## DESCRIPTION

# TITLE OF INVENTION: METHOD FOR GENERATING VIDEO BASED ON SCENE RECONSTRUCTION, STORAGE MEDIUM AND SYSTEM

## CROSS-REFERENCE TO RELATED APPLICATIONS

The present disclosure claims priority of Chinese Patent Application No. 202310467335.3, filed to China National Intellectual Property Administration on April 25, 2023 and titled "METHOD FOR GENERATING VIDEO BASED ON SCENE RECONSTRUCTION, STORAGE MEDIUM AND SYSTEM", the content of which is hereby incorporated by reference in its entirety.

## TECHNICAL FIELD

The present disclosure relates to the field of computer technologies, artificial intelligence technologies, and virtual reality (VR)/augmented reality (AR) technologies, and in particular, to a method for generating a video based on scene reconstruction, a storage medium and a system.

## BACKGROUND OF THE INVENTION

In recent years, a technical concept of reconstructing a three-dimensional virtual scene based on image data (for example, aerial data or remote sensing data) has increasing application value in the fields of oblique photography, topographic surveying and mapping, cultural tourism promotion, scenic spot/campus/building reconstruction, VR roaming, and the like. However, in an image data–based scene reconstruction method provided in the related art, a reconstruction concept for an explicit mesh model (also referred to as explicit modeling) is usually used. The explicit mesh model obtained in this method has many mesh defects (for example, a geometrical error or a hole), and it is difficult to provide high-quality scene roaming effect at a free viewpoint. Even if manual optimization is performed, a sense of reality of a scene roaming video rendered based on the mesh model is still poor.

As to the problem mentioned above, no effective solution have been provided yet.

## SUMMARY OF THE INVENTION

Embodiments of the present disclosure provide a method for generating a video based on scene reconstruction, a storage medium and a system, to solve at least technical problems of many defects in a modeling result and a poor sense of reality of a scene roaming video in the related art due to scene reconstruction performed through explicit modeling.

According to an aspect of the embodiments of the present disclosure, a method for generating a video based on scene reconstruction is provided, and includes the following steps. Image data to be reconstructed is obtained, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed. Sparse reconstruction is performed on the image data to be reconstructed to obtain a reconstruction result. Implicit modeling is performed based on the image data to be reconstructed and the reconstruction result to obtain a modeling result. A target video is generated by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene.

According to another aspect of the embodiments of the present disclosure, a method for generating a video based on scene reconstruction is further provided, and includes the following steps. Image data to be reconstructed is displayed on an operation interface in response to an input instruction performed on the operation interface, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed. A target video on the operation interface is displayed in response to a reconstruction instruction performed on the operation interface, where the target video is used for performing scene roaming on the preset three-dimensional scene, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on the image data to be reconstructed and a reconstruction result, and the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed.

According to another aspect of the embodiments of the present disclosure, a method for generating a video based on scene reconstruction is further provided, and includes the following steps. An initial default picture is displayed on a displayed picture of a VR device or an AR device. The VR device or the AR device is driven to jump from displaying the initial

default picture to displaying a target video, where the target video is used for performing scene roaming on a preset three-dimensional scene to be reconstructed, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on image data to be reconstructed and a reconstruction result, the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed, and the image data to be reconstructed is used for recording an image of the preset three-dimensional scene.

According to another aspect of the embodiments of the present disclosure, a computer-readable storage medium is further provided. The computer-readable storage medium stores a program. When the program is run, a device on which the computer-readable storage medium is located is controlled to perform the method for generating a video based on scene reconstruction according to any one of the implementations.

According to another aspect of the embodiments of the present disclosure, a system for generating a video based on scene reconstruction is further provided, and includes: a processor; and a memory, connected with the processor and configured to provide the processor with instructions for processing the following processing steps: obtaining image data to be reconstructed, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; performing sparse reconstruction on the image data to be reconstructed to obtain a reconstruction result; performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain a modeling result; and generating a target video by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene.

In the embodiments of the present disclosure, the image data to be reconstructed is obtained, where the image data to be reconstructed is used for recording the image of the preset three-dimensional scene to be reconstructed; sparse reconstruction is further performed on the image data to be reconstructed to obtain the reconstruction result; the modeling result is obtained by performing implicit modeling based on the image data to be reconstructed and the reconstruction result; and then the target video is generated by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene. Therefore, in the embodiments of the present disclosure, an objective of obtaining a scene

roaming video through rendering by performing scene modeling based on the image data to be reconstructed in a manner of sparse reconstruction and implicit modeling is achieved, so that technical effect of improving accuracy of a scene modeling result and a sense of reality of the scene roaming video is achieved, and technical problems of many defects in a modeling result and a poor sense of reality of the scene roaming video in the related art due to scene reconstruction performed through explicit modeling are further solved.

It is easily noted that the general descriptions and the following detailed descriptions are intended to explain the present disclosure by using examples, but do not constitute a limitation on the present disclosure.

## BRIEF DESCRIPTION OF DRAWINGS

The accompanying drawings described herein are intended for further understanding of the present disclosure, and constitute a part of the present disclosure. Exemplary embodiments of the present disclosure and descriptions thereof are used for explaining the present disclosure, but do not constitute any inappropriate limitation on the present disclosure. In the accompanying drawings:

FIG. 1 is a block diagram of a hardware structure of a computer terminal (or a mobile device) configured to implement a method for generating a video based on scene reconstruction;

FIG. 2 is a diagram of a hardware environment of a VR device for a method for generating a video based on scene reconstruction according to Embodiment One of the present disclosure;

FIG. 3 is a structural block diagram of a computing environment for a method for generating a video based on scene reconstruction according to Embodiment One of the present disclosure;

FIG. 4 is a flowchart of a method for generating a video based on scene reconstruction according to Embodiment One of the present disclosure;

FIG. 5 is a diagram of an optional scene reconstruction process according to Embodiment One of the present disclosure;

FIG. 6 is a diagram of an optional implicit modeling process according to Embodiment One of the present disclosure;

FIG. 7 is a flowchart of a method for generating a video based on scene reconstruction according to Embodiment Two of the present disclosure;

FIG. 8 is a flowchart of a method for generating a video based on scene reconstruction according to Embodiment Three of the present disclosure;

FIG. 9 is a structural schematic diagram of an apparatus for generating a video based on scene reconstruction according to Embodiment Four of the present disclosure;

FIG. 10 is a structural schematic diagram of an optional apparatus for generating a video based on scene reconstruction according to Embodiment Four of the present disclosure;

FIG. 11 is a structural schematic diagram of another optional apparatus for generating a video based on scene reconstruction according to Embodiment Four of the present disclosure; and

FIG. 12 is a structural block diagram of a computer terminal according to Embodiment Five of the present disclosure.

## DETAILED DESCRIPTION OF THE INVENTION

To make a person having ordinary skill in the art better understand the solutions of the present disclosure, the following clearly and thoroughly describes the technical solutions in the embodiments of the present disclosure with reference to the accompanying drawings in the embodiments of the present disclosure. Apparently, the described embodiments are some but not all of the embodiments of the present disclosure. All other embodiments obtained by a person having ordinary skill in the art based on the embodiments of the present disclosure without creative efforts shall fall within the protection scope of the present disclosure.

It should be noted that the terms “first”, “second”, and the like in the specification, claims, and accompanying drawings of the present disclosure are intended to distinguish between similar objects but do not necessarily indicate a specific order or sequence. It should be understood that data used in this way is interchangeable in appropriate circumstances, so that the embodiments of the present disclosure described herein can be implemented in an order other than the order illustrated or described herein. In addition, the terms “include”, “have”, and any variants thereof are intended to cover a non-exclusive inclusion. For example, a process, method, system, product, or device that includes a list of steps or units is not necessarily limited

to those expressly listed steps or units, but may include other steps or units that are not expressly listed or are inherent to the process, method, product, or device.

First, the following explanations are applicable to some nouns or terms that appear during description of the embodiments of the present disclosure.

Neural radiance field (NeRF): a method for integrating an image and a deformation of three-dimensional space by using a deep learning technology to generate a high-quality three-dimensional scene. In the present disclosure, the NeRF is a method for reconstructing a three-dimensional scene based on multiple images and an implicit mesh model.

Structure from motion (SfM): a technology for estimating a structure of a three-dimensional scene based on a sequence of multiple two-dimensional images.

Multi-layer perceptron (MLP): a neural network including multiple fully-connected layers.

In recent years, a technical concept of reconstructing a three-dimensional virtual scene based on image data (for example, aerial data or remote sensing data) has increasing application value in the fields of oblique photography, topographic surveying and mapping, cultural tourism promotion, scenic spot/campus/building reconstruction, VR roaming, and the like. However, in an image data–based method for generating a video based on scene reconstruction in the related art, a reconstruction concept for an explicit mesh model is usually used. The explicit mesh model obtained in this method has many mesh defects (for example, a geometrical error or a hole), and it is difficult to provide high-quality scene roaming effect at a free viewpoint. Even if manual optimization is performed, a sense of reality of a scene roaming video rendered based on the mesh model is still poor.

Based on the related art, a person having ordinary skill in the art proposes a NeRF algorithm. The NeRF algorithm is used during scene reconstruction to obtain a modeling result through implicit modeling, and then a scene image is obtained through rendering based on the modeling result. However, the existing NeRF algorithm prior to the present disclosure is used for object model reconstruction or small-scale scene reconstruction, and it is difficult to apply the existing NeRF algorithm to a reconstruction task of a large-scale scene (for example, a scene, such as a campus or a city, corresponding to aerial data, or a scene, such as a terrain or a district, corresponding to remote sensing data). In addition, it is difficult to obtain a scene roaming

video with a strong sense of reality at the free viewpoint through rendering based on a reconstruction result obtained by using the existing NeRF algorithm.

Embodiment One

According to embodiments of the present disclosure, a method for generating a video based on scene reconstruction is further provided. It should be noted that steps shown in flowcharts of the accompanying drawings are performed, for example, in a computer system with a group of computer-executable instructions. In addition, although a logical order is shown in the flowcharts, in some cases, the shown or described steps are performed in an order different from the order herein.

The method provided in Embodiment One of the present disclosure is performed on a mobile terminal, a computer terminal, or a similar operation apparatus. FIG. 1 is a block diagram of a hardware structure of a computer terminal (or a mobile device) configured to implement a method for generating a video based on scene reconstruction. As shown in FIG. 1, the computer terminal 10 (or the mobile device 10) may include at least one (indicated by 102a, 102b, ..., and 102n in the figure) processor 102 (the processor 102 may include but is not limited to a processing apparatus such as a microcontroller unit (MCU) or a field programmable gate array (FPGA)), a memory 104 configured to store data, and a transmission apparatus 106 configured to perform a communication function. In addition, the computer terminal 10 may further include a display, an input/output interface (I/O interface), a universal serial bus (USB) port (which is included as one of ports of a bus), a network interface, a cursor control device (for example, a mouse or a touchpad), a keyboard, a power supply, and/or a camera.

A person having ordinary skill in the art may understand that the structure shown in FIG. 1 is an example and does not constitute a limitation on a structure of the electronic apparatus. For example, the computer terminal 10 may alternatively include more or fewer components than those shown in FIG. 1, or have a configuration different from that shown in FIG. 1.

It should be noted that the at least one processor 102 and/or other data processing circuits are usually referred to as a “data processing circuit” in the present disclosure. All or a part of the data processing circuit is embodied as software, hardware, firmware, or any other combination. In addition, the data processing circuit is a single independent processing module, or all or a part of the data processing circuit is integrated into any one of other components in the

computer terminal 10 (or the mobile device). As involved in the embodiments of the present disclosure, the data processing circuit serves as a processor for control (for example, selection of a variable-resistance terminal path connected with an interface).

The memory 104 is configured to store a software program and a module of application software, for example, a program instruction/data storage apparatus corresponding to the method for generating a video based on scene reconstruction in the embodiments of the present disclosure. The processor 102 runs the software program and the module stored in the memory 104 to perform various functional applications and data processing, in other words, implement the method for generating a video based on scene reconstruction. The memory 104 may include a high-speed random memory, or may include a non-transitory memory, for example, at least one magnetic storage apparatus, a flash memory, or another non-transitory solid-state memory. In some examples, the memory 104 may further include memories remotely disposed relative to the processor 102, and the remote memories are connected with the computer terminal 10 through a network. Examples of the network include but are not limited to the Internet, an intranet, a local area network, a mobile communication network, and a combination thereof.

The transmission apparatus 106 is configured to be connected with a network through a network interface to receive or send data. Specific examples of the network may include a wired network and/or a wireless network provided by a communication provider of the computer terminal 10. In an example, the transmission apparatus 106 includes a network interface controller (NIC), which is connected with another network device through a base station, to communicate with the Internet. In an example, the transmission apparatus 106 is a radio frequency (RF) module, which is configured to communicate with the Internet in a wireless manner.

The display shown in FIG. 1 is, for example, a touchscreen liquid crystal display (LCD). The LCD enables a user to interact with a user interface of the computer terminal 10 (or the mobile device).

It should be noted herein that, in some optional embodiments, the computer device (or the mobile device) shown in FIG. 1 may include a hardware component (including a circuit), a software component (including computer code stored in a computer-readable medium), or a combination of a hardware component and a software component. It should be noted that FIG.

1 is one of specific examples and is intended to show types of components that may exist in the computer device (or the mobile device).

It should be noted herein that, in some embodiments, the computer device (or the mobile device) shown in FIG. 1 has a touch display (also referred to as a "touchscreen" or a "touch display screen"). In some embodiments, the computer device (or the mobile device) shown in FIG. 1 has a graphical user interface (GUI). A user may perform human-computer interaction with the GUI through a finger touch and/or a gesture on a touch-sensitive surface of the touch. The human-computer interaction function herein optionally includes the following interaction: web page creation, drawing, text processing, electronic document production, gaming, video conferencing, instant messaging, email receiving/sending, a call interface, digital video playing, digital music playing, web browsing, and/or the like. An executable instruction used for performing the human-computer interaction function is configured/stored in at least one processor-executable computer program product or readable storage media.

The method embodiment provided in Embodiment One of the present disclosure is performed on a VR device or an AR device. The VR device is used as an example. FIG. 2 is a diagram of a hardware environment of a VR device for a method for generating a video based on scene reconstruction according to Embodiment One of the present disclosure. As shown in FIG. 2, a VR device 204 is connected with a terminal 206, and the terminal 206 is connected with a server 202 through a network. The VR device 204 is not limited to a VR helmet, VR glasses, a VR all-in-one machine, and the like. The terminal 204 is not limited to a PC, a mobile phone, a tablet computer, and the like. The server 202 is a server corresponding to a media file operator. The network includes but is not limited to a wide area network, a metropolitan area network, or a local area network.

Optionally, the VR device 204 in the embodiments of the present disclosure includes a memory, a processor, and a transmission apparatus (not shown in FIG. 2). The memory is configured to store an application program, and the application program is used for performing the method for generating a video based on scene reconstruction provided in the embodiments of the present disclosure. The processor may invoke and execute the application program stored in the memory, to obtain target grids point corresponding to a position to be recognized

by using the method for generating the video based on scene reconstruction provided in the embodiments of the present disclosure.

Optionally, the terminal 206 in the embodiments of the present disclosure may alternatively be configured to perform the method for generating the video based on scene reconstruction, and display, through a display picture of the VR device or the AR device, target grids point corresponding to a position to be recognized, or display, through an image output device associated with the VR device or the AR device, the target grid point corresponding to the position to be recognized generated by performing the method for generating the video based on scene reconstruction.

The block diagram of the hardware structure shown in FIG. 2 not only is used as an exemplary block diagram of a hardware environment of the AR/VR device (or mobile device), but also is used as an exemplary block diagram of the server.

FIG. 3 is a structural block diagram of a computing environment for a method for generating a video based on scene reconstruction according to Embodiment One of the present disclosure. FIG. 3 shows, by using a block diagram, an embodiment in which the VR device (or the mobile device) shown in FIG. 2 is used as a computing node in a computing environment 301.

Still as shown in FIG. 3, the computing environment 301 includes multiple (indicated by 310-1, 310-2, ... in FIG. 3) computing nodes (for example, servers) running on a distributed network. Each computing node includes local processing and memory resources, and a terminal user 302 may remotely run an application program or store data in the computing environment 301. The application program is provided as multiple services 320-1 (representing a service A), 320-2 (representing a service D), 320-3 (representing a service E), and 320-4 (representing a service H) in the computing environment 301.

Still as shown in FIG. 3, the terminal user 302 may provide and access a service by using a web browser or another software application program on a client. In some embodiments, a supply and/or a request of the terminal user 302 are provided for an ingress gateway 330. The ingress gateway 330 may include a corresponding proxy to process a supply and/or a request for the service 320 (at least one service provided in the computing environment 301).

Still as shown in FIG. 3, the service 320 is provided or deployed based on various virtualization technologies supported by the computing environment 301. In some embodiments, the service 320 is provided based on virtual machine (VM)-based virtualization, container-based virtualization, and/or a similar manner. In the VM-based virtualization, a VM is initialized to simulate a real computer, to execute a program and an application program without direct contact with any actual hardware resource. When a machine is virtualized by using the VM, based on the container-based virtualization, a container is started to virtualize an entire operating system (OS), so that multiple workloads can run on a single OS instance.

In an embodiment of container-based virtualization, several containers of the service 320 is assembled into one pod (for example, a Kubernetes pod). For example, as shown in FIG. 3, the service 320-2 is configured with at least one pod 340-1, 340-2, ..., and 340-N (collectively referred to as a pod 340). Each pod 340 may include a proxy 345 and at least one container 342-1, 342-2, ..., and 342-M (collectively referred to as a container 342). One or more containers 342 in the pod 340 process a request related to at least one corresponding function of the service. The proxy 345 usually controls network functions, such as routing and load balancing, related to the service. Other services 320 may also have pods similar to the pod 340.

During an operation, to execute a user request from the terminal user 302, at least one services 320 in the computing environment 301 may need to be invoked; and to perform at least one functions of one service 320, at least one functions of another service 320 may need to be invoked. As shown in FIG. 3, the service 320-1 (representing the service A) receives a user request from the terminal user 302 through the ingress gateway 330, the service 320-1 (representing the service A) may invoke the service 320-2 (representing the service D), and the service 320-2 (representing the service D) may request the service 320-3 (representing the service E) to perform at least one functions.

The computing environment is a cloud computing environment. Resource allocation is managed by a cloud service provider, and a function is allowed to be developed without considering implementation, adjustment, or extension of a server. The computing environment allows a developer to execute code of a response event without building or maintaining complex infrastructure. A service is divided into a group of functions that can be automatically

and independently scaled, instead of extending a single hardware device to process a potential load.

In the running environment, the present disclosure provides a method for generating a video based on scene reconstruction, as shown in FIG. 4. FIG. 4 is a flowchart of a method for generating a video based on scene reconstruction according to Embodiment One of the present disclosure. As shown in FIG. 4, the method for generating a video based on scene reconstruction includes the following steps.

In step S41, image data to be reconstructed is obtained, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed.

In step S42, sparse reconstruction is performed on the image data to be reconstructed to obtain a reconstruction result.

In step S43, implicit modeling is performed based on the image data to be reconstructed and the reconstruction result to obtain a modeling result.

In step S44, a target video is generated by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene.

The method for generating the video based on scene reconstruction is applied to but is not limited to the following scenarios: a visual navigation scenario (for example, constructing a three-dimensional virtual scene based on a map to perform autonomous navigation for a robot or an uncrewed aerial vehicle), a VR/AR scene production scenario (for example, VR exhibition hall production or VR/AR game scene design), a three-dimensional printing scenario (for example, reconstructing a virtual model based on an image for entity production), a three-dimensional movie and television creation and post-production scenario (for example, reconstructing a three-dimensional scene or a virtual object in a three-dimensional scene based on a given image), and a cultural tourism promotion scenario (for example, reconstructing a three-dimensional virtual scene of an area such as a scenic spot, a campus, an airport, a station, or a road, or producing a cultural tourism promotion video).

In some exemplary embodiments, the sparse reconstruction is a three-dimensional scene reconstruction operation based on a SfM algorithm, and the reconstruction result includes a

camera pose corresponding to the image data to be reconstructed. The implicit modeling is a three-dimensional scene modeling optimization operation based on a NeRF algorithm. The modeling result includes a scene model corresponding to the preset three-dimensional scene to be reconstructed. The target video is a scene roaming video, at a free viewpoint, corresponding to the preset three-dimensional scene to be reconstructed. The target video is a video obtained through rendering by using a rendering tool (for example, a game engine or image editing software).

It should be noted that a reconstruction algorithm used in the sparse reconstruction is not limited to the SfM algorithm. For example, the reconstruction algorithm may alternatively be: a sparse representation based on a greedy algorithm, such as, a matching pursuit (MP) algorithm or an orthogonal matching pursuit (OMP) algorithm; a sparse representation based on iterative optimization, such as, an L1 norm–based least absolute shrinkage and selection operator (LASSO) algorithm or an L0 norm–based backpropagation (BP) algorithm; a sparse representation based on a compressive sensing theory, such as, a compressive sampling matching pursuit (CoSaMP) algorithm based on a polynomial deflation algorithm, or a successive projection (SP) algorithm based on single orthogonal matching pursuit; a sparse representation based on dictionary learning, such as, a K-singular value decomposition (K-SVD) algorithm or a matching pursuit algorithm with orthogonal decision (MOD) algorithm; a sparse representation based on a convolutional neural network, such as, a sparse coding–based CNN algorithm; or a sparse representation based on a Bayesian method, such as, a Bayesian sparse coding algorithm based on a variational auto-encoder. A modeling algorithm used in the implicit modeling is not limited to the NeRF algorithm. For example, the modeling algorithm may alternatively be a hidden Markov model (HMM) algorithm, a decision tree algorithm, a random forest algorithm, a support vector machine (SVM) algorithm, or a neural network algorithm. The reconstruction algorithm and modeling algorithm are not limited in the present disclosure.

The method for generating the video based on scene reconstruction provided in the embodiments of the present disclosure is run on a client corresponding to the application scenario. The client determines the image data to be reconstructed in the application scenario from a preset database or data inputted by a user in real time, performs sparse reconstruction on the image data to be reconstructed to obtain the reconstruction result, performs implicit

modeling based on the image data to be reconstructed and the reconstruction result to obtain the modeling result, and then generates the target video by using the modeling result. Further, optionally, the client may further display the target video on a GUI, and provide the target video for the user through an output device (for example, a display or VR glasses).

The method for generating the video based on scene reconstruction provided in the embodiments of the present disclosure is run on a server corresponding to the application scenario. The server is an independent server or a server cluster, and generate the target video corresponding to the application scenario in the manner of sparse reconstruction and implicit modeling based on the image data to be reconstructed given by a client. The server may alternatively be a cloud server, and perform real-time interaction with the client in a software as a service (SaaS) manner, perform, based on the image data to be reconstructed given by the client, sparse reconstruction on the image data to be reconstructed to obtain the reconstruction result, perform implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain the modeling result, generate the target video by using the modeling result, and then return the target video to the client to provide the target video for a user.

In the embodiments of the present disclosure, the image data to be reconstructed is obtained, where the image data to be reconstructed is used for recording the image of the preset three-dimensional scene to be reconstructed; sparse reconstruction is further performed on the image data to be reconstructed to obtain the reconstruction result; the modeling result is obtained by performing implicit modeling based on the image data to be reconstructed and the reconstruction result; and then the target video is generated by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene. Therefore, in the embodiments of the present disclosure, an objective of obtaining a scene roaming video through rendering by performing scene modeling based on the image data to be reconstructed in a manner of sparse reconstruction and implicit modeling is achieved, so that technical effect of improving accuracy of a scene modeling result and a sense of reality of a scene roaming video is achieved, and technical problems of many defects in the modeling result and a poor sense of reality of the scene roaming video in the related art due to scene reconstruction performed through explicit modeling are further solved.

According to the method for generating the video based on scene reconstruction provided in the embodiments of the present disclosure, image data–based scene reconstruction is performed by using a neural implicit mesh model, to obtain the scene roaming video with a strong sense of reality through rendering. In particular, the method for generating the video based on scene reconstruction is especially suitable for large-scale scene reconstruction based on aerial data.

FIG. 5 is a diagram of an optional scene reconstruction process according to Embodiment One of the present disclosure. As shown in FIG. 5, taking performing scene reconstruction on a campus based on aerial data of the campus as example, the scene reconstruction process includes the following four stages. A first stage is an aerial data capture stage. A second stage is a SfM sparse reconstruction stage. A third stage is a neural implicit optimization modeling stage. A fourth stage is a roaming video rendering stage.

In some exemplary embodiments, the preset three-dimensional scene is a campus scene, and the image data to be reconstructed is aerial data corresponding to the campus scene. As shown in FIG. 5, in the first stage, the aerial data is captured by an aerial photography device in real time, and the aerial data includes at least one aerial image. In the second stage, SfM sparse reconstruction is performed on the aerial data to obtain a camera pose (equivalent to the reconstruction result) corresponding to each of the at least one aerial image. In the third stage, implicit modeling is performed based on the aerial data and the camera pose corresponding to the at least one aerial image, that is, modeling optimization is performed on the reconstruction result, to obtain a modeling result. In the fourth stage, a target video is generated based on the modeling result by using a video editor, where the target video is a campus roaming video at a free viewpoint.

In some exemplary embodiments, an operation of generating the target video based on the modeling result includes the following step. An angle trajectory corresponding to the free viewpoint is obtained. The target video is obtained through rendering based on the modeling result and the angle trajectory.

In some optional embodiments, a process of performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain the modeling result include the following steps.

First, grids corresponding to the preset three-dimensional scene are constructed, on multiple scale features based on the image data to be reconstructed and the reconstruction result, to obtain a construction result, where the construction result is used for determining three-dimensional space information and light field information of the preset three-dimensional scene on the multiple scale features.

Then, implicit modeling is performed by using the construction result to obtain the modeling result.

In some optional embodiments, the multiple scale features are multiple different data features obtained through division based on multiple different feature scales. Correspondingly, the construction result includes grids, corresponding to each of the multiple scale features, of the preset three-dimensional scene, grids corresponding to each scale feature include the same grid size, and the multiple scale features correspond to multiple grid sizes. Each of grids corresponding to the preset three-dimensional scene stores corresponding three-dimensional space information and light field information. The three-dimensional space information includes geometric shape data and spatial position data of the preset three-dimensional scene. The light field information is optical representation data (for example, multiple optical parameters, including opacity, luminance, saturation, and the like) of the preset three-dimensional scene. Further, implicit modeling is performed based on an optimized NeRF algorithm by using the construction result to obtain the modeling result.

FIG. 6 is a diagram of an optional implicit modeling process according to Embodiment One of the present disclosure. As shown in FIG. 6, a process of performing implicit modeling in the third stage in FIG. 5 includes the following steps. First, scene occupancy grids corresponding to the preset three-dimensional scene are constructed. Then, a multi-scale feature hash table is determined based on the scene occupancy grid, where the hash table stores encoded data of a multi-scale feature. Further, neural volume rendering optimization is performed based on the multi-scale feature hash table to obtain a modeling result, where an implicit modeling algorithm used in the neural volume rendering optimization includes a large-scale scene sampling policy, a sparsity constraint, and a depth constraint.

In some optional embodiments, a process of constructing, on the multiple scale features based on the image data to be reconstructed and the reconstruction result, the grids corresponding to

the preset three-dimensional scene to obtain the construction result may further include the following steps.

First, a bounding box corresponding to the preset three-dimensional scene is determined based on the image data to be reconstructed and the reconstruction result.

Then, multi-scale rasterizing division is performed on three-dimensional space in the bounding box to obtain a division result.

Finally, the grids corresponding to the preset three-dimensional scene are obtained by using light field information corresponding to the division result, to obtain the construction result.

Still in the examples in which scene reconstruction is performed on a campus based on aerial data of the campus, the bounding box corresponding to the campus is determined based on the aerial data of the campus and the reconstruction result. For example, the bounding box is a polygonal cylinder, a size and a shape of a bottom surface of the polygon are determined based on a floor area of the campus, and a height of the polygon is a preset scene height. For example, for a campus occupying an area with a length of 500 meters and a width of 400 meters, a maximum aerial photography height corresponding to the campus is set to 200 meters. In this case, the bounding box of the campus is a cuboid bounding box of 500 meters × 400 meters × 200 meters.

In some exemplary embodiments, an operation of determining the bounding box corresponding to the preset three-dimensional scene based on the image data to be reconstructed and the reconstruction result includes the following steps. Camera pose information corresponding to the image of the preset three-dimensional scene to be reconstructed are determined based on the image data to be reconstructed. Three-dimensional space information and light field information of the preset three-dimensional scene on the multiple scale features are determined based on the reconstruction result, where the three-dimensional space information includes scene size data of the preset three-dimensional scene. The bounding box corresponding to the preset three-dimensional scene are determined based on the camera pose information, the three-dimensional space information, and the light field information.

In some exemplary embodiments, the division result includes multiple division sub-results, and an operation of performing multi-scale rasterizing division on the three-dimensional space

in the bounding box to obtain the division result includes the following steps. Multiple division scales are determined, and the three-dimensional space is divided in the bounding box based on the multiple division scales to obtain the multiple division sub-results. Still in the example in which scene reconstruction is performed on a campus based on aerial data of the campus, five division scales are respectively determined as follows: 1 meter, 2 meters, 5 meters, 10 meters, and 50 meters. A cuboid bounding box of 500 meters × 400 meters × 200 meters is divided based on each of the five division scales to obtain five division sub-results. For example, when the division scale is 1 meter, the cuboid bounding box of 500 meters × 400 meters × 200 meters is divided into 500 × 400 × 200 grids of 1 meter × 1 meter × 1 meter, to obtain a first division sub-result. For example, when the division scale is 2 meters, the cuboid bounding box of 500 meters × 400 meters × 200 meters is divided into 250 × 200 × 100 grids of 2 meters × 2 meters × 2 meters, to obtain a second division sub-result.

It should be noted that, in the division result, each grid correspondingly stores light field information. For example, in the first division sub-result, grids of 1 meter × 1 meter × 1 meter correspondingly stores light field information of multiple sampling points included in the grid. In the second division sub-result, grids of 2 meters × 2 meters × 2 meters correspondingly stores light field information of multiple sampling points included in the grid. The light field information includes at least one of the following: opacity, luminance, saturation, and the like.

In some exemplary embodiments, the light field information includes the opacity, and an operation of obtaining the grids corresponding to the preset three-dimensional scene by using the light field information corresponding to the division result, to obtain the construction result includes the following steps. Some or all of multiple grids in the division result are determined as target grids based on opacity corresponding to each of the multiple grids. The target grids are determined as the grids corresponding to the preset three-dimensional scene, to obtain the construction result. An operation of selecting the target grids from the multiple grids based on the opacity corresponding to each of the multiple grids in the division result includes the following steps. The opacity corresponding to each of the multiple grids is sequentially determined. A current grid is determined as a target grid in response to the opacity corresponding to the current grid meeting a preset condition. Optionally, the preset condition is that the opacity is greater than a preset threshold. The preset condition is used for

determining whether the current grid is an empty grid that is not occupied by the preset three-dimensional scene.

In some exemplary embodiments, the method for generating the video based on scene reconstruction further includes the following step. The light field information corresponding to the division result is optimized, based on the construction result, to obtain updated light field information. Specifically, multiple rounds of iterative optimization and adjustment are performed, based on the construction result, on the opacity corresponding to each of the multiple grids in the division result, until opacity of the grids corresponding to the preset three-dimensional scene is 1 or is within a preset range (for example, 0.9 to 1). Each of the multiple rounds of iterative optimization and adjustment includes the following steps. The opacity of each grid corresponding to the preset three-dimensional scene in the construction result is increased, and the opacity of each grid in the division result other than the grids corresponding to the preset three-dimensional scene is decreased.

It is easily understood that, after the grids (that is, the scene occupancy grids) corresponding to the preset three-dimensional scene are determined, during scene reconstruction, the scene occupancy grids are sampled, and grids in the division result other than the scene occupancy grid are skipped, to reduce data sampling overheads during the scene reconstruction, thereby improving scene reconstruction efficiency and rendering efficiency of the target video.

According to the steps provided in the optional embodiments, accuracy of each grid, that is determined in the construction result, corresponding to the preset three-dimensional scene is high, and grids, which are unoccupied by the preset three-dimensional scene, among the multiple grids in the division result are less likely or are not to be determined as the grids corresponding to the preset three-dimensional scene, to ensure accuracy of a subsequent scene reconstruction result.

In some optional embodiments, an operation of performing implicit modeling by using the construction result to obtain the modeling result further includes the following steps.

First, the three-dimensional space information and the light field information in the construction result are divided into multiple partitions based on the multiple scale features to obtain a division result, where different partitions in the division result are used for respectively

storing three-dimensional space information and light field information of different scale features.

Then, the construction result is sampled, according to a preset sampling policy, based on a capture angle of an aerial image capture device and a target rendering position to obtain a sampling result, where the preset sampling policy is used for determining a sampling count and a sampling step that are to be used for the target rendering position during sampling and a ray count emitted from the target rendering position.

Then, a first multi-scale feature is obtained from the division result by using the sampling result.

Finally, implicit modeling is performed by using the first multi-scale feature to obtain the modeling result.

In some exemplary embodiments, an operation of dividing the three-dimensional space information and the light field information in the construction result into the multiple partitions based on the multiple scale features to obtain the division result includes the following steps. The multiple partitions are determined based on multiple division scales corresponding to the construction result. The three-dimensional space information and the light field information on the multiple scale features in the construction result are divided into the multiple partitions based on the multiple partitions to obtain the division result. The first multi-scale feature includes multiple groups of features corresponding to the image data to be reconstructed, and the multiple groups of features correspond to the multiple division scales. For example, each division scale corresponds to one group of features in the first multi-scale feature. The multiple partitions is multiple storage areas corresponding to the multiple scale features.

Still in the example in which scene reconstruction is performed on a campus based on aerial data of the campus, five division scales are respectively determined as follows: 1 meter, 2 meters, 5 meters, 10 meters, and 50 meters. Correspondingly, five partitions are determined. When a division scale is 1 meter, a cuboid bounding box of 500 meters × 400 meters × 200 meters is divided into 500 × 400 × 200 grids of 1 meter × 1 meter × 1 meter, and a dimension of a corresponding first partition is 500 × 400 × 200. When a division scale is 2 meters, a cuboid bounding box of 500 meters × 400 meters × 200 meters is divided into 250 × 200 × 100 grids of 2 meters × 2 meters × 2 meters, and a dimension of a corresponding second partition

is 250 × 200 × 100. That is, for a division sub-result including N1 × N2 × N3 grids, a dimension of a corresponding partition is N1 × N2 × N3.

In some exemplary embodiments, the multiple partitions is represented as a multi-scale feature hash table. For example, each partition corresponds to a hash table of one feature scale. As shown in FIG. 6, the first multi-scale feature is a multi-scale feature hash table, and the multi-scale feature hash table includes multiple feature tables in different dimensions. A feature table 1, a feature table 2, ..., and a feature table n shown in FIG. 6 correspond to n grid division scales.

It is easily understood that, the three-dimensional space information and the light field information on the multiple scale features in the construction result are divided into the multiple partitions by using the multi-scale feature hash table, so that three-dimensional space information and light field information on different scale features can be conveniently and flexibly accessed by using the multi-scale feature hash table during scene reconstruction, to enhance a representation capability of feature coding during scene reconstruction.

In some exemplary embodiments, the preset sampling policy is obtained according to a requirement in an application scenario. The sampling count and the sampling step that are to be used for the target rendering position during sampling and the ray count emitted from the target rendering position are determined according to the preset sampling policy. The construction result is sampled based on the capture angle of the aerial image capture device, the target rendering position, the sampling herein, and the sampling step along a direction of each of multiple rays emitted from the target rendering position, to obtain the sampling result. A direction of each ray emitted from the target rendering position is determined based on the target rendering position and the capture angle of the aerial image capture device. For example, the aerial image capture device is an aerial camera. The target sampling position is a current rendering pixel during rendering.

It should be noted that sampling the construction result includes: sampling grids through which each ray passes among grids occupied by the preset three-dimensional scene in the construction result. A sampling result corresponding to each ray includes multiple sampling points.

In some exemplary embodiments, an operation of obtaining the first multi-scale feature from the division result by using the sampling result includes the following steps. The first multi-

scale feature is obtained from the division result by using multiple sampling points corresponding to each of the multiple rays in the sampling result based on position information of the sampling points. Specifically, the division result is a multi-scale feature hash table. Feature values of multiple scales are obtained from the multi-scale feature hash table based on position information of each sampling point, and then the feature values of the multiple scales are combined to obtain a first multi-scale feature of the sampling point. Further, implicit modeling is performed by using the first multi-scale feature to obtain the modeling result.

It is easily understood that, in the embodiments of the present disclosure, during implicit modeling by using the construction result, the implicit modeling based on the first multi-scale feature is implemented in the partition-based storage manner, the sampling manner determined according to the preset sampling policy, and the manner of obtaining the first multi-scale feature. Compared with the method for generating the video based on scene reconstruction in the related art, in the embodiments of the present disclosure, scene feature sampling accuracy is high, optimization effect of the implicit modeling is strong, and a more accurate and real scene reconstruction result can be obtained.

In some optional embodiments, an operation of performing implicit modeling by using the first multi-scale feature to obtain the modeling result may further include the following steps.

First, pixel value prediction is performed on the first multi-scale feature by using a target fully-connected neural network model to obtain a target pixel value, where the target fully-connected neural network model is obtained through deep learning training by using multiple groups of data, and each group of data includes a training image and a second multi-scale feature corresponding to the training image.

Then, the modeling result is generated based on the target pixel value.

The second multi-scale feature includes multiple groups of features corresponding to the training image, and the multiple groups of features correspond to the multiple division scales. For example, each division scale corresponds to one group of features in the second multi-scale feature.

In the optional embodiments, the target fully-connected neural network model is an MLP. The MLP is obtained through deep learning training by using the multiple groups of data, and each

groups of data includes the training image and the second multi-scale feature corresponding to the training image. During training of the MLP, for a manner of obtaining the second multi-scale feature, refer to the step of performing implicit modeling by using the construction result corresponding to the image data to be reconstructed to obtain the modeling result.

As shown in FIG. 6, neural volume rendering optimization is performed based on the multi-scale feature hash table to obtain the modeling result. Specifically, a large-scale sampling policy is determined for a large-scale scene (for example, a campus scene, a city scene, or a district scene). Further, scene occupancy grids corresponding to the construction result are sampled according to the large-scale sampling policy to obtain the sampling result, and the first multi-scale feature is obtained from the multi-scale feature hash table based on the sampling result. Then, pixel value prediction is performed on the first multi-scale feature by using the MLP to obtain the target pixel value, where the target pixel value includes an RGB attribute value of a pixel. The modeling result is generated based on the target pixel value.

In some optional embodiments, the method for generating the video based on scene reconstruction further includes the following method steps.

In step S451, pixel value prediction is performed on the second multi-scale feature by using an initial fully-connected neural network model to obtain a training pixel value.

In step S452, a first loss is determined based on the training pixel value and a real pixel value corresponding to the training image.

In step S453, model parameters of the initial fully-connected neural network model are updated based on the first loss to obtain the target fully-connected neural network model.

For example, the process of step S451 to step S453 is further specifically shown in FIG. 6. In a process of neural volume rendering optimization, during training of the MLP, an initial MLP is updated by using the first loss determined based on the training pixel value and the real pixel value, to obtain a target MLP. The training pixel value is a pixel value obtained by the initial MLP by performing pixel value prediction on the second multi-scale feature corresponding to the training image, and the real pixel value is a real pixel value corresponding to the training image. For example, a calculation manner for the first loss is an absolute value of a difference between the training pixel value and the real pixel value.

According to the method provided in step S451 to step S453 in the embodiments of the present disclosure, the target MLP to be used is trained during scene reconstruction, so that model performance of the target MLP can be enhanced, to improve prediction accuracy of pixel value prediction performed on the first multi-scale feature.

In some optional embodiments, the method for generating the video based on scene reconstruction further includes the following method steps.

In step S461, depth value prediction is performed on the second multi-scale feature by using the initial fully-connected neural network model to obtain a training depth value.

In step S462, a second loss is determined based on the training depth value and a real depth value corresponding to the training image.

In step S463, a sampling ray is determined by using the capture angle and the target rendering position.

In step S464, a third loss is determined based on a sampling weight, on the sampling ray, of a sampling point in the sampling result.

In step S465, model parameters of the target fully-connected neural network model are updated based on at least one of the second loss and the third loss to obtain an updated target fully-connected neural network model.

For example, the process of step S461 to step S465 is further specifically shown in FIG. 6. During training of an initial MLP, the second loss is determined based on a depth constraint by using the training depth value and the real depth value. The third loss is determined based on a sparsity constraint by using the sampling ray, the sampling result, and the sampling weight that correspond to the second multi-scale feature. Model parameter adjustment is performed on the initial MLP based on at least one of the second loss and the third loss to obtain a target MLP, or model parameter update is performed on the target MLP based on at least one of the second loss and the third loss to obtain an updated target MLP.

It should be noted that the depth constraint and the sparsity constraint are constraint rules preset for a large-scale scene (for example, a campus scene, a city scene, or a district scene). The depth constraint is used for removing some grids (for example, grids that is high up in the air that is irrelevant to scene reconstruction), to reduce possible incorrect floating objects in a

scene reconstruction result. The sparsity constraint is used for determining that sampling weights of multiple sampling points on each sampling ray are concentrated in a specified range, to further reduce possible incorrect floating objects in the scene reconstruction result.

For example, the depth constraint is a depth map constraint based on a preset depth map. An operation of determining the second loss based on the depth constraint by using the training depth value and the real depth value includes the following steps. The second loss is determined based on the preset depth map, the training depth value obtained through performing, by the initial MLP, depth value prediction on the second multi-scale feature, and the real depth value corresponding to the training image. A calculation manner for the second loss is an absolute value of a difference between the training depth value and the real depth value when the depth map constraint of the preset depth map is satisfied.

For example, the sparsity constraint is a weight threshold constraint. A calculation manner for the third loss is, for each sampling ray, when sampling weights of multiple sampling points on the sampling ray satisfy the weight threshold constraint, calculating entropy losses of the sampling weights of the multiple sampling points.

According to the method provided in step S461 to step S465 in the embodiments of the present disclosure, the target MLP to be used is trained during scene reconstruction based on the depth constraint and the sparsity constraint that are designed for the large-scale scene, so that model performance of the target MLP can be enhanced, to improve prediction accuracy of pixel value prediction performed on the first multi-scale feature.

In some optional embodiments, an operation of obtaining the image data to be reconstructed in step S41 includes the following method steps.

In step S411, aerial data capturing is performed on the preset three-dimensional scene by using an aerial image capture device under a preset condition to obtain the image data to be reconstructed, where the preset condition is used for determining a weather condition and a flight condition of the aerial image capture device during aerial data capturing.

In the optional embodiments, the preset condition is further used for determining to perform scene modeling by using a neural implicit modeling method. For example, in an application scenario of performing scene reconstruction on a campus based on aerial image data, the preset

condition includes at least the following. Weather during photographing is overcast. A flight altitude of the aerial image capture device is 110 meters. A lateral overlap rate of the aerial image capture device is 80%. A longitudinal overlap rate of the aerial image capture device is 85%. A pan-tilt-zoom tilt angle of the aerial image capture device is –50 degrees. When photographing a preset key scene area, the aerial image capture device performs photographing at a preset short distance.

In some optional embodiments, an operation of performing sparse reconstruction on the image data to be reconstructed to obtain the reconstruction result in step S42 includes the following method steps.

In step S421, feature extraction is performed on the image data to be reconstructed to obtain an extraction result.

In step S422, feature matching is performed on the extraction result to obtain a matching result.

In step S423, sparse reconstruction is performed on the matching result to obtain the reconstruction result, where the reconstruction result is used for determining pose information, corresponding to the image data to be reconstructed, of the aerial image capture device.

Still as shown in FIG. 5, in the second stage, during SfM sparse reconstruction, the following process is performed based on the captured aerial data: feature extraction, feature matching, and sparse reconstruction, to solve the pose information of the aerial image capture device based on the aerial data.

In some optional implementations of step S421, performing feature extraction on the image data to be reconstructed to obtain the extraction result includes the following method steps.

In step S4211, the image data to be reconstructed is preprocessed to obtain a preprocessing result.

In step S4212, feature extraction is performed on the preprocessing result to obtain the extraction result.

In some exemplary embodiments, an operation of preprocessing the image data to be reconstructed to obtain the preprocessing result includes the following steps. Adaptive adjustment is performed based on a resolution of an image in the image data to be reconstructed to obtain an adjustment result. Exchangeable image file format (EXIF) information

corresponding to an image in the image data to be reconstructed is written to the adjustment result, to obtain the preprocessing result, where the EXIF information includes a device manufacturer, a device model, a software version number, an aperture, a shutter speed, a focal length, sensitivity, a metering mode, and the like.

In some exemplary embodiments, the extraction result includes an image point feature and an image descriptor, and the image descriptor is used for determining a local feature of an image. An operation of performing feature extraction on the preprocessing result to obtain the extraction result includes the following steps. Point feature extraction is performed on each image in the preprocessing result to obtain the image point feature. Local feature extraction is performed on each image in the preprocessing result to obtain the image descriptor.

In some optional implementations of step S422, performing feature matching on the extraction result to obtain the matching result includes the following method steps.

In step S4221, image pairing is performed on the extraction result to obtain a pairing result.

In step S4222, feature matching is performed on the pairing result to obtain the matching result.

In some exemplary embodiments, an operation of performing image pairing on the extraction result to obtain the pairing result includes the following steps. Satellite positioning data corresponding to the preset three-dimensional scene is obtained. Image matching is performed on the extraction result based on the satellite positioning data to obtain the pairing result. Specifically, pairing is performed, based on the satellite positioning data and the image point feature and the image descriptor that are in the extraction result, between a current image corresponding to the extraction result and another scene image (which is selected from the preprocessing result) that is close to the current image. Therefore, after pairing is performed between every two of multiple images corresponding to the extraction result, inter-image constraint calculation is performed on two paired images.

In some exemplary embodiments, an operation of performing feature matching on the pairing result to obtain the matching result includes the following steps. Image feature point matching is performed, based on the image descriptor, on two paired images in the pairing result, to obtain the matching result. The matching result is used for calculating a relative rotation relationship and a relative translation relationship between the two images that are paired.

In some optional implementations of step S423, performing sparse reconstruction on the matching result to obtain the reconstruction result includes the following method steps.

In step S4231, geometrical filtering is performed on the matching result to obtain a filtering result.

In step S4232, global three-dimensional structural estimation is performed on the filtering result to obtain an estimation result.

In step S4233, the reconstruction result is determined based on the estimation result.

In some exemplary embodiments, an operation of performing geometrical filtering on the matching result to obtain the filtering result includes the following steps. Matching filtering is determined on the matching result based on a preset geometrical constraint to filter out an incorrectly matched image feature point from the matching result, to obtain the filtering result. For example, the preset geometrical constraint includes an essential matrix constraint or a fundamental matrix constraint.

In some exemplary embodiments, an operation of performing global three-dimensional structural estimation on the filtering result to obtain the estimation result includes the following steps. A relative rotation relationship and a relative translation relationship between two paired images are calculated based on the filtering result. Rotation averaging information corresponding to the two paired images is determined based on the relative rotation relationship. Translation averaging information corresponding to the two paired images is determined based on the relative translation relationship. An estimation result of a global three-dimensional structure is calculated based on the rotation averaging information and the translation averaging information by using a global SfM algorithm, where the estimation result includes internal parameters of the aerial image capture device and a capture point pose corresponding to the aerial image capture device. The global SfM algorithm is obtained by optimizing a conventional SfM algorithm, and the global SfM algorithm is an algorithm based on the rotation averaging information and the translation averaging information.

For example, an operation of determining, based on the relative rotation relationship, the rotation averaging information corresponding to the two paired images includes the following step. A rotation optimization problem is constructed by using a rotation constraint condition

corresponding to the relative rotation relationship based on a vector norm (an L2 norm is used in this example). The rotation optimization problem is solved to obtain the rotation averaging information corresponding to the two paired images.

For example, an operation of determining, based on the relative translation relationship, the translation averaging information corresponding to the two paired images includes the following steps. A translation optimization problem is constructed by using a translation constraint condition corresponding to the relative translation relationship based on a vector norm (an L1 norm is used in this example). The translation optimization problem is solved to obtain the translation averaging information corresponding to the two paired images.

To further improve accuracy of the reconstruction result, in some optional implementations of step S4233, determining the reconstruction result based on the estimation result includes the following method steps.

In step S4234, horizon correction is performed on the estimation result to obtain a first correction result.

In step S4235, heading angle correction is performed on the first correction result to obtain a second correction result.

In step S4236, point cloud partitioning processing is performed on the second correction result to obtain a partitioning result.

In step S4237, data format adjustment is performed on the partitioning result to obtain the reconstruction result.

In some exemplary embodiments, an operation of performing horizon correction on the estimation result to obtain the first correction result includes the following steps. Robust plane fitting is performed based on the internal parameter of the aerial image capture device in the estimation result to obtain a target robust plane. Rotation estimation is performed by using a normal direction of the target robust plane to obtain a rotation matrix, where the rotation matrix is configured to determine a rotation deviation angle between the aerial image capture device and a real ground plane. Horizon correction is performed on the estimation result based on the rotation matrix to obtain the first correction result.

In some exemplary embodiments, an operation of performing heading angle correction on the first correction result to obtain the second correction result includes the following steps. Heading angle correction is performed on the first correction result based on the capture point pose corresponding to the aerial image capture device in the estimation result to obtain the second correction result. Through the heading angle correction, it is ensured that a small bounding box of the preset three-dimensional scene is determined in a subsequent process, to improve accuracy of a scene reconstruction result. The second correction result includes a point cloud distributed in a preset three-dimensional scene obtained through the heading angle correction. The point cloud is configured to store point cloud position information, the internal parameter of the aerial image capture device, and the capture point pose corresponding to the aerial image capture device. Each point in the point cloud corresponds to one aerial image capture device.

In some exemplary embodiments, an operation of performing point cloud partitioning processing on the second correction result to obtain the partitioning result includes the following steps. Overlappable partitioning is performed on the point cloud in the second correction result to obtain the partitioning result. The overlappable partitioning means that a point in the point cloud is distributed to at least one point cloud block. The point cloud partitioning processing is performed on the second correction result to further improve a capability of reconstructing a large-scale scene in the embodiments of the present disclosure.

In some exemplary embodiments, an operation of performing data format adjustment on the partitioning result to obtain the reconstruction result includes the following steps. Point cloud data corresponding to multiple point cloud blocks in the partitioning result is adjusted, based on a preset data format, to obtain the reconstruction result, where the preset data format is a unified output format corresponding to the point cloud position information, the internal parameter of the aerial image capture device, and the capture point pose corresponding to the aerial image capture device.

In some optional embodiments, generating the target video by using the modeling result in step S44 includes the following method steps.

In step S441, trajectory information corresponding to the aerial image capture device is obtained.

In step S442, video rendering is performed on the modeling result based on the trajectory information to obtain the target video.

In some exemplary embodiments, the target video is generated by using an image editing tool based on the modeling result. An image rendering trajectory corresponding to the aerial image capture device is determined in the image editing tool, and video rendering is performed on the modeling result based on the image rendering trajectory to obtain the target video. Specifically, an operation of determining, in the image editing tool, the image rendering trajectory corresponding to the aerial image capture device includes the following steps. A pre-drawn image rendering trajectory is obtained. Video rendering preview is performed based on the image rendering trajectory to obtain a preview result. It is determined, based on the preview result, whether the image rendering trajectory is available. The pre-drawn image rendering trajectory is freely drawn by a user, or is a randomly generated free roaming trajectory. Therefore, the target video obtained by performing video rendering on the modeling result based on the trajectory information is the scene roaming video of the preset three-dimensional scene at the free viewpoint.

It is easily understood that, according to the method for generating the video based on scene reconstruction provided in the embodiments of the present disclosure, image data–based scene reconstruction is performed by using an improved neural implicit mesh model, to obtain a scene roaming video with a strong sense of reality through rendering. In particular, the method for generating the video based on scene reconstruction is especially applicable to large-scale scene reconstruction based on aerial data.

It should be noted that user information (including but not limited to user equipment information, personal information of users, and the like) and data (including but not limited to data for analysis, stored data, displayed data, and the like) in the present disclosure are information and data used under authorization by users or full authorization by all parties. In addition, collection, use, and processing of related data need to comply with related laws, regulations, and standards in related countries and regions, and a corresponding operation entry is provided for a user to choose to grant a permission or reject.

It should be noted that, for ease of description, the method embodiments are described as a series of combinations of actions. However, a person having ordinary skill in the art should be

aware that the present disclosure is not limited to the described order of the actions, because some steps is performed in another order or simultaneously according to the present disclosure. In addition, a person having ordinary skill in the art should also be aware that the embodiments described in the present disclosure are all exemplary embodiments, and the described actions and modules are not necessarily required in the present disclosure.

According to the descriptions of the implementations, a person having ordinary skill in the art may clearly understand that the methods in the embodiments are implemented by software in combination with a necessary common hardware platform, or certainly are implemented by hardware. Based on such an understanding, the technical solutions of the present disclosure essentially or the part contributing to the related art are implemented in a form of a software product. The computer software product is stored in a storage medium (for example, a read-only memory (ROM), a random access memory (RAM), a magnetic disk, or a compact disc), and includes several instructions for instructing a terminal device (which is a mobile phone, a computer, a server, a network device, or the like) to perform the methods described in the embodiments of the present disclosure.

Embodiment Two

In the running environment in Embodiment One, the present disclosure provides another method for generating a video based on scene reconstruction, as shown in FIG. 7. FIG. 7 is a flowchart of a method for generating a video based on scene reconstruction according to Embodiment Two of the present disclosure. As shown in FIG. 7, the method for generating the video based on scene reconstruction includes the following steps.

In step S71, image data to be reconstructed is displayed on an operation interface in response to an input instruction performed on the operation interface, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed.

In step S72, a target video is displayed on the operation interface in response to a reconstruction instruction performed on the operation interface, where the target video is used for performing scene roaming on the preset three-dimensional scene, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on the image

data to be reconstructed and a reconstruction result, and the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed.

The operation interface is a GUI of a client that performs the method for generating a video based on scene reconstruction. The input instruction is generated based on an input operation performed on the operation interface, and the input operation is a touch operation or an operation implemented by using a device such as a keyboard, a mouse, or a joystick. The image data to be reconstructed is displayed on the operation interface when the input instruction performed on the operation interface is detected. The reconstruction instruction is generated based on a reconstruction operation performed on the operation interface. The reconstruction operation is a touch operation or an operation implemented by using a device such as a keyboard, a mouse, or a joystick (for example, by clicking an OK button). The target video is displayed on the operation interface when the reconstruction instruction performed on the operation interface is detected. The target video is generated by the client according to a specific implementation of the method for generating a video based on scene reconstruction provided in Embodiment One of the present disclosure. The target video may alternatively be generated by a server corresponding to a request of the client, and the server generates the target video according to a specific implementation of the method for generating a video based on scene reconstruction provided in Embodiment One of the present disclosure.

The method for generating the video based on scene reconstruction is applied to but is not limited to the following scenarios: a visual navigation scenario (for example, constructing a three-dimensional virtual scene based on a map to perform autonomous navigation for a robot or an uncrewed aerial vehicle), a VR/AR scene production scenario (for example, VR exhibition hall production or VR/AR game scene design), a three-dimensional printing scenario (for example, reconstructing a virtual model based on an image for entity production), a three-dimensional movie and television creation and post-production scenario (for example, reconstructing a three-dimensional scene or a virtual object in a three-dimensional scene based on a given image), and a cultural tourism promotion scenario (for example, reconstructing a three-dimensional virtual scene of an area such as a scenic spot, a campus, an airport, a station, or a road, or producing a cultural tourism promotion video). The image data to be reconstructed is used for recording an image (for example, an aerial image) of a preset three-dimensional scene to be reconstructed in the application scenario. The target video is a scene roaming video,

at a free viewpoint, corresponding to the preset three-dimensional scene to be reconstructed in the application scenario.

Therefore, according to Embodiment Two of the present disclosure, a user can perform a control operation on an operation interface of a client to generate at least one of an input instruction and a reconstruction instruction, specify image data to be reconstructed, and trigger the client to perform the following steps. Sparse reconstruction is performed on the image data to be reconstructed to obtain a reconstruction result. Implicit modeling is performed based on the image data to be reconstructed and the reconstruction result to obtain a modeling result. A target video is generated by using the modeling result. The target video is displayed on the operation interface. In the process, the user has a high degree of freedom and flexibility, and can customize a target video corresponding to personalized three-dimensional scene reconstruction according to a requirement of an application scenario. The method provided in the embodiments of the present disclosure is conducive to application in an actual scenario.

It should be noted that performing sparse reconstruction on the image data to be reconstructed to obtain the reconstruction result, performing implicit modeling on the image data to be reconstructed and the reconstruction result to obtain the modeling result, and generating the target video by using the modeling result further include other method steps. Refer to related descriptions in Embodiment One of the present disclosure. Details are not described again.

Embodiment Three

In the running environment in Embodiment One, the present disclosure provides another method for generating a video based on scene reconstruction, as shown in FIG. 8. FIG. 8 is a flowchart of a method for generating a video based on scene reconstruction according to Embodiment Three of the present disclosure. As shown in FIG. 8, the method for generating the video based on scene reconstruction includes the following steps.

In step S81, an initial default picture is displayed on a displayed picture of a VR device or an AR device.

In step S82, the VR device or the AR device is driven to jump from displaying the initial default picture to displaying a target video, where the target video is used for performing scene roaming on a preset three-dimensional scene to be reconstructed, the target video is generated

by using a modeling result, the modeling result is obtained by performing implicit modeling on image data to be reconstructed and a reconstruction result, the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed, and the image data to be reconstructed is used for recording an image of the preset three-dimensional scene.

The method for generating the video based on scene reconstruction is applied to but is not limited to the following scenarios: a VR/AR visual navigation scenario (for example, constructing a three-dimensional virtual scene based on a map, and performing VR/AR autonomous navigation for a robot or an uncrewed aerial vehicle based on the three-dimensional virtual scene), a VR/AR scene production scenario (for example, VR exhibition hall production or VR/AR game scene design), a VR/AR three-dimensional movie and television creation and post-production scenario (for example, reconstructing a three-dimensional scene or a virtual object in a three-dimensional scene based on a given image, and providing a reconstruction result for an audience through a VR/AR device), and a VR/AR cultural tourism promotion scenario (for example, reconstructing a three-dimensional virtual scene of an area such as a scenic spot, a campus, an airport, a station, or a road, and providing a reconstruction result for a traveler through a VR/AR device). The image data to be reconstructed is used for recording an image (for example, an aerial image) of a preset three-dimensional scene to be reconstructed in the application scenario. The target video is a scene roaming video, at a free viewpoint, corresponding to the preset three-dimensional scene to be reconstructed in the application scenario.

The initial default picture is a scene picture displayed by the VR device or the AR device at a current moment. The scene picture displayed at the current moment is a historical target video generated by performing scene reconstruction on the preset three-dimensional scene. When a switching instruction is detected, the VR device or the AR device is driven to jump from displaying the initial default picture to displaying the target video. The switching instruction is used for determining to display the target video to be displayed on the display picture of the VR device or the AR device. The target video is obtained by a processor associated with the VR device or the AR device by performing the following steps. Sparse reconstruction is performed on the image data to be reconstructed to obtain the reconstruction result. Implicit modeling is performed based on the image data to be reconstructed and the reconstruction

result to obtain the modeling result. The target video is generated by using the modeling result. The target video is displayed on an operation interface.

According to the virtual model display method in the embodiments of the present disclosure, in a VR or AR scenario, a scene reconstruction result with high accuracy is generated, and a scene roaming video with a strong sense of reality is generated, so that a scene roaming video, at a free viewpoint, corresponding to the preset three-dimensional scene is dynamically and accurately displayed on the VR device or the AR device. In particular, the method for generating the video based on scene reconstruction is especially applicable to large-scale scene reconstruction based on aerial data.

It should be noted that, for an exemplary implementation of the embodiments, reference is made to related descriptions in Embodiment One or Embodiment Two. Details are not described herein again.

Embodiment Four

According to embodiments of the present disclosure, an embodiment of an apparatus configured to implement the method for generating a video based on scene reconstruction is further provided. FIG. 9 is a structural schematic diagram of an apparatus for generating a video based on scene reconstruction according to Embodiment Four of the present disclosure. As shown in FIG. 9, the apparatus includes: an obtaining module 901, configured to obtain image data to be reconstructed, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; a reconstruction module 902, configured to perform sparse reconstruction on the image data to be reconstructed to obtain a reconstruction result; a modeling module 903, configured to perform implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain a modeling result; and a generation module 904, configured to generate a target video by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene.

Optionally, the modeling module 903 is further configured to: construct, on multiple scale features based on the image data to be reconstructed and the reconstruction result, grids corresponding to the preset three-dimensional scene to obtain a construction result, where the construction result is used for determining three-dimensional space information and light field

information of the preset three-dimensional scene on the multiple scale features; and perform implicit modeling by using the construction result to obtain the modeling result.

Optionally, the modeling module 903 is further configured to: determine a bounding box corresponding to the preset three-dimensional scene based on the image data to be reconstructed and the reconstruction result; perform multi-scale rasterizing division on three-dimensional space in the bounding box to obtain a division result; and obtain the grids corresponding to the preset three-dimensional scene by using light field information corresponding to the division result, to obtain the construction result.

Optionally, the modeling module 903 is further configured to: divide the three-dimensional space information and the light field information in the construction result into multiple partitions based on the multiple scale features to obtain a division result, where different partitions in the division result are used for respectively storing three-dimensional space information and light field information of different scale features; sample, according to a preset sampling policy, the construction result based on a capture angle of an aerial image capture device and a target rendering position to obtain a sampling result, where the preset sampling policy is used for determining a sampling count and a sampling step that are to be used for the target rendering position during sampling and a ray count emitted from the target rendering position; obtain a first multi-scale feature from the division result by using the sampling result; and perform implicit modeling by using the first multi-scale feature to obtain the modeling result.

Optionally, the modeling module 903 is further configured to: perform pixel value prediction on the first multi-scale feature by using a target fully-connected neural network model to obtain a target pixel value, where the target fully-connected neural network model is obtained through deep learning training by using multiple groups of data, and each of the multiple groups of data include a training image and a second multi-scale feature corresponding to the training image; and generate the modeling result based on the target pixel value.

Optionally, FIG. 10 is a structural schematic diagram of an optional apparatus for generating a video based on scene reconstruction according to Embodiment Four of the present disclosure. As shown in FIG. 10, in addition to all the modules shown in FIG. 9, the apparatus further includes: a first update module 905, configured to: perform pixel value prediction on the

second multi-scale feature by using an initial fully-connected neural network model to obtain a training pixel value; determine a first loss based on the training pixel value and a real pixel value corresponding to the training image; and update model parameters of the initial fully-connected neural network model based on the first loss to obtain the target fully-connected neural network model.

Optionally, FIG. 11 is a structural schematic diagram of another optional apparatus for generating a video based on scene reconstruction according to Embodiment Four of the present disclosure. As shown in FIG. 11, in addition to all the modules shown in FIG. 10, the apparatus further includes: a second update module 906, configured to: perform depth value prediction on the second multi-scale feature by using the initial fully-connected neural network model to obtain a training depth value; determine a second loss based on the training depth value and a real depth value corresponding to the training image; determine a sampling ray by using the capture angle and the target rendering position; determine a third loss based on a sampling weight, on the sampling ray, of a sampling point in the sampling result; and update model parameters of the target fully-connected neural network model based on at least one of the second loss and the third loss to obtain an updated target fully-connected neural network model.

Optionally, the obtaining module 901 is further configured to: perform aerial data capturing on the preset three-dimensional scene by using an aerial image capture device under a preset condition to obtain the image data to be reconstructed, where the preset condition is used for determining a weather condition and a flight condition of the aerial image capture device during aerial data capturing.

Optionally, the reconstruction module 902 is further configured to: perform feature extraction on the image data to be reconstructed to obtain an extraction result; perform feature matching on the extraction result to obtain a matching result; and perform sparse reconstruction on the matching result to obtain the reconstruction result, where the reconstruction result is used for determining pose information, corresponding to the image data to be reconstructed, of the aerial image capture device.

Optionally, the reconstruction module 902 is further configured to: preprocess the image data to be reconstructed to obtain a preprocessing result; and perform feature extraction on the preprocessing result to obtain the extraction result.

Optionally, the reconstruction module is further configured to: perform image pairing on the extraction result to obtain a pairing result; and perform feature matching on the pairing result to obtain the matching result.

Optionally, the reconstruction module 902 is further configured to: perform geometrical filtering on the matching result to obtain a filtering result; perform global three-dimensional structural estimation on the filtering result to obtain an estimation result; and determine the reconstruction result based on the estimation result.

Optionally, the reconstruction module 902 is further configured to: perform horizon correction on the estimation result to obtain a first correction result; perform heading angle correction on the first correction result to obtain a second correction result; perform point cloud partitioning processing on the second correction result to obtain a partitioning result; and perform data format adjustment on the partitioning result to obtain the reconstruction result.

Optionally, the generation module 904 is further configured to: obtain trajectory information corresponding to the aerial image capture device; and perform video rendering on the modeling result based on the trajectory information to obtain the target video.

In the embodiments of the present disclosure, the image data to be reconstructed is obtained, where the image data to be reconstructed is used for recording the image of the preset three-dimensional scene to be reconstructed; sparse reconstruction is further performed on the image data to be reconstructed to obtain the reconstruction result; the modeling result is obtained by performing implicit modeling based on the image data to be reconstructed and the reconstruction result; and then the target video is generated by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene. Therefore, in the embodiments of the present disclosure, an objective of obtaining a scene roaming video through rendering by performing scene modeling based on image data to be reconstructed in a manner of sparse reconstruction and implicit modeling is achieved, so that technical effect of improving accuracy of a scene modeling result and a sense of reality of a scene roaming video is achieved, and technical problems of many defects in a modeling result

and a poor sense of reality of the scene roaming video in the related art due to scene reconstruction performed through explicit modeling are further solved.

It should be noted herein that the obtaining module 901, the reconstruction module 902, the modeling module 903, and the generation module 904 correspond to step S41 to step S44 in Embodiment One. Implementation examples and application scenarios of the four modules are the same as those of corresponding steps, but are not limited to the content disclosed in Embodiment One. It should be noted that the modules or units is hardware components or software components that are stored in a memory (for example, the memory 104) and that are processed by at least one processors (for example, the processors 102a, 102b, ..., and 102n). The modules may alternatively serve as a part of an apparatus and run in the computer terminal 10 provided in Embodiment One.

It should be noted that, for an exemplary implementation of the embodiments, reference is made to related descriptions in Embodiment One. Details are not described herein again.

Embodiment Five

According to embodiments of the present disclosure, a computer terminal is further provided. The computer terminal is any computer terminal device in a computer terminal group. Optionally, in the embodiments, the computer terminal may alternatively be replaced with a terminal device such as a mobile terminal.

Optionally, in the embodiments, the computer terminal is located in at least one of multiple network devices in a computer network.

In the embodiments, the computer terminal may execute program code of the following steps in a method for generating a video based on scene reconstruction: obtaining image data to be reconstructed, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; performing sparse reconstruction on the image data to be reconstructed to obtain a reconstruction result; performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain a modeling result; and generating a target video by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene.

Optionally, FIG. 12 is a structural block diagram of a computer terminal according to Embodiment Five of the present disclosure. As shown in FIG. 12, the computer terminal 120 may include at least one (one is shown in the figure) processor 1202, a memory 1204, a storage controller 1206, and a peripheral interface 1208. The peripheral interface 1208 is connected with an RF module, an audio module, and a display.

The memory 1204 is configured to store a software program and a module, for example, a program instruction/module corresponding to the method and the apparatus for generating a video based on scene reconstruction in the embodiments of the present disclosure. The processor runs the software program and the module stored in the memory to perform various functional applications and data processing, in other words, implement the method for generating a video based on scene reconstruction. The memory 1204 may include a high-speed random memory, or may include a non-transitory memory, for example, at least one magnetic storage apparatus, a flash memory, or another non-transitory solid-state memory. In some examples, the memory 1204 may further include memories remotely disposed relative to the processor, and the remote memories are connected with the computer terminal 120 through a network. Examples of the network include but are not limited to the Internet, an intranet, a local area network, a mobile communication network, and a combination thereof.

The processor 1202 may invoke, through a transmission apparatus, information and an application program that are stored in the memory, to perform the following steps: obtaining image data to be reconstructed, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; performing sparse reconstruction on the image data to be reconstructed to obtain a reconstruction result; performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain a modeling result; and generating a target video by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene.

Optionally, the processor 1202 may further execute program code of the following steps: constructing, on multiple scale features based on the image data to be reconstructed and the reconstruction result, grids corresponding to the preset three-dimensional scene to obtain a construction result, where the construction result is used for determining three-dimensional

space information and light field information of the preset three-dimensional scene on the multiple scale features; and performing implicit modeling by using the construction result to obtain the modeling result.

Optionally, the processor 1202 may further execute program code of the following steps: determining a bounding box corresponding to the preset three-dimensional scene based on the image data to be reconstructed and the reconstruction result; performing multi-scale rasterizing division on three-dimensional space in the bounding box to obtain a division result; and obtaining the grids corresponding to the preset three-dimensional scene by using light field information corresponding to the division result, to obtain the construction result.

Optionally, the processor 1202 may further execute program code of the following steps: dividing the three-dimensional space information and the light field information in the construction result into multiple partitions based on the multiple scale features to obtain a division result, where different partitions in the division result are used for respectively storing three-dimensional space information and light field information of different scale features; sampling, according to a preset sampling policy, the construction result based on a capture angle of an aerial image capture device and a target rendering position to obtain a sampling result, where the preset sampling policy is used for determining a sampling count and a sampling step that are to be used for the target rendering position during sampling and a ray count emitted from the target rendering position; obtaining a first multi-scale feature from the division result by using the sampling result; and performing implicit modeling by using the first multi-scale feature to obtain the modeling result.

Optionally, the processor 1202 may further execute program code of the following steps: performing pixel value prediction on the first multi-scale feature by using a target fully-connected neural network model to obtain a target pixel value, where the target fully-connected neural network model is obtained through deep learning training by using multiple groups of data, and each of the multiple groups of data include a training image and a second multi-scale feature corresponding to the training image; and generating the modeling result based on the target pixel value.

Optionally, the processor 1202 may further execute program code of the following steps: performing pixel value prediction on the second multi-scale feature by using an initial fully-

connected neural network model to obtain a training pixel value; determining a first loss based on the training pixel value and a real pixel value corresponding to the training image; and updating model parameters of the initial fully-connected neural network model based on the first loss to obtain the target fully-connected neural network model.

Optionally, the processor 1202 may further execute program code of the following steps: performing depth value prediction on the second multi-scale feature by using the initial fully-connected neural network model to obtain a training depth value; determining a second loss based on the training depth value and a real depth value corresponding to the training image; determining a sampling ray by using the capture angle and the target rendering position; determining a third loss based on a sampling weight, on the sampling ray, of a sampling point in the sampling result; and updating model parameters of the target fully-connected neural network model based on at least one of the second loss and the third loss to obtain an updated target fully-connected neural network model.

Optionally, the processor 1202 may further execute program code of the following steps: performing aerial data capturing on the preset three-dimensional scene by using an aerial image capture device under a preset condition to obtain the image data to be reconstructed, where the preset condition is used for determining a weather condition and a flight condition of the aerial image capture device during aerial data capturing.

Optionally, the processor 1202 may further execute program code of the following steps: performing feature extraction on the image data to be reconstructed to obtain an extraction result; performing feature matching on the extraction result to obtain a matching result; and performing sparse reconstruction on the matching result to obtain the reconstruction result, where the reconstruction result is used for determining pose information, corresponding to the image data to be reconstructed, of the aerial image capture device.

Optionally, the processor 1202 may further execute program code of the following steps: preprocessing the image data to be reconstructed to obtain a preprocessing result; and performing feature extraction on the preprocessing result to obtain the extraction result.

Optionally, the processor 1202 may further execute program code of the following steps: performing image pairing on the extraction result to obtain a pairing result; and performing feature matching on the pairing result to obtain the matching result.

Optionally, the processor 1202 may further execute program code of the following steps: performing geometrical filtering on the matching result to obtain a filtering result; performing global three-dimensional structural estimation on the filtering result to obtain an estimation result; and determining the reconstruction result based on the estimation result.

Optionally, the processor 1202 may further execute program code of the following steps: performing horizon correction on the estimation result to obtain a first correction result; performing heading angle correction on the first correction result to obtain a second correction result; performing point cloud partitioning processing on the second correction result to obtain a partitioning result; and performing data format adjustment on the partitioning result to obtain the reconstruction result.

Optionally, the processor 1202 may further execute program code of the following steps: obtaining trajectory information corresponding to the aerial image capture device; and performing video rendering on the modeling result based on the trajectory information to obtain the target video.

The processor 1202 may invoke, through a transmission apparatus, information and an application program that are stored in the memory, to perform the following steps: displaying image data to be reconstructed on an operation interface in response to an input instruction performed on the operation interface, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; and displaying a target video on the operation interface in response to a reconstruction instruction performed on the operation interface, where the target video is used for performing scene roaming on the preset three-dimensional scene, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on the image data to be reconstructed and a reconstruction result, and the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed.

The processor 1202 may invoke, through a transmission apparatus, information and an application program that are stored in the memory, to perform the following steps: displaying an initial default picture on a displayed picture of a VR device or an AR device; and driving the VR device or the AR device to jump from displaying the initial default picture to displaying a target video, where the target video is used for performing scene roaming on a preset three-

dimensional scene to be reconstructed, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on image data to be reconstructed and a reconstruction result, the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed, and the image data to be reconstructed is used for recording an image of the preset three-dimensional scene.

According to the embodiments of the present disclosure, a scene reconstruction solution is provided. Image data to be reconstructed is obtained, where the image data to be reconstructed is used for recording the image of the preset three-dimensional scene to be reconstructed; sparse reconstruction is further performed on the image data to be reconstructed to obtain the reconstruction result; the modeling result is obtained by performing implicit modeling based on the image data to be reconstructed and the reconstruction result; and then the target video is generated by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene. Therefore, in the embodiments of the present disclosure, an objective of obtaining a scene roaming video through rendering by performing scene modeling based on image data to be reconstructed in a manner of sparse reconstruction and implicit modeling is achieved, so that technical effect of improving accuracy of a scene modeling result and a sense of reality of the scene roaming video is achieved, and technical problems of many defects in a modeling result and a poor sense of reality of the scene roaming video in the related art due to scene reconstruction performed through explicit modeling are further solved.

A person having ordinary skill in the art may understand that the structure shown in FIG. 12 is an example, and the computer terminal may alternatively be a terminal device such as a smartphone (for example, an Android-based mobile phone or an iOS-based mobile phone), a tablet computer, a palmtop computer, or a mobile Internet device (MID). FIG. 12 does not constitute a limitation on a structure of the computer terminal. For example, the computer terminal 120 may alternatively include more or fewer components (for example, a network interface or a display apparatus) than those shown in FIG. 12, or have a configuration different from that shown in FIG. 12.

A person having ordinary skill in the art may understand that all or some of the steps in the methods in the embodiments are performed by a program instructing related hardware of a

terminal device. The program is stored in a computer-readable storage medium. The storage medium may include a flash memory, a ROM, a RAM, a magnetic disk, a compact disc, or the like.

Embodiment Six

According to embodiments of the present disclosure, a computer-readable storage medium is further provided. Optionally, in the embodiments, the storage medium is configured to store program code executed in the method for generating a video based on scene reconstruction provided in Embodiment One, Embodiment Two, or Embodiment Three.

Optionally, in the embodiments, the storage medium is located in any computer terminal in a computer terminal group in a computer network, or located in any mobile terminal in a mobile terminal group.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: obtaining image data to be reconstructed, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; performing sparse reconstruction on the image data to be reconstructed to obtain a reconstruction result; performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain a modeling result; and generating a target video by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: constructing, on multiple scale features based on the image data to be reconstructed and the reconstruction result, grids corresponding to the preset three-dimensional scene to obtain a construction result, where the construction result is used for determining three-dimensional space information and light field information of the preset three-dimensional scene on the multiple scale features; and performing implicit modeling by using the construction result to obtain the modeling result.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: determining a bounding box corresponding to the preset three-dimensional scene based on the image data to be reconstructed and the

reconstruction result; performing multi-scale rasterizing division on three-dimensional space in the bounding box to obtain a division result; and obtaining the grids corresponding to the preset three-dimensional scene by using light field information corresponding to the division result, to obtain the construction result.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: dividing the three-dimensional space information and the light field information in the construction result into multiple partitions based on the multiple scale features to obtain a division result, where different partitions in the division result are used for respectively storing three-dimensional space information and light field information of different scale features; sampling, according to a preset sampling policy, the construction result based on a capture angle of an aerial image capture device and a target rendering position to obtain a sampling result, where the preset sampling policy is used for determining a sampling count and a sampling step that are to be used for the target rendering position during sampling and a ray count emitted from the target rendering position; obtaining a first multi-scale feature from the division result by using the sampling result; and performing implicit modeling by using the first multi-scale feature to obtain the modeling result.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing pixel value prediction on the first multi-scale feature by using a target fully-connected neural network model to obtain a target pixel value, where the target fully-connected neural network model is obtained through deep learning training by using multiple groups of data, and each of the multiple groups of data include a training image and a second multi-scale feature corresponding to the training image; and generating the modeling result based on the target pixel value.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing pixel value prediction on the second multi-scale feature by using an initial fully-connected neural network model to obtain a training pixel value; determining a first loss based on the training pixel value and a real pixel value corresponding to the training image; and updating model parameters of the initial fully-connected neural network model based on the first loss to obtain the target fully-connected neural network model.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing depth value prediction on the second multi-scale feature by using the initial fully-connected neural network model to obtain a training depth value; determining a second loss based on the training depth value and a real depth value corresponding to the training image; determining a sampling ray by using the capture angle and the target rendering position; determining a third loss based on a sampling weight, on the sampling ray, of a sampling point in the sampling result; and updating model parameters of the target fully-connected neural network model based on at least one of the second loss and the third loss to obtain an updated target fully-connected neural network model.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing aerial data capturing on the preset three-dimensional scene by using an aerial image capture device under a preset condition to obtain the image data to be reconstructed, where the preset condition is used for determining a weather condition and a flight condition of the aerial image capture device during aerial data capturing.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing feature extraction on the image data to be reconstructed to obtain an extraction result; performing feature matching on the extraction result to obtain a matching result; and performing sparse reconstruction on the matching result to obtain the reconstruction result, where the reconstruction result is used for determining pose information, corresponding to the image data to be reconstructed, of the aerial image capture device.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: preprocessing the image data to be reconstructed to obtain a preprocessing result; and performing feature extraction on the preprocessing result to obtain the extraction result.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing image pairing on the extraction

result to obtain a pairing result; and performing feature matching on the pairing result to obtain the matching result.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing geometrical filtering on the matching result to obtain a filtering result; performing global three-dimensional structural estimation on the filtering result to obtain an estimation result; and determining the reconstruction result based on the estimation result.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: performing horizon correction on the estimation result to obtain a first correction result; performing heading angle correction on the first correction result to obtain a second correction result; performing point cloud partitioning processing on the second correction result to obtain a partitioning result; and performing data format adjustment on the partitioning result to obtain the reconstruction result.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: obtaining trajectory information corresponding to the aerial image capture device; and performing video rendering on the modeling result based on the trajectory information to obtain the target video.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: displaying image data to be reconstructed on an operation interface in response to an input instruction performed on the operation interface, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; and displaying a target video on the operation interface in response to a reconstruction instruction performed on the operation interface, where the target video is used for performing scene roaming on the preset three-dimensional scene, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on the image data to be reconstructed and a reconstruction result, and the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed.

Optionally, in the embodiments, the computer-readable storage medium is configured to store program code for performing the following steps: displaying an initial default picture on a

displayed picture of a VR device or an AR device; and driving the VR device or the AR device to jump from displaying the initial default picture to displaying a target video, where the target video is used for performing scene roaming on a preset three-dimensional scene to be reconstructed, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on image data to be reconstructed and a reconstruction result, the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed, and the image data to be reconstructed is used for recording an image of the preset three-dimensional scene.

The sequence numbers of the embodiments of the present disclosure are intended for description, but do not indicate a preference among the embodiments.

In the embodiments of the present disclosure, the descriptions in the embodiments have respective focuses. For a part that is not described in detail in an embodiment, refer to related descriptions in other embodiments.

In the several embodiments provided in the present disclosure, it should be understood that the disclosed technical content is implemented in other manners. The described apparatus embodiments are exemplary. For example, division into the units is logical function division, and other division manners are used during actual implementation. For example, multiple units or components are combined, or are integrated into another system, or some features is omitted or not performed. In addition, the coupling, or direct coupling, or communication connection between the displayed or discussed components is the indirect coupling or communication connection by means of some interfaces, units, or modules, and is electrical or of other forms.

The units described as separate components may or may not be physically separated, and the components displayed as units may or may not be physical units, and are located in one place or are distributed over multiple network units. Some or all of the units are selected according to actual needs to achieve the objectives of the solutions of the embodiments.

In addition, functional units in the embodiments of the present disclosure are integrated into one processing unit, or each of the units may exist alone physically, or two or more units are integrated into one unit. The integrated unit is implemented in a form of hardware, or is implemented in a form of a software functional unit.

When the integrated unit is implemented in the form of a software functional unit and sold or used as an independent product, the integrated unit is stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the present disclosure essentially, or the part contributing to the related art, or all or some of the technical solutions are implemented in a form of a software product. The computer software product is stored in a storage medium, and includes several instructions for instructing a computer device (which is a personal computer, a server, a network device, or the like) to perform all or some of the steps of the methods described in the embodiments of the present disclosure. The storage medium includes any medium that can store program code, such as a USB flash drive, a ROM, a RAM, a removable hard disk, a magnetic disk, or a compact disc.

The descriptions are exemplary implementations of the present disclosure. It should be noted that a person having ordinary skill in the art may make several improvements or polishing without departing from the principle of the present disclosure, and the improvements and polishing instructed shall also fall within the protection scope of the present disclosure.

**Industrial Applicability**

According to the method and the system for generating a video based on scene reconstruction and the storage medium that are provided in the embodiments of the present disclosure, image data to be reconstructed is obtained, where the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; sparse reconstruction is further performed on the image data to be reconstructed to obtain a reconstruction result; a modeling result is obtained by performing implicit modeling based on the image data to be reconstructed and the reconstruction result; and then a target video is generated by using the modeling result, where the target video is used for performing scene roaming on the preset three-dimensional scene. Therefore, in the embodiments of the present disclosure, an objective of obtaining a scene roaming video through rendering by performing scene modeling based on image data to be reconstructed in a manner of sparse reconstruction and implicit modeling is achieved, so that technical effect of improving accuracy of a scene modeling result and a sense of reality of the scene roaming video is achieved, and technical problems of many defects in a modeling result and a poor sense of reality of the scene roaming

video in the related art due to scene reconstruction performed through explicit modeling are further solved.

## Claims

1. A method for generating a video based on scene reconstruction, comprising:

   obtaining image data to be reconstructed, wherein the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed;

   performing sparse reconstruction on the image data to be reconstructed to obtain a reconstruction result;

   performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain a modeling result; and

   generating a target video by using the modeling result, wherein the target video is used for performing scene roaming on the preset three-dimensional scene.

2. The method according to claim 1, wherein performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain the modeling result comprises:

   constructing, on a plurality of scale features based on the image data to be reconstructed and the reconstruction result, grids corresponding to the preset three-dimensional scene to obtain a construction result, wherein the construction result is used for determining three-dimensional space information and light field information of the preset three-dimensional scene on the plurality of scale features; and

   performing implicit modeling by using the construction result to obtain the modeling result.

3. The method according to claim 2, wherein constructing, on the plurality of scale features based on the image data to be reconstructed and the reconstruction result, the grids corresponding to the preset three-dimensional scene to obtain the construction result comprises:

   determining a bounding box corresponding to the preset three-dimensional scene based on the image data to be reconstructed and the reconstruction result;

   performing multi-scale rasterizing division on three-dimensional space in the bounding box to obtain a division result; and

   obtaining the grids corresponding to the preset three-dimensional scene by using light field information corresponding to the division result, to obtain the construction result.

4. The method according to claim 2, wherein performing implicit modeling by using the construction result to obtain the modeling result comprises:

dividing the three-dimensional space information and the light field information in the construction result into a plurality of partitions based on the plurality of scale features to obtain a division result, wherein different partitions in the division result are used for respectively storing three-dimensional space information and light field information of different scale features;

sampling, according to a preset sampling policy, the construction result based on a capture angle of an aerial image capture device and a target rendering position to obtain a sampling result, wherein the preset sampling policy is used for determining a sampling count and a sampling step that are to be used for the target rendering position during sampling and a ray count emitted from the target rendering position;

obtaining a first multi-scale feature from the division result by using the sampling result; and

performing implicit modeling by using the first multi-scale feature to obtain the modeling result.

5. The method according to claim 4, wherein performing implicit modeling by using the first multi-scale feature to obtain the modeling result comprises:

performing pixel value prediction on the first multi-scale feature by using a target fully-connected neural network model to obtain a target pixel value, wherein the target fully-connected neural network model is obtained through deep learning training by using a plurality of groups of data, and each of the plurality of groups of data comprise a training image and a second multi-scale feature corresponding to the training image; and

generating the modeling result based on the target pixel value.

6. The method according to claim 5, wherein the method further comprises:

performing pixel value prediction on the second multi-scale feature by using an initial fully-connected neural network model to obtain a training pixel value;

determining a first loss based on the training pixel value and a real pixel value corresponding to the training image; and

updating model parameters of the initial fully-connected neural network model based on the first loss to obtain the target fully-connected neural network model.

7. The method according to claim 6, wherein the method further comprises:

performing depth value prediction on the second multi-scale feature by using the initial fully-connected neural network model to obtain a training depth value;

determining a second loss based on the training depth value and a real depth value corresponding to the training image;

determining a sampling ray by using the capture angle and the target rendering position;

determining a third loss based on a sampling weight, on the sampling ray, of a sampling point in the sampling result; and

updating model parameters of the target fully-connected neural network model based on at least one of the second loss and the third loss to obtain an updated target fully-connected neural network model.

8. The method according to claim 1, wherein obtaining the image data to be reconstructed comprises:

performing aerial data capturing on the preset three-dimensional scene by using an aerial image capture device under a preset condition to obtain the image data to be reconstructed, wherein the preset condition is used for determining a weather condition and a flight condition of the aerial image capture device during aerial data capturing.

9. The method according to claim 8, wherein performing sparse reconstruction on the image data to be reconstructed to obtain the reconstruction result comprises:

performing feature extraction on the image data to be reconstructed to obtain an extraction result;

performing feature matching on the extraction result to obtain a matching result; and

performing sparse reconstruction on the matching result to obtain the reconstruction result, wherein the reconstruction result is used for determining pose information, corresponding to the image data to be reconstructed, of the aerial image capture device.

10. The method according to claim 8, wherein generating the target video by using the modeling result comprises:

obtaining trajectory information corresponding to the aerial image capture device; and

performing video rendering on the modeling result based on the trajectory information to obtain the target video.

11. The method according to claim 1, wherein the sparse reconstruction is a three-dimensional scene reconstruction operation based on a SfM algorithm, the reconstruction result

comprises a camera pose corresponding to the image data to be reconstructed, the implicit modeling is a three-dimensional scene modeling optimization operation based on a NeRF algorithm, the modeling result comprises a scene model corresponding to the preset three-dimensional scene to be reconstructed, and the target video is a scene roaming video, at a free viewpoint, corresponding to the preset three-dimensional scene to be reconstructed.

12. A method for generating a video based on scene reconstruction, comprising:

displaying image data to be reconstructed on an operation interface in response to an input instruction performed on the operation interface, wherein the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed; and

displaying a target video on the operation interface in response to a reconstruction instruction performed on the operation interface, wherein the target video is used for performing scene roaming on the preset three-dimensional scene, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on the image data to be reconstructed and a reconstruction result, and the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed.

13. A method for generating a video based on scene reconstruction, comprising:

displaying an initial default picture on a displayed picture of a VR device or an AR device; and

driving the VR device or the AR device to jump from displaying the initial default picture to displaying a target video, wherein the target video is used for performing scene roaming on a preset three-dimensional scene to be reconstructed, the target video is generated by using a modeling result, the modeling result is obtained by performing implicit modeling on image data to be reconstructed and a reconstruction result, the reconstruction result is obtained by performing sparse reconstruction on the image data to be reconstructed, and the image data to be reconstructed is used for recording an image of the preset three-dimensional scene.

14. A computer-readable storage medium, wherein the computer-readable storage medium stores an executable program, and when the executable program is run, a device on which

the computer-readable storage medium is located is controlled to perform the method for generating the video based on scene reconstruction as claimed in any one of claims 1 to 13.

15. A system for generating a video based on scene reconstruction, comprising:

a processor; and

a memory, connected with the processor and configured to provide the processor with instructions for processing the following processing steps:

obtaining image data to be reconstructed, wherein the image data to be reconstructed is used for recording an image of a preset three-dimensional scene to be reconstructed;

performing sparse reconstruction on the image data to be reconstructed to obtain a reconstruction result;

performing implicit modeling based on the image data to be reconstructed and the reconstruction result to obtain a modeling result; and

generating a target video by using the modeling result, wherein the target video is used for performing scene roaming on the preset three-dimensional scene.