims 1 to 3, wherein the generating the virtual object based on the second prompt information and the first virtual model comprises:
 - for each first observation view angle of at least one first observation view angle, determining a first observation image of the first virtual model at the each first observation view angle;
 - for each the first observation image, updating the first observation image for at least one time by using a preset first neural network, to obtain a second observation image;
 - determining a first target loss value based on each the second observation image and the second prompt information; and
 - in a case that the first target loss value satisfies a first preset condition, rendering the first virtual model based on each the second observation image, to generate the virtual object.
5. The method of claim 4, wherein the generating the virtual object based on the second prompt information and the first virtual model further comprises:
 - in a case that the first target loss value fails to satisfy the first preset condition, updating a network parameter of the first neural network based on the first target loss value;
 - updating each the first observation image for at least one time by using the updated first neural network, to obtain an updated second observation image; and
 - determining an updated first target loss value based on each the updated second observation image and the second prompt information.

6. The method of claim 4 or 5, wherein the first observation view angle comprises a first observation point and at least one first observation direction corresponding to the first observation point; the first observation image comprises at least one first pixel point, each the first pixel point corresponds to one first intersection point on the first virtual model, each the first intersection point represents the first intersection point between the first virtual model and a line coming from the first observation point and extending along one first observation direction;

wherein the updating the first observation image for at least one time by using a preset first neural network, to obtain a second observation image comprises:

determining at least one first spatial point corresponding to each the first observation direction based on the first observation point and each the first observation direction;

for each the first spatial point of each the first observation direction, determining, by using the preset first neural network, a first distance value between the first intersection point corresponding to the first observation direction where the first spatial point is located and the first spatial point, and determining a first color value obtained by observing the first virtual model according to the first observation direction at the first spatial point; and

updating the first observation image for at least one time based on the first distance value and the first color value corresponding to each the first spatial point, to obtain the second observation image.

7. The method of claim 6, wherein the updating the first observation image for at least one time based on the first distance value and the first color value corresponding to each the first spatial point, to obtain the second observation image comprises:

for each the first observation direction, determining a first weight value of each the first spatial point based on the first distance value corresponding to each the first spatial point in the first observation direction; and

updating the at least one first pixel point in the first observation image based on the first weight value and the first color value of each the first spatial point, to obtain the second observation image.

8. The method of claim 6 or 7, wherein the determining a first target loss value based on each the second observation image and the second prompt information comprises:

for each the first observation image, determining a first loss sub-value based on the second prompt information and the first observation image;

for each the second observation image, determining a second loss sub-value based on the second prompt information and the second observation image; and

determining the first target loss value based on each the first loss sub-value and each the second loss sub-value.

9. The method of claim 8, wherein the determining a first loss sub-value based on the second prompt information and the first observation image comprises:

determining a third observation image based on the first observation image; and

determining the first loss sub-value based on the second prompt information and the third observation image.

10. The method of claim 9, wherein the determining a third observation image based on the first observation image comprises:

for each the first observation direction, determining a distance gradient value of each

the first spatial point based on the first distance value corresponding to each first spatial point in the first observation direction; and

updating the at least one first pixel point in the first observation image based on the first weight value of each the first spatial point and the distance gradient value, to obtain the third observation image.

11. The method of any of claims 1 to 10, further comprising:

acquiring third prompt information, the third prompt information representing animation information of the virtual object; and

generating a target animation of the virtual object based on the third prompt information.

12. The method of claim 11, wherein the generating a target animation of the virtual object based on the third prompt information comprises:

determining a first candidate pose image set based on the third prompt information;

generating a first animation based on a preset first parameter; and

determining the target animation of the virtual object based on the first candidate pose image set and the first animation.

13. The method of claim 12, wherein the determining a first candidate pose image set based on the third prompt information comprises:

querying a preset second correspondence by using the third prompt information, to obtain at least one candidate pose image matching the third prompt information, the second correspondence representing a correspondence between the third prompt information and the pose image; and

determining the first candidate pose image set based on each the candidate pose image.

14. The method of claim 12 or 13, wherein the determining the target animation of the virtual object based on the first candidate pose image set and the first animation comprises:

determining a second target loss value based on the first candidate pose image set and the first animation; and

using the first animation as the target animation in a case that the second target loss value satisfies a second preset condition.

15. The method of claim 14, wherein the determining the target animation of the virtual object based on the first candidate pose image set and the first animation further comprises:

in a case that the second target loss value fails to satisfy the second preset condition, updating the first parameter to obtain the updated first parameter;

generating an updated first animation based on the updated first parameter; and

determining the second target loss value based on the first candidate pose image set and the updated first animation.

16. The method of claim 14 or 15, wherein the determining a second target loss value based on the first candidate pose image set and the first animation comprises:

determining a third loss sub-value based on the first candidate pose image set and each frame of image in the first animation;

determining a fourth loss sub-value based on each two adjacent frame images in the

first animation;

determining a fifth loss sub-value based on the third prompt information and each frame of image in the first animation; and

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

17. An apparatus for generating a virtual object, comprising:

a first determination module, configured to acquire first prompt information and second prompt information, the first prompt information representing body shape information of the virtual object, and the second prompt information representing appearance information of the virtual object;

a second determination module, configured to determine a first virtual model based on the first prompt information; and

a first generation module, configured to generate the virtual object based on the second prompt information and the first virtual model.

18. An electronic device, comprising a processor and a memory, wherein the memory stores a computer program executable on the processor, wherein the processor is configured to execute the computer program to:

acquire first prompt information and second prompt information, the first prompt information representing body shape information of the virtual object, and the second prompt information representing appearance information of the virtual object;

determine a first virtual model based on the first prompt information; and

generate the virtual object based on the second prompt information and the first virtual model.

19. A computer-readable storage medium, having a computer program stored thereon, wherein the computer program is performed by a processor to:

acquire first prompt information and second prompt information, the first prompt information representing body shape information of the virtual object, and the second prompt information representing appearance information of the virtual object;

determine a first virtual model based on the first prompt information; and

generate the virtual object based on the second prompt information and the first virtual model.

20. A computer program, wherein the computer program is executable by a processor to:

acquire first prompt information and second prompt information, the first prompt information representing body shape information of the virtual object, and the second prompt information representing appearance information of the virtual object;

determine a first virtual model based on the first prompt information; and

generate the virtual object based on the second prompt information and the first virtual model.