

PANORAMA GENERATION METHOD, NETWORK TRAINING METHOD, APPARATUSES, DEVICE, AND MEDIUM

TECHNICAL FIELD

- [0001] Embodiments of the application relate to the field of computer vision, and relate, but not limited, to a panorama generation method, a network training method, apparatuses, a device, and a medium.

BACKGROUND

- [0002] In graphics, high-quality panoramas are one of the most popular methods for providing 360-degree realistic environment textures and light effects for 3-dimension (3D) scenes. In the related art, the panoramas are usually obtained by using an expensive panorama camera and then performing post-calibration and retouching by professionals, which leads to a higher threshold for obtaining the panoramas.

SUMMARY

- [0003] The embodiments of the application provide a technical solution for panorama generation.
- [0004] An embodiment of the application provides a panorama generation method, which may include the following operations. A scene description text is acquired. Based on the scene description text, feature sampling is performed in a global feature book corresponding to a preset scene image set to obtain a global image feature. Based on the global image feature and spherical positional encoding corresponding to a preset panorama, feature sampling is performed in a local feature book corresponding to the preset scene image set to obtain a target panorama corresponding to the scene description text.
- [0005] An embodiment of the application provides a method for training a panorama generation network, which may include the following operations. A sample text feature is acquired, the sample text feature including a pseudo text feature and a text-like feature, the pseudo text feature being obtained by perturbing a text feature corresponding to a sample image, and the text-like feature being a text feature corresponding to an image whose similarity with the sample image satisfies a preset condition. According to the sample text feature and a global feature book, a global sampler to be trained in a panorama generation network to be trained is trained to obtain a trained global sampler, the global feature book being obtained by performing feature coding on a preset scene image set based on a trained global encoder. A global sample feature matching the sample text feature on scene semantics is determined in the global feature book using the trained global sampler. According to the global sample feature, spherical positional encoding corresponding to a preset panorama and a local feature book, a local sampler to be trained in the panorama generation network to be trained is trained to obtain a trained local sampler, the local feature book being obtained by performing feature coding on the

preset scene image set based on a trained local encoder.

- [0006] An embodiment of the application provides a panorama generation apparatus, which may include: a first acquisition module, a first sampling module, and a second sampling module. The first acquisition module is configured to acquire a scene description text. The first sampling module is configured to perform, based on the scene description text, feature sampling in a global feature book corresponding to a preset scene image set to obtain a global image feature. The second sampling module is configured to perform, based on the global image feature and spherical positional encoding corresponding to a preset panorama, feature sampling in a local feature book corresponding to the preset scene image set to obtain a target panorama corresponding to the scene description text.
- [0007] An embodiment of the application provides an apparatus for training a panorama generation network, which may include: a second acquisition module, a first training module, a first determination module, and a second training module. The second acquisition module is configured to acquire a sample text feature, the sample text feature including a pseudo text feature and a text-like feature, the pseudo text feature being obtained by perturbing a text feature corresponding to a sample image, and the text-like feature being a text feature corresponding to an image whose similarity with the sample image satisfies a preset condition. The first training module is configured to train, according to the sample text feature and a global feature book, a global sampler to be trained in a panorama generation network to be trained to obtain a trained global sampler, the global feature book being obtained by performing feature coding on a preset scene image set based on a trained global encoder. The first determination module is configured to determine, in the global feature book, a global sample feature matching the sample text feature on scene semantics using the trained global sampler. The second training module is configured to train a local sampler to be trained in the panorama generation network to be trained according to the global sample feature, the spherical positional encoding corresponding to a preset panorama and a local feature book to obtain a trained local sampler, the local feature book being obtained by performing feature coding on the preset scene image set based on a trained local encoder.
- [0008] An embodiment of the application provides a computer device, which may include a memory and a processor. A computer-executable instruction is stored in the memory. The processor can execute the computer-executable instruction in the memory to implement the panorama generation method or the method for training a panorama generation network.
- [0009] An embodiment of the application provides a computer storage medium, in which a computer-executable instruction is stored. The computer-executable instruction can be executed to implement the panorama generation method or the method for training a panorama generation network.
- [0010] The embodiments of the application provide a panorama generation method, a network training method, apparatuses, a device, and a medium. The panorama generation method may include the following operations. Firstly, a scene description text is acquired; in this way, a basis is provided for the subsequent generation of a corresponding panorama based on the scene description text. Secondly, based on the scene description text, feature sampling is performed in a global feature book corresponding to a preset scene image set to obtain a global image feature; here, a global

image feature matching the input scene description text can be determined in the global feature book, which is generated in advance, corresponding to the preset scene image set, the global feature book being used for representing rough global feature information of an image; in this way, global scene semantics of an image can be provided for the subsequent construction of the panorama. Finally, based on the global image feature and spherical positional encoding corresponding to a preset panorama, feature sampling is performed in a local feature book corresponding to the preset scene image set to obtain a target panorama corresponding to the scene description text; in this way, the local image feature that matches the input scene description text and corresponds to each spherical positional encoding can be determined in the local feature book, which is generated in advance, corresponding to the preset scene image set, the local feature book being used for representing local feature information of the fine granularity of an image, and the spherical positional coding corresponding to the preset panorama; in this way, based on that the global feature information of an image is obtained, the panorama that matches the scene description text can be further obtained based on the local feature information of the image corresponding to each spherical positional encoding. Thus, the rough global feature of the subsequent constructed image can be determined according to the input scene description text; according to the rough global feature obtained by sampling, the local image feature that represents the local fine granularity of an image is sampled at the corresponding position of the panorama based on the spherical positional coding, and then by performing feature decoding and feature synthesis on the acquired local image feature, the panorama that has a spherical structure and matches text semantics of the scene description text can be obtained. In this way, based on input text information, it is relatively more convenient to obtain the panorama that matches the text semantics of the text information.

[0011] It should be understood that the above general descriptions and detailed descriptions below are only exemplary and explanatory and not intended to limit the technical solutions of the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] In order to more clearly illustrate technical solutions in embodiments of the application, the accompanying drawings needed in the description of the embodiments are simply introduced below. It is apparent that the accompanying drawings in the following description are only embodiments of the application, for the ordinary skill in the art, some other accompanying drawings can also be obtained according to these on the premise of not contributing creative effort.

[0013] FIG. 1 is a flowchart of a first panorama generation method provided by an embodiment of the application.

[0014] FIG. 2 is a flowchart of a second panorama generation method provided by an embodiment of the application.

[0015] FIG. 3 is a flowchart of a method for training a panorama generation network provided by an embodiment of the application.

[0016] FIG. 4 is a framework diagram corresponding to a method for generating a panorama based on text drive that includes two implementation stages provided by an

embodiment of the application.

- [0017] FIG. 5 is a panorama generated based on a panorama generation method provided by an embodiment of the application.
- [0018] FIG. 6 is a schematic diagram of an implementation framework for generating a panorama with Low Dynamic Range (LDR) and Low Resolution (LR) based on an input text provided by an embodiment of the application.
- [0019] FIG. 7A is a schematic diagram of an implementation framework for adjusting the resolution and dynamic range for a generated panorama provided by an embodiment of the application.
- [0020] FIG. 7B is a schematic diagram of a framework for generating a training sample of an image resolution and dynamic range adjustment part in a panorama generation network provided by an embodiment of the application.
- [0021] FIG. 8A is a structural schematic diagram of a panorama generation apparatus provided by an embodiment of the application.
- [0022] FIG. 8B is a structural schematic diagram of an apparatus for training a panorama generation network provided by an embodiment of the application.
- [0023] FIG. 9 is a structure diagram of a computer device provided by an embodiment of the application.

DETAILED DESCRIPTION

- [0024] In order to make the purpose, technical solutions and advantages of the embodiments of the application clearer, the specific technical solutions of the disclosure will be described below in more detail in combination with the accompanying drawings in the embodiments of the application. The following embodiments are used for illustrating the embodiments of the application, but not for limiting the scope of the embodiments of the application.
- [0025] "Some embodiments" involved in the following descriptions describes a subset of all possible embodiments. However, it can be understood that "some embodiments" may be the same subset or different subsets of all the possible embodiments, and may be combined without conflicts.
- [0026] Term "first/second/third" involved in the following descriptions is only for distinguishing similar objects and does not represent a specific sequence of the objects. It can be understood that "first/second/third" may be interchanged to specific sequences or orders if allowed to implement the embodiments of the application described herein in sequences except the illustrated or described ones.
- [0027] Unless otherwise defined, all technical and scientific terms in the specification have the same meaning as those skilled in the art, belonging to the embodiments of the application, usually understand. Terms used in the specification are only used for describing the purpose of the embodiments of the application, but not intended to limit the embodiments of the application.
- [0028] Nouns and terms involved in the embodiments of the application will be described before the embodiments of the application are further described in detail. The

nouns and terms involved in the embodiments of the application are suitable to be explained as follows.

- [0029] 1) Panorama: surroundings are shown as much as possible through a wide angle expression and the forms like painting, photos, videos, and three-dimensional models. As for a 360-degree panorama, pictures, which are obtained by capturing image information of a whole scene by means of a professional camera or rendering through modeling software, are spliced by means of software and played with a specific player, that is, plane photos or computer modeling pictures are transformed into a 360-degree view for virtual reality browsing, and 2-Dimension (2D) plane pictures are simulated into a real 3D space for a presentation to the viewers.
- [0030] 2) Zero-Shot Learning (ZSL): past knowledge is used for reasoning out a specific form of a new object, so as to recognize the new object. It is hoped that ZSL imitates this reasoning process of humans, so that computers have the capability of recognizing new things.
- [0031] 3) Dynamic range of image: it represents the range from the "darkest" to the "brightest" contained in the image. The larger the dynamic range is, the more gradations can be expressed, and the wider the color space contained is.
- [0032] 4) Super-Resolution (SR): the process of improving the resolution of the original image by hardware or software, and creating a high-resolution image from a series of low-resolution images is called SR reconstruction, for example, LR to HR.
- [0033] 5) Inverse Tone Mapping (ITM): it is a technology for converting a source signal with a Standard Dynamic Range (SDR) into a source signal with a High Dynamic Range (HDR).
- [0034] 6) Image interpolation: the gray values of the known neighboring pixels are used to generate the gray values of the unknown pixels, so as to regenerate an image with a higher resolution from the original image.
- [0035] Exemplary applications of a panorama generation apparatus provided by the embodiments of the application are described below. The device provided by the embodiments of the application may be implemented as a laptop with an image acquisition function, a tablet computer, a desktop computers, a camera, mobile devices (for example, a personal digital assistant, a dedicated message device, and a portable game device) and other types of user terminals, and may also be implemented as a server. The exemplary application of the device implemented as the terminal or the server will be described below.
- [0036] The method may be applied to a computer device. The function realized by the method may be realized by a processor in the computer device by calling a program code. Of course, the program code may be stored in a computer storage medium. It can be seen that the computer device at least includes the processor and the storage medium.
- [0037] An embodiment of the application provides a panorama generation method. As shown in FIG. 1 which is a flowchart of a first panorama generation method provided by an embodiment of the application, descriptions will be made in combination with the operations shown in FIG. 1.
- [0038] At S101, a scene description text is acquired.

- [0039] In some embodiments, a panorama generation apparatus may receive text information input from the outside based on its input and output module, and perform semantic parsing to the text information, so as to obtain the scene description text, or directly receive the scene description text input from the outside based on the input and output module; or, the panorama generation apparatus receives the scene description text sent by the apparatus with which information is exchanged.
- [0040] In some embodiments, the scene description text may be a sentence or a paragraph; at the same time, the scene description text may be expressed in any form, for example, "the full moon is like a yellow lantern, rising from the eastern horizon in the sky", "a vast prairie, a vast expanse, a gentle breeze with the sweet smell of the earth ruffled a wide field of crops", "brown wooden dock on lake surrounded by green trees during daytime", and "ocean waves crashing on shore under blue and white cloudy sky during daytime".
- [0041] In some embodiments, the scene description text may describe related scene information; for example, "blue sky and white clouds", "wooden boat by the lake", "grassland after rain" and so on can be known from the scene description text. In the embodiments of the application, the scene description text is text information with preset scene semantics.
- [0042] In some embodiments, the number of scene description texts may be one, or may be two or more.
- [0043] At S102, based on the scene description text, feature sampling is performed in a global feature book corresponding to a preset scene image set to obtain a global image feature.
- [0044] In some embodiments, the preset scene image set may be any image set. The preset scene image set may include at least two preset images; at the same time, a frame of image of any preset scene image included in the preset scene image set may include various scene information, which may be determined according to actual requirements.
- [0045] In some embodiments, the preset scene image set may include: a preset image 1 which includes a puppy running in a park, a preset image 2 which includes a wooden boat on a lake, a preset image 3 which includes trees at sunset, etc.
- [0046] In some embodiments, the global feature book corresponding to the preset scene image set may refer to feature coding performed on each preset scene image in the preset scene image set using a trained encoder, so as to obtain a feature corresponding to each preset scene image in the preset scene image set. The feature here specifically refers to the global feature representing rough semantics of the preset scene image. Each global feature in the global feature book corresponds to a preset scene image in the preset scene image set.
- [0047] Here, any preset scene image in the preset scene image set may be a 2D image or a 3D image, or may be a gray scale image or a color image.
- [0048] In some embodiments, for each preset scene image in the preset scene image set, the global image feature (global codebook) representing the rough semantics of the image and the local image feature (local codebook) representing fine granularity information of the image can be obtained by means of different encoders.

- [0049] In some embodiments, it is possible to search in the global feature book for multiple pieces of feature information that match the text semantics carried by the scene description text, and then the multiple pieces of feature information are combined to obtain the global image feature. Here, the scene description text may also be transformed into an embedded feature vector, and then the matched global image feature is sampled in the global feature book based on the embedded feature vector.
- [0050] In some embodiments, the global image feature may be a feature sequence, which is used for representing the global information of an image corresponding to the scene description text. At the same time, feature decoding may be performed on the global image feature, that is, image reconstruction may be realized; and then the image with rough semantics corresponding to the scene description text is obtained.
- [0051] At S103, based on the global image feature and spherical positional encoding corresponding to a preset panorama, feature sampling is performed in a local feature book corresponding to the preset scene image set to obtain a target panorama corresponding to the scene description text.
- [0052] In some embodiments, the preset panorama may refer to a panorama that does not include any frame information, or to a panorama that includes any frame information; for example, the pixel value of each pixel in the preset panorama is 255 or 0; at the same time, the spherical positional encoding corresponding to the preset panorama may refer to position information of each pixel in the preset panorama.
- [0053] In some embodiments, the spherical positional encoding corresponding to the preset panorama, namely any pixel P, may be represented by $S(\theta, \phi, r)$, where r is the distance between the origin O and the point P of the preset panorama, θ is the included angle between the directed line segment OP and the forward direction of the z-axis, ϕ is the angle at which the x-axis goes counterclockwise to OM looking from the forward direction of the z-axis, M is the projection of the point P on the plane xOy. At the same time, r has a variation range of $[0, +\infty)$, θ has a variation range of $[0, \pi]$, ϕ has a variation range of $[0, 2\pi]$. In the embodiments of the application, r may be 1 by default.
- [0054] In some embodiments, the local feature book corresponding to the preset scene image set may refer to feature coding performed on each preset scene image in the preset scene image set using a trained encoder, so as to obtain a feature corresponding to each preset scene image in the preset scene image set. The feature here specifically refers to the local feature representing the fine granularity of the preset scene image. Each local feature in the local feature book corresponds to any image block in a preset scene image in the preset scene image set.
- [0055] In some embodiments, the global image feature may be taken as a selection condition, and the local image feature corresponding to each spherical positional encoding may be determined in the local feature book corresponding to the preset scene image set; then, feature decoding is performed on the local image feature corresponding to each spherical positional encoding, and the image synthesis is performed according to the position information corresponding to the spherical positional encoding to obtain the corresponding panorama, namely the target panorama corresponding to the scene description text.
- [0056] In some embodiments, firstly, a piece of scene description information is acquired, which may be presented in the form of scene description text, for example,

"ocean waves crashing on shore under blue and white cloudy sky during daytime"; in this way, a basis is provided for the subsequent generation of the corresponding panorama based on the scene description information; secondly, based on the scene description text, feature sampling is performed in the global feature book corresponding to the preset scene image set to obtain the global image feature, that is, the global image feature that match the text semantics of "ocean waves crashing on shore under blue and white cloudy sky during daytime" is determined based on the global feature book corresponding to the preset scene image set, the global feature book being used for representing the rough global feature information of an image; in this way, global scene semantics can be provided for the subsequent construction of the panorama; finally, based on the global image feature and the spherical positional encoding corresponding to the preset panorama, feature sampling is performed in the local feature book corresponding to the preset scene image set to obtain the target panorama corresponding to the scene description text; in this way, the global image feature corresponding to the "ocean waves crashing on shore under blue and white cloudy sky during daytime" at each spherical positional encoding can be sampled based on the local feature book corresponding to the preset scene image set, the local feature book being used for representing the local feature information of the fine granularity of an image, and the spherical positional encoding corresponding to the preset panorama; in this way, based on that the global feature information of an image is acquired, the panorama that matches the scene description text can be obtained based on the local feature information of the image corresponding to each spherical positional encoding. Thus, the rough global feature of the subsequent constructed images can be determined according to the input scene description text; and according to the rough global feature obtained by sampling, the local image feature of the fine granularity of an image is sampled at the corresponding position of the panorama based on the spherical positional encoding, and then by performing feature decoding and feature synthesis to the acquired local image feature, the panorama that has a spherical structure and the scene semantics covered by the text semantics of "ocean waves crashing on shore under blue and white cloudy sky during daytime" can be obtained. In this way, with only one piece of text information, the panorama aligned with its text semantics information, namely the panorama matching the text semantics information, can be obtained without the need of relying on professionals and expensive panorama acquisition devices. Thus, the labor cost of panorama generation can be reduced, and the implementation scheme of panorama generation can be simplified.

[0057] In the panorama generation method provided by the embodiment of the application, firstly, a scene description text is acquired; in this way, a basis is provided for the subsequent generation of the corresponding panorama based on the scene description text; secondly, based on the scene description text, feature sampling is performed in the global feature book corresponding to the preset scene image set to obtain the global image feature; here, the global image feature that matches the input scene description text can be determined in the global feature book, which is generated in advance, corresponding to the preset scene image set, the global feature book being used for representing the rough global feature information of an image; in this way, the global scene semantics of an image can be provided for the subsequent construction of the panorama; finally, based on the global image feature and the spherical positional encoding corresponding to the preset panorama, feature sampling is performed in the

local feature book corresponding to the preset scene image set to obtain the target panorama corresponding to the scene description text; in this way, the local image feature that matches the input scene description text and corresponds to each spherical positional encoding can be determined in the local feature book, which is generated in advance, corresponding to the preset scene image set, the local feature book being used for representing the local feature information of the fine granularity of an image, and the spherical positional encoding corresponding to the preset panorama; in this way, based on that the global feature information of an image is acquired, the panorama that matches the scene description text can be further obtained based on the local feature information of an image corresponding to each spherical positional encoding. Thus, the rough global feature of the subsequent constructed image can be determined according to the input scene description text; and according to the rough global feature obtained by sampling and the spherical positional encoding, the local image feature representing the local fine granularity of an image is sampled at the corresponding position of the panorama, and then by performing feature decoding and feature synthesis on the acquired local image feature, the panorama that has a spherical structure and matches the text semantics of the scene description text can be obtained. In this way, based on the input text information, it is relatively more convenient to obtain the panorama that matches the text semantics of the text information.

[0058] In some embodiments, firstly, feature coding may be performed on the scene description text, that is, a preset text encoder may be used to perform feature coding on the scene description text, so as to obtain a corresponding text feature; in this way, it is convenient to search for the feature matching or similar to the text feature in the global feature book later; secondly, it is possible to search for, based on the determined text feature, the feature matching the text feature on the scene semantics in the global feature book and to sample the feature, and then the global image feature is obtained; in this way, the global image feature matching the text feature on the scene semantics can be determined in the global codebook corresponding to the preset scene image set, that is, the subsequent provision of the global scene semantics of an image for the determined target panorama is realized. S102 provided in the above embodiment may also be implemented by performing S201 and S202 below. FIG. 2 is a flowchart of a second panorama generation method provided by an embodiment of the application. Descriptions will be made in combination with the operations shown in FIG. 1 and FIG. 2.

[0059] At S201, a text feature corresponding to the scene description text is determined.

[0060] For the determination of the text feature corresponding to the scene description text, feature coding may be performed to the scene description text in any way, and then the corresponding text feature can be obtained. For example, the text feature may be represented by a discrete feature sequence. Generally, a text encoder in a Contrastive Language Image Pre-training (CLIP) model may be used to perform feature coding on the scene description text, and then the text feature is obtained.

[0061] At S202, a feature that matches the text feature on scene semantics is sampled in the global feature book to obtain the global image feature.

[0062] In some embodiments, it is possible to search for the feature that matches the text feature on the scene semantics in the global feature book, and then the feature may

be sampled to obtain the global image feature. Here, the feature that matches the text feature on the scene semantics may be image scene semantics carried by the relevant global feature in the global feature book corresponding to the preset scene image set, which is similar to the text semantics carried by the text feature. The global image feature may be represented by an off-line feature. Generally, a global sampler in the trained panorama generation network may be used to sample, in the global feature book, the global image feature that matches the text feature, namely the corresponding feature for representing the global rough semantics of an image to be constructed later.

[0063] In some possible implementation modes, firstly, the local image feature corresponding to each positional code in Spherical Positional Encoding (SPE) may be determined in the local feature book based on the global image feature acquired in the previous operation; in this way, the local image feature that corresponds to each spherical positional encoding in the target panorama to be constructed later and represents the fine granularity of an image can be obtained; thus, based on the consideration of spherical structure information of a 360-degree scene, the feature information representing the fine granularity on each part (positional code) of the image can be obtained; secondly, feature decoding and feature synthesis are performed on the obtained local image feature corresponding to each positional code, so as to obtain the target panorama; in this way, feature decoding and feature fusion can be performed on the local feature corresponding to each spherical positional encoding of the panorama, and then the target panorama that has the spherical structure information and matches the scene semantics of the scene description text is obtained; here, the scene semantics includes a relevant rough global feature and a local feature of the fine granularity, thus, on the basis of not using a large number of text-image pairs for training, the text semantics of the input scene description text can be aligned with the scene semantics of the generated target panorama in the process of ZSL, that is, based on the input text information, the target panorama aligned with its text semantics is generated in the process of ZSL. S103 provided in the above embodiment may be implemented by S203 and S204 below.

[0064] At S203, based on the global image feature, a local image feature corresponding to each positional code in the SPE is determined in the local feature book.

[0065] In some embodiments, the global image feature may be taken as a selection condition, and the local image feature corresponding to each positional code in the SPE can be determined in the local feature book, that is, the local feature representing the fine granularity of an image is sampled according to the corresponding position of the spherical positional encoding in the panorama, and then the local image feature is obtained. The local image feature may be represented by an off-line feature. Generally, based on the global image feature obtained in the previous operation, a local sampler in the trained panorama generation network may be used to sample, in the local feature book, the local image feature corresponding to each positional code in the SPE, namely the feature corresponding to local semantics of an image.

[0066] At S204, feature decoding and feature synthesis are performed on the local image feature corresponding to each positional code to obtain the target panorama.

[0067] In some embodiments, firstly, feature decoding may be performed on the local image feature corresponding to each positional code, and the image information corresponding to each positional code is obtained; secondly, image synthesis is performed on the image information corresponding to each positional code to obtain the

corresponding target panorama.

- [0068] In some embodiments, the scene semantics represented by the target panorama matches the text semantics corresponding to the scene description text.
- [0069] In some possible implementation modes, the relevant information, for example, the resolution of the image and the dynamic range of the image, of the obtained target panorama may also be adjusted; for example, the resolution and dynamic range of the image are improved, so that the resolution of the adjusted target panorama is higher than that of the target panorama, and the dynamic range corresponding to the adjusted target panorama is higher than that of the target panorama, and then the panorama with high quality and HDR is obtained; here, firstly, the continuous feature corresponding to each pixel in the target panorama may be determined; in this way, for the continuous feature corresponding to each pixel, up-sampling is performed in terms of the resolution and the dynamic range, that is, the resolution and the dynamic range are adjusted for the continuous feature corresponding to each pixel to obtain the adjusted target panorama with a higher resolution and a higher dynamic range compared with the target panorama; that is, after S103, the panorama generation method provided by the embodiment of the application may also perform S205 and S206 below.
- [0070] At S205, the continuous feature corresponding to each pixel in the target panorama is determined.
- [0071] In some embodiment, the continuous feature corresponding to each pixel may be represented based on a potential feature corresponding to each pixel in the target panorama. The gray value of unknown pixels may also be generated based on the gray values of the known neighboring pixels in an image, so as to regenerate an image with a higher resolution from the original image.
- [0072] In some possible implementation modes, firstly, regions division may be performed on the target panorama, and then multiple image blocks corresponding to the target panorama are obtained, that is, an image block set is obtained; in this way, processing an image in blocks can improve the speed and accuracy of image processing; secondly, feature coding is performed on each pixel in each image block in the image block set, and at least one potential feature surrounding the pixel can be obtained; in this way, based on the potential features corresponding to the pixel, a structured potential feature on the sphere of the target panorama with the spherical structure information can be obtained; finally, the ratio between the area of each potential feature surrounding the pixel and a preset area is correspondingly fused with each potential feature surrounding the pixel to obtain the continuous feature corresponding to the pixel, the preset area being the sum of the areas of potential codes that include the pixel. Thus, the continuous feature corresponding to each positional code (spherical positional code) in the target panorama may be represented, that is, image interpolation and representation of the continuous image feature can be realized conveniently based on the area parameter corresponding to each potential feature, and then a basis is provided for obtaining the panorama with a higher resolution and higher dynamic range. S205 may be implemented by S2051 to S2053 below (not shown in the accompanying drawings).
- [0073] At S2051, region division is performed on the target panorama to obtain an image block set.
- [0074] In some embodiments, region division is performed on the target panorama to

obtain at least two image blocks, and then an image block set is obtained based on the at least two image blocks. The target panorama may also be segmented with a sliding window to obtain multiple image blocks. There may be similar parts among the multiple image blocks.

[0075] At S2052, for each pixel in each image block in the image block set, feature coding is performed on the pixel to obtain at least one potential feature surrounding the pixel.

[0076] In some embodiments, feature coding may be performed on each pixel in each image block in the image block set to obtain a preset feature corresponding to each pixel, and then at least one potential feature surrounding the pixel is obtained based on the position of the preset feature corresponding to each pixel. The encoder in the trained panorama generation network may be used to perform feature coding on each pixel in the obtained target panorama, and then at least one potential feature corresponding to the pixel is obtained.

[0077] At S2053, the continuous feature corresponding to the pixel is obtained based on each potential feature surrounding the pixel and the ratio between the area of each potential feature surrounding the pixel and the preset area.

[0078] In some embodiments, the preset area is the sum of the areas of potential codes surrounding the pixel. Here, firstly, the ratio between the area of each potential feature surrounding the pixel and the preset area may be determined, that is, the area ratio of each potential feature surrounding the pixel may be obtained, the preset area being the sum of the areas of the potential codes surrounding the pixel; secondly, the area ratio of each potential feature surrounding the pixel is accordingly fused with each potential feature surrounding the pixel to obtain feature to be fused corresponding to each potential feature; finally, the features to be fused corresponding to multiple potential features surrounding the pixel are combined to obtain the continuous feature corresponding to the pixel.

[0079] In some embodiments, for example, the at least one potential feature of any pixel include: z_{00} , z_{01} , z_{11} , and z_{10} , and the area ratios (i.e. the ratio between the area of each potential feature and the preset area) corresponding to the potential features z_{00} , z_{01} , z_{11} , and z_{10} are a_0 , a_1 , a_2 and a_3 , respectively; the preset area is the sum of the corresponding areas of the four potential features; then, the area ratio corresponding to each potential feature is fused with each potential feature to form continuous features, and then the continuous features corresponding to the pixel may be expressed as $\{a_0z_{00}, a_1z_{01}, a_2z_{11}, a_3z_{10}\}$.

[0080] At S206, the resolution and dynamic range of the continuous feature corresponding to each pixel are adjusted to obtain the adjusted target panorama.

[0081] In some embodiments, the resolution and dynamic range of the continuous feature corresponding to each pixel may be amplified successively, and then the adjusted panorama with improved resolution and dynamic range compared with the target panorama is obtained. Moreover, different multi-layer perceptrons may be used successively to adjust the continuous feature corresponding to each pixel to obtain the adjusted target panorama.

[0082] In some possible implementation modes, firstly, up-sampling may be

performed on the continuous feature corresponding to each pixel in terms of resolution to obtain an intermediate feature corresponding to each pixel and an intermediate panorama; in this way, by performing up-sampling on the continuous feature corresponding to each pixel in the target panorama in terms of resolution, namely improving its corresponding resolution, the intermediate panorama with high quality, namely higher resolution, compared with the target panorama, can be obtained; thus, the panorama which includes more information can be obtained; at the same time, the corresponding dynamic range can be improved while the resolution is improved, so as to obtain the adjusted target panorama with a higher resolution and a higher dynamic range; secondly, based on the positional code of each pixel in the target panorama, up-sampling is performed on the intermediate feature corresponding to each pixel in terms of dynamic range to obtain the adjusted target panorama; in this way, the dynamic range is further adjusted for the pixel with a higher resolution, and then the adjusted target panorama with the higher resolution as well as the higher dynamic range compared with the target panorama can be obtained; thus, the panorama with high quality and HDR can be obtained more conveniently. S206 may be implemented by S2061 and S2062 below (not shown in the accompanying drawings).

- [0083] At S2061, up-sampling is performed on the continuous feature corresponding to each pixel in terms of resolution to obtain the intermediate feature corresponding to each pixel and the intermediate panorama.
- [0084] In some embodiments, by performing up-sampling on the continuous feature corresponding to pixel at any positional code in the target panorama in terms of resolution, namely improving the corresponding resolution, the continuous feature with the higher resolution corresponding to each pixel can be obtained, and image reconstruction may also be performed based on the continuous feature with the higher resolution corresponding to each pixel, so as to obtain the intermediate panorama with the higher resolution compared with the target panorama.
- [0085] In some embodiments, the number of pixels included in the intermediate panorama is more than that in the target panorama, and the resolution in the intermediate panorama corresponding to each positional code in the spherical positional encoding is higher than the resolution in the target panorama corresponding to each positional code in the spherical positional encoding.
- [0086] At S2062, based on the positional code of each pixel in the target panorama, up-sampling is performed on the intermediate feature corresponding to each pixel in terms of dynamic range to obtain the adjusted target panorama.
- [0087] In some embodiments, based on the positional code of each pixel in the target panorama, up-sampling may further be performed on the intermediate feature corresponding to each pixel in terms of dynamic range, that is, the corresponding dynamic range is improved, and then the continuous feature with the higher resolution as well as the higher dynamic range is obtained; furthermore, based on the continuous feature with the higher resolution as well as the higher dynamic range, feature decoding and feature fusion are performed to obtain the corresponding image, namely the adjusted target panorama.
- [0088] In some embodiments, based on the input scene description text, the global feature and the local feature of relevant images are successively acquired in the global

feature book and the local feature book corresponding to the preset scene image set, and the spherical positional encoding corresponding to the preset panorama is also introduced in the process of acquiring the features; in this way, on the basis of performing feature decoding and feature synthesis on the acquired relevant features, the panorama that has the spherical structure and matches the text semantics of the scene description text is obtained; at the same time, up-sampling may also be performed on the obtained panorama in terms of resolution and dynamic range, so as to obtain the high-quality panorama with the higher resolution and the higher dynamic range.

[0089] In some embodiments, the panorama generation network may be used to process the obtained scene description text, and then the target panorama corresponding to the scene description text and the corresponding adjusted target panorama are obtained. The panorama generation network is obtained by training a panorama generation network to be trained, and a training process of the panorama generation network to be trained may be implemented by the following process.

[0090] Firstly, a sample text feature corresponding to a sample image may be obtained, here, the sample text feature may include: a pseudo text feature and a text-like feature, the pseudo text feature being obtained by perturbing a text feature corresponding to the sample image, and the text-like feature being a text feature corresponding to an image whose similarity with the sample image satisfies a preset condition; in this way, the panorama generation network to be trained is trained through the sample text feature obtained by fusing the pseudo text feature and the text-like feature, so that the scene semantics of the corresponding panorama generated by the trained panorama generation network based on the relevant text matches the text semantics of the relevant text better; at the same time, the purpose of using the sample text feature obtained based on the fusion of the pseudo text feature and the text-like feature for training is to obtain the global sample feature that is closer to and even aligned with the global scene semantics of the sample image, and then obtain the more accurate local sample feature based on the aligned global sample feature. Secondly, according to the sample text feature and the global feature book, the global sampler to be trained in the panorama generation network to be trained is trained to obtain the trained global sampler, the global feature book being obtained by performing feature coding on the preset scene image set based on the trained global encoder; in this way, taking the sample text feature as a condition, the relevant global feature associated with the sample text feature is determined in the global feature book, and the global sampler to be trained is trained to obtain the trained global sampler, so as to use the trained global sampler later to determine, in the global feature book, the global sample feature aligned with the text semantics of the sample text feature. Finally, according to the global sample feature, the spherical positional encoding corresponding to the preset panorama and the local feature book, the local sampler to be trained in the panorama generation network to be trained is trained to obtain the trained local sampler, the local feature book being obtained by performing feature coding on the preset scene image set based on the trained local encoder; in this way, the local sampler to be trained is trained, that is, taking the global sample feature and the spherical positional encoding as conditions, the local feature corresponding to each positional code in the spherical positional encoding is determined in the local feature book, and by training the local sampler to be trained, the trained local sampler obtained later determines, in the local feature book, the local feature corresponding to each positional code, so that the corresponding panorama can be obtained later based on the local feature corresponding

to each positional code. Thus, by training the panorama generation network to be trained in blocks and in stages, the speed and accuracy of training the panorama generation network to be trained can be improved; at the same time, in the process of sampling the features, the accuracy of the generated panorama can be improved by classifying the features corresponding to the image into the global feature and the local feature, namely classifying into the rough global feature corresponding to the image and the local feature of the fine granularity corresponding to the image, and training related modules of the panorama generation network to be trained in stages. The training process of the panorama generation network to be trained may be implemented by the operations shown in FIG. 3. FIG. 3 is a flowchart of a method for training a panorama generation network provided by an embodiment of the application. Descriptions will be made in combination with the operations shown in FIG. 3.

[0091] At S31, a sample text feature is acquired.

[0092] In some embodiments, the sample text feature includes a pseudo text feature and a text-like feature. The pseudo text feature is obtained by perturbing a text feature corresponding to a sample image. The text-like feature is a text feature corresponding to an image whose similarity with the sample image satisfies a preset condition.

[0093] In some embodiments, the text feature corresponding to the sample image may refer to the text feature obtained by performing feature coding on the sample image by a sampled image encoder; at the same time, the pseudo text feature refers to the text feature obtained by perturbing the text feature corresponding to the sample image; here, perturbing may refer to adjusting some features in the text feature corresponding to the sample image, so as to obtain the text feature that is partially different from text adjustment corresponding to the sample image.

[0094] In some embodiments, the process of acquiring the text-like feature may refer to the process of acquiring the text feature corresponding to the sample image, which will not be elaborated here. The image whose similarity with the sample image satisfies a preset condition is the similar image associated with the sample image, and the number of similar images may be either one or two or more. Here, when the number of similar images is two or more, the text-like feature may refer to the text feature obtained by fusing the text feature corresponding to each similar image and then averaging, and may also refer to a text feature set corresponding to each similar image.

[0095] In some embodiments, the image whose similarity with the sample image satisfies the preset condition may refer to the image whose value corresponding to the similarity with the sample image is greater than a preset value, and may also refer to the image whose similarity with the content included in the picture of the sample image satisfies the preset condition.

[0096] In some embodiments, the sample text feature is the text feature obtained by fusing the pseudo text feature and the text-like feature, and the sample text feature may also be obtained by sorting the pseudo text feature and the text-like feature in a preset order.

[0097] At S32, according to the sample text feature and the global feature book, a global sampler to be trained in the panorama generation network to be trained is trained to obtain a trained global sampler.

- [0098] In some embodiments, the global feature book is obtained by encoding the preset scene image set based on the trained global encoder.
- [0099] In some embodiments, the global sampler to be trained may also be used to sample, in the global feature book, the global feature associated with the sample text feature. In the sampling process, until the corresponding semantics of the global feature obtained by sampling and the sample text feature are aligned with, namely infinitely close to, each other, the corresponding global sampler may be considered as the trained global sampler.
- [00100] In some embodiments, the global feature book is obtained by performing feature coding on the preset scene image set based on the trained global encoder. Here, feature coding may be performed on each preset scene image in the preset scene image set based on the trained global encoder to obtain the global feature corresponding to each preset scene image, and then the global feature book is obtained.
- [00101] In some possible implementation modes, in the process of training the global encoder to be trained, the image to be trained may be input into the global encoder to be trained, so as to obtain the corresponding global continuous sample feature and global discrete sample feature; then, feature decoding is performed based on the obtained relevant features to reconstruct the image; finally, based on the loss of the similarity between the reconstructed image and the image to be trained, the global encoder to be trained is trained to obtain the trained global encoder. The training process of the trained global encoder may be implemented by the following procedure.
- [00102] At the first step, the global encoder to be trained is used to perform feature coding on the image to be trained, so as to obtain a global continuous sample feature and a global discrete sample feature corresponding to the global continuous sample feature.
- [00103] In some embodiments, the global encoder to be trained may be used to perform feature coding on the image to be trained, and then global continuous sample code corresponding to the image to be trained is obtained; then, the global continuous sample code may be represented discretely, and global discrete sample code corresponding to the global continuous sample code may be obtained.
- [00104] Here, the global encoder to be trained may be a Vector-Quantized Variational AutoEncoder (VQVAE).
- [00105] At the second step, feature decoding is performed on the global discrete sample feature to obtain a second reconstructed image.
- [00106] In some embodiments, feature decoding may be performed on the global discrete sample feature, that is, image reconstruction may be performed on the global discrete sample feature, so as to obtain a third reconstructed image.
- [00107] At the third step, a third reconstruction loss of the similarity between the second reconstructed image and the image to be trained and a first adversarial loss are determined.
- [00108] In some embodiments, the third reconstruction loss and the first adversarial loss between the second reconstructed image and the image to be trained may be determined respectively. The third reconstruction loss represents the loss of supervising the similarity between the two, and the first adversarial loss represents the loss of

supervising true and false labels between the two.

- [00109] At the fourth step, a first quantization loss between the global discrete sample feature and the global continuous sample feature is determined.
- [00110] The first quantization loss represents the loss of supervising the quantization between the global discrete sample feature and the global continuous sample feature.
- [00111] At the fifth step, the third reconstruction loss, the first adversarial loss and the first quantization loss are fused to obtain a first overall loss.
- [00112] In some embodiments, the third reconstruction loss, the first adversarial loss and the first quantization loss are fused to obtain the first overall loss. Here, the third reconstruction loss, the first adversarial loss and the first quantization loss may be summed up to obtain the first overall loss; or, the corresponding weights of the third reconstruction loss, the first adversarial loss and the first quantization loss may be determined respectively, and the corresponding weights are multiplied correspondingly with the third reconstruction loss, the first adversarial loss and the first quantization loss to obtain three corresponding losses; then, the three losses are summed up to obtain the first overall loss.
- [00113] At the sixth step, based on the first overall loss, a network parameter of the global encoder to be trained is adjusted, so that the loss of a feature output by the trained global encoder satisfies a convergence condition.
- [00114] In some embodiments, the first overall loss may be used to adjust the network parameter, e.g., weight coefficient, of the global encoder to be trained, so that the loss of the feature output by the trained global encoder converges.
- [00115] In some possible implementation modes, firstly, the global sampler to be trained is used to determine, in the global feature book, an initial global feature matching the sample text feature; in this way, the global sampler to be trained is used to find the initial global feature matching the sample text feature, which provides the basis for subsequent image reconstruction; secondly, feature decoding is performed on the initial global feature to obtain a first reconstructed image, and a first reconstruction loss of the cosine similarity between the first reconstructed image and the sample image is further determined; thus, the loss of supervising the similarity between the sample image and the first reconstructed image is determined, so as to train the global sampler to be trained later based on the loss; that is, based on the first reconstruction loss, the network parameter of the global sampler to be trained is adjusted, so that the first reconstruction loss output by the trained global sampler satisfies the convergence condition; in this way, taking the sample text feature as the condition, the relevant global feature associated with the sample text feature is determined in the global feature book, that is, by training the global sampler to be trained, the trained global sampler obtained later obtains, by sampling in the global feature book, the relevant global sample feature whose text semantic is aligned with the text semantics of the sample text feature. S32 may be implemented by S321 to S324 below (not shown in the accompanying drawings).
- [00116] At S321, the global sampler to be trained is used to determine, in the global feature book, the initial global feature that matches the sample text feature.
- [00117] In some embodiments, the global sampler to be trained is used to search, in the global feature book, for the initial global feature associated with the sample text feature.

- [00118] In some possible implementation modes, firstly, the global sampler to be trained may be used to determine an empty sequence with empty information, that is, the first feature in a first sequence, according to the sample text feature, the first sequence here being determined based on an index sequence of the global feature book; in this way, the first feature in the first sequence is determined; secondly, based on the i -th feature in the first sequence and the sample text feature, the $(i+1)$ -th feature in the first sequence is predicted in the global feature book until a first feature sequence containing the same number of features as the first sequence is obtained, that is, the corresponding first feature sequence is determined in an autoregressive way, so that the determined first feature sequence matches the sample text feature better, and then the first feature sequence is determined as the initial global feature. S321 may be implemented by the following process.
- [00119] At the first step, the global sampler to be trained is used to determine the first feature in the first sequence according to the sample text feature.
- [00120] The first sequence represents the empty sequence associated with the index sequence of the global feature book. The first sequence is an empty sequence. Here, the first feature in the first sequence may be predicted based on the sample text feature, or the sample text feature may be directly determined as the first feature in the first sequence.
- [00121] At the second step, based on the i -th feature in the first sequence and the sample text feature, the $(i+1)$ -th feature in the first sequence is predicted in the global feature book until the first feature sequence containing the same number of features as the first sequence is obtained. i is a positive integer greater than or equal to 1.
- [00122] In some embodiments, based on the i -th feature in the first sequence and the sample text feature, the $(i+1)$ -th feature in the first sequence is predicted in the global feature book until the first feature sequence is obtained, which is an autoregressive process; that is, based on the i -th determined feature in the first sequence and the input sample text feature, the global feature book is searched for the corresponding feature, and the feature is determined as the $(i+1)$ -th feature in the first sequence; this step is performed repeatedly until the first feature sequence is obtained.
- [00123] At the third step, the first feature sequence is determined as the initial global feature.
- [00124] In some embodiments, the first feature sequence may be taken as the initial global feature, and the initial global feature is a discrete feature sequence.
- [00125] At S322, feature decoding is performed on the initial global feature to obtain the first reconstructed image.
- [00126] In some embodiments, feature decoding is performed to the initial global feature, that is, image reconstruction is implemented based on the initial global feature to obtain the first reconstructed image.
- [00127] At S323, the first reconstruction loss of the cosine similarity between the first reconstructed image and the sample image is determined.
- [00128] That is, the first reconstruction loss of supervising the cosine similarity between the first reconstructed image and the sample image is determined.

- [00129] At S324, based on the first reconstruction loss, the network parameter of the global sampler to be trained is adjusted, so that the first reconstruction loss output by the trained global sampler satisfies the convergence condition.
- [00130] In some embodiments, the first reconstruction loss may be used to adjust the network parameter, for example, the weight coefficient, of the global sampler to be trained, so that the first reconstruction loss output by the trained global sampler converges, or the output loss of the global feature converges.
- [00131] At S33, the trained global sampler is used to determine, in the global feature book, the global sample feature that matches the sample text feature.
- [00132] In some embodiments, the implementation process of using the trained global sampler to determine, in the global feature book, the global sample feature that matches the sample text feature may refer to the implementation process of S202 provided in the above embodiment, which will not be elaborated here.
- [00133] At S34, according to the global sample feature, the spherical positional encoding corresponding to the preset panorama and the local feature book, a local sampler to be trained in the panorama generation network to be trained is trained to obtain a trained local sampler.
- [00134] The local feature book is obtained by performing feature coding on the preset scene image set based on the trained local encoder.
- [00135] In some embodiments, the local sampler to be trained may be used to sample, in the local feature book, the local feature corresponding to each positional code in the SPE based on the global sample feature. The local feature book is obtained by performing feature coding on the preset scene image set based on the trained local encoder. Here, feature coding may be performed on each preset scene image in the preset scene image set based on the trained local encoder to obtain the local feature corresponding to each preset scene image, and then the local feature book is obtained.
- [00136] In some possible implementation modes, in the process of training the local encoder to be trained, a preset region in the image to be trained, namely any image block, may be input into the local encoder to be trained, so as to obtain the corresponding local continuous sample feature and local discrete sample feature; then, feature decoding and image reconstruction are performed based on the obtained relevant features; finally, based on the loss of the similarity between the reconstructed image block and the preset region in the image to be trained, the local encoder to be trained is trained to obtain the trained local encoder. The training process of the trained local encoder may be implemented by the following procedure.
- [00137] At the first step, the local encoder to be trained is used to perform feature coding on the preset region in the image to be trained, so as to obtain the local continuous sample feature and the local discrete sample feature corresponding to the local continuous sample feature.
- [00138] In some embodiments, the local encoder to be trained may be used to perform feature coding on the preset region in the image to be trained, and then the corresponding local continuous sample code is obtained; then, the local continuous sample code may be represented discretely, and the local discrete sample code corresponding to the local continuous sample code may be obtained. Here, the local encoder to be trained may be

the VQVAE.

- [00139] At the second step, feature decoding is performed on the local discrete sample feature to obtain a second reconstructed image block.
- [00140] In some embodiments, feature decoding may be performed on the local discrete sample feature, that is, image reconstruction may be performed on the local discrete sample feature, so as to obtain the second reconstructed image block.
- [00141] At the third step, a fourth reconstruction loss of the similarity between the second reconstructed image block and the preset region in the image to be trained and a second adversarial loss are determined.
- [00142] In some embodiments, the fourth reconstruction loss and the second adversarial loss between the second reconstructed image block and the preset region in the image to be trained may be determined respectively. The fourth reconstruction loss represents the loss of supervising the similarity between the two, and the second adversarial loss represents the loss of supervising the true and false labels between the two.
- [00143] At the fourth step, a second quantization loss between the local discrete sample feature and the local continuous sample feature is determined.
- [00144] The second quantization loss represents the loss of supervising the quantization between the local discrete sample feature and the local continuous sample feature.
- [00145] At the fifth step, the fourth reconstruction loss, the second adversarial loss and the second quantization loss are fused to obtain a second overall loss.
- [00146] In some embodiments, the fourth reconstruction loss, the second adversarial loss and the second quantization loss are fused to obtain the second overall loss. Here, the fourth reconstruction loss, the second adversarial loss and the second quantization loss may be summed up to obtain the second overall loss; or, the corresponding weights of the fourth reconstruction loss, the second adversarial loss and the second quantization loss may be determined respectively, and the corresponding weights are multiplied correspondingly with the fourth reconstruction loss, the second adversarial loss and the second quantization loss to obtain three corresponding losses; then, the three losses are summed up to obtain the second overall loss.
- [00147] At the sixth step, based on the second overall loss, the network parameter of the local encoder to be trained are adjusted, so that the loss of the feature output by the trained local encoder satisfies the convergence condition.
- [00148] In some embodiments, the second overall loss may be used to adjust the network parameter, for example, the weight coefficient, of the local encoder to be trained, so that the loss of the feature output by the trained local encoder converges.
- [00149] In some possible implementation modes, firstly, taking the global sample feature as the condition, the local sampler to be trained is used to determine, in the local feature book, the local sample feature corresponding to a preset positional code, and feature coding is performed on the sample feature to obtain a first reconstructed image block; secondly, the second reconstruction loss of the cosine similarity between the first reconstructed image block and the sample image block is further determined; in this way, the loss of supervising the similarity between the sample image block and the first

reconstructed image block is determined, so as to train the local sampler to be trained based on the loss later; finally, based on the second reconstruction loss, the network parameter of the local sampler to be trained is adjusted, so that the second reconstruction loss output by the trained local sampler satisfies the convergence condition; in this way, taking the global sample feature as the condition, the local sample feature corresponding to each position in the image with the spherical structure (spherical positional encoding) is determined in the local feature book, and the local sample feature is used to train the local sampler to be trained, so that the local feature corresponding to the trained global sampler obtained later can be more accurate. S34 may be implemented by S341 to S344 below (not shown in the accompanying drawings).

[00150] At S341, according to the global sample feature, the local sampler to be trained is used to determine, in the local feature book, the local sample feature corresponding to a preset positional code in the SPE.

[00151] In some possible implementation modes, firstly, the global sampler to be trained may be used to determine an empty sequence with empty information, that is, the first feature in a second sequence, according to the sample text feature, the second sequence here being determined based on an index sequence of the local feature book; in this way, the first feature in the second sequence is determined; secondly, based on the i -th feature in the second sequence and the sample text feature, the $(i+1)$ -th feature in the second sequence is predicted in the global feature book until a second feature sequence is obtained, that is, the corresponding second feature sequence is determined in the autoregressive way, so that the determined second feature sequence can better reflect the details corresponding to the sample text feature, and then the second feature sequence is determined as the local sample feature. S341 may be implemented by the following process.

[00152] At the first step, the local sampler to be trained is used to determine the first feature in the second sequence according to the global sample feature and the preset positional code.

[00153] The second sequence represents the empty sequence associated with the index sequence of the local feature book. The second sequence may be an empty sequence. Here, the first feature in the second sequence may be predicted based on the global sample feature and the preset positional code, or the global sample feature and the preset positional code may be directly determined as the first feature in the second sequence.

[00154] At the second step, based on the i -th feature in the second sequence, the global sample feature and the preset positional code, the $(i+1)$ -th feature in the second sequence is predicted in the local feature book until the second feature sequence containing the same number of features as the second sequence is obtained. i is a positive integer greater than or equal to 1.

[00155] In some embodiments, based on the i -th feature in the second sequence, the global sample feature and the preset positional code, the $(i+1)$ -th feature in the second sequence is predicted in the global feature book until the second feature sequence is obtained, which is an autoregressive process, that is, based on the i -th determined feature in the second sequence, the global sample feature and the preset positional code, the local feature book is searched for the corresponding feature, and the feature is determined as the $(i+1)$ -th feature in the second sequence. This step is performed repeatedly until the

first feature sequence is obtained.

- [00156] At the third step, the second feature sequence is determined as the local sample feature.
- [00157] In some embodiments, the second feature sequence may be taken as the local sample feature, and the local sample feature is a discrete feature sequence.
- [00158] At S342, feature decoding is performed on the local sample feature to obtain a first reconstructed image block.
- [00159] In some embodiments, feature decoding is performed on the local sample feature, that is, image reconstruction is implemented based on the initial global feature to obtain the first reconstructed image block.
- [00160] At S343, the second reconstruction loss of the cosine similarity between the first reconstructed image block and the sample image block is determined.
- [00161] In some embodiments, the sample image block is the image region at the preset position in the sample image. There is a mapping relationship between the preset position and the preset positional code.
- [00162] In some embodiments, the second reconstruction loss of supervising the cosine similarity between the first reconstructed image block and the sample image block is determined.
- [00163] At S344, based on the second reconstruction loss, the network parameter of the local sampler to be trained is adjusted, so that the second reconstruction loss output by the trained local sampler satisfies the convergence condition.
- [00164] In some embodiments, the second reconstruction loss may be used to adjust the network parameter, for example, the weight coefficient, of the local sampler to be trained, so that the second reconstruction loss output by the trained local sampler converges, or the output loss of the local feature converges.
- [00165] In some embodiments, in the case where the panorama generation network to be trained includes an image adjustment sub-network to be trained, the image adjustment sub-network to be trained may be trained to obtain the trained image adjustment sub-network, and then the trained image adjustment sub-network may be used to adjust the resolution and/or the dynamic range of a relevant initial panorama, so as to obtain the target panorama with higher resolution and/or higher dynamic range compared with the initial panorama. The method for training a panorama generation network provided by the embodiment of the present may also include the following process.
- [00166] At the first step, a panoramic sample image block and a first panoramic image block and a second panoramic image block that correspond to the panoramic sample image block and have different dynamic ranges are acquired.
- [00167] In some embodiments, the panoramic sample image block may be obtained by performing region division on a panoramic sample image to obtain the corresponding image blocks; at the same time, the first panoramic image block and the second panoramic image block that correspond to the panoramic sample image block and have different dynamic ranges may be obtained by adjusting the dynamic ranges corresponding to the panoramic sample image block with different parameters.

- [00168] In some possible implementation processes, the panoramic sample image block, the first panoramic image block and the second panoramic image block may be all obtained based on the same sample image, which can improve the accuracy of the acquired data to be trained and the convenience of acquiring the data to be trained. The process of obtaining the relevant image blocks in the first step may be implemented by the following process.
- [00169] Firstly, a panoramic sample image is acquired.
- [00170] In some embodiments, a panoramic sample image is acquired first. The panoramic sample image may be obtained by splicing multiple images, or directly using a panorama camera to acquire images of a related object or scenery.
- [00171] Next, tone mapping is performed on the panoramic sample image and any image block is selected to obtain the first panoramic image block.
- [00172] In some embodiments, adjustment of a relevant parameter, e.g., tone mapping, of the panoramic sample image may be performed, and then the region selection is performed on the adjusted panoramic sample image to obtain the corresponding first panoramic image block.
- [00173] Then, down-sampling is performed on the first panoramic image block to obtain the panoramic sample image block.
- [00174] In some embodiments, down-sampling, which may refer to the resolution, is performed on the obtained first panoramic image block, and then the panoramic sample image block is obtained.
- [00175] Finally, brightness calibration is performed on the panoramic sample image, and the second panoramic image block corresponding to the first panoramic image block is selected.
- [00176] In some embodiments, brightness calibration may be performed simultaneously on the obtained panoramic sample image, and the second panoramic image block in the same region as the first panoramic image block in the panoramic sample image is selected.
- [00177] At the second step, the image adjustment sub-network to be trained is used to perform feature coding and tone adjustment on each pixel in the panoramic sample image block, so as to obtain a third panoramic image block and a fourth panoramic image block.
- [00178] In some embodiments, the image adjustment sub-network to be trained is used to perform feature coding on each pixel in the panoramic sample image block in turn to obtain the continuous sample feature corresponding to each pixel, and then the resolution and the dynamic range are adjusted based on the continuous sample feature corresponding to each pixel, so as to obtain the third panoramic image block with changed resolution and the fourth panoramic image block with changed resolution and dynamic range compared with the panoramic sample image block.
- [00179] In some possible implementation modes, firstly, the image adjustment sub-network to be trained is used to perform feature coding on each pixel in the panoramic sample image block to obtain the continuous sample feature corresponding to each pixel, so as to adjust the relevant parameter based on the continuous sample feature later; secondly, the resolution and the dynamic range are adjusted based on the continuous

sample feature corresponding to each pixel, so as to obtain the third panoramic image block with changed resolution and the fourth panoramic image block with changed resolution and dynamic range compared with the panoramic sample image block; in this way, multiple adjusted panoramic image blocks are obtained, so that similarity comparison can be performed between multiple adjusted panoramic image blocks and the panoramic sample image block, and then the image adjustment sub-network to be trained is trained based on the loss corresponding to a comparison result. Thus, the accuracy of training the image adjustment sub-network to be trained can be improved. The operation that the image adjustment sub-network to be trained is used to perform feature coding and tone adjustment on each pixel in the panoramic sample image block, so as to obtain a third panoramic image block and a fourth panoramic image block may be implemented by the following process.

[00180] Firstly, the image adjustment sub-network to be trained is used to perform feature coding on each pixel in the panoramic sample image block, so as to obtain the continuous sample feature corresponding to each pixel.

[00181] Here, the corresponding implementation process of performing feature coding on each pixel in the panoramic sample image block to obtain the continuous sample feature corresponding to each pixel is similar to the implementation process of determining the continuous feature corresponding to each pixel in the target panorama in S205 provided in the above embodiment, which will not be elaborated here.

[00182] Secondly, up-sampling is performed on the continuous sample feature corresponding to each pixel in terms of resolution, so as to obtain an intermediate sample feature corresponding to each pixel and the corresponding third panoramic image block.

[00183] Here, the corresponding implementation process of obtaining the corresponding intermediate sample feature and the third panoramic image block based on the continuous sample feature corresponding to each pixel is similar to the implementation process of determining the intermediate feature corresponding to each pixel and the intermediate panorama in S2061 provided in the above embodiment, which will not be elaborated here.

[00184] Finally, based on the spherical code corresponding to each pixel, up-sampling is performed on the intermediate sample feature corresponding to each pixel in terms of dynamic range to obtain the fourth panoramic image block.

[00185] Here, the corresponding implementation process of obtaining the fourth panoramic image block based on the spherical codes corresponding to each pixel and the intermediate sample feature corresponding to each pixel is similar to the implementation process of determining the adjusted target panorama in S2062 provided in the above embodiment, which will not be elaborated here.

[00186] At the third step, a first loss of the similarity between the first panoramic image block and the third panoramic image block is determined, and a second loss of the similarity between the second panoramic image block and the fourth panoramic image block is determined.

[00187] In some embodiments, the first loss of the similarity between the first panoramic image block and the third panoramic image block, and the second loss of the similarity between the second panoramic image block and the fourth panoramic image

block are determined respectively. The first loss represents the loss of supervising the similarity between the first panoramic image block and the third panoramic image block. The second loss represents the loss of supervising the similarity between the second panoramic image block and the fourth panoramic image block.

[00188] At the fourth step, based on the first loss and the second loss, the network parameter of the image adjustment sub-network to be trained is adjusted, so that the loss of the image output by the trained image adjustment sub-network satisfies the convergence condition.

[00189] In some embodiments, the network parameter, for example, the weight coefficient, of the image adjustment sub-network to be trained may be adjusted using the first loss and the second loss, so that the loss of the image output by the trained image adjustment sub-network satisfies the convergence condition.

[00190] The panorama generation method and the method for training a panorama generation network are illustrated below with a specific embodiment. However, it should be noted that the specific embodiment is intended only to better illustrate the embodiments of the application and does not form an improper limit to the embodiments of the application.

[00191] The panorama with high quality and HDR can provide sufficient scene texture details and realistic lighting because of having the HDR and HR. In the related art, the acquisition of the panorama with high quality and HDR often requires an expensive panorama camera and needs professionals for post-calibration and retouching. So the cost of acquiring the panorama with high quality and HDR is high. For example, in a practical application process, it is difficult to generate a high-resolution scene image with an extremely high resolution, and the high-resolution scene image does not have the relevant spherical structural transcendence when it is generated, so the spherical structure information of a 360-degree scene cannot be maintained; moreover, in the practical application process, multiple common images are also spliced to obtain a panorama, so it is difficult for the generated panorama to maintain the consistency of single scene semantics, and the generated panorama generally cannot represent a true illumination value in the scene of the image.

[00192] At the same time, a large number of text-image pairs are usually used for model training in the process of generating a panorama based on text drive, which is usually difficult to align the scene semantics and the text semantics involved in the text-image pairs in the ZSL. As a result, the process of generating a panorama based on a text is tedious, and the quality of the generated panorama is not high.

[00193] Based on this, an embodiment of the application provides a panorama generation method and a method for training a panorama generation network to build a model of generating a panorama based on text drive; the model takes an input scene text as a scene description to directly acquire a panorama with high quality and HDR by synthesizing; that is, the purpose of the application is to generate a panorama with high quality and HDR which has a resolution of at least 4K when the given scene description text is taken as the input. The application may be used for generating a panorama with high quality and HDR taking a text in a free form as the scene description text. FIG. 4 is a frame diagram corresponding to a method for generating a panorama based on text drive that includes two implementation stages provided by an embodiment of the

application. In the first stage, an input text is converted into a panorama with LDR based on dual-codebook discrete representation. As shown in FIG. 4, firstly, feature coding is performed on the input scene description text 401 "ocean waves crashing on shore under blue and white cloudy sky during daytime" by means of a text editor 402 to obtain the text feature 403; secondly, the global sampler 404 is used to sample the overall semantic feature, namely the global feature (global condition), of the subsequent constructed image from the global feature book 405 according to the input text feature 403. Then, the local sampler 407 is used to sample the local feature in the local feature book 406 according to the obtained global condition and the spherical positional encoding 409, and feature decoding and feature synthesis are performed, so as to obtain a panorama with LDR 408. In the second stage, the panorama with LDR 408 obtained in the first stage is adjusted to the panorama with HDR 414. As shown in FIG. 4, a structured potential code 413 is taken as each pixel (i.e. the pixel on each spherical positional encoding) in part of region image in the preset panorama 412. Before the continuous feature of each pixel is determined, part of image blocks may be selected for the panorama with LDR 408 first, then the encoder 411 is used to perform feature coding on each pixel, and the continuous feature is determined based on the potential code after the corresponding feature is obtained; at the same time, a Super-Resolution Inverse Tone Mapping Operator (SR-ITMO) 410 is used to adjust the pixel resolution and dynamic range of the panorama with LDR 408, so as to improve the pixel resolution and dynamic range of the panorama with LDR 408, and then obtain the panorama with HDR 414.

[00194] Here, the spherical positional encoding 409 is the corresponding spherical positional encoding of each pixel in the panorama in the spherical structure. The features involved in the first stage are all discrete features, and the features involved in the second stage are all continuous features. At the same time, in the first stage, a panorama with LDR and LR is generated taking the input scene description text as the input; in the second stage, the panorama generated in the first stage is amplified by the SR-ITMO in terms of the resolution and the dynamic range to obtain a panorama with HDR and HR. In the first stage, a panorama with maintained image structure consistency and overall semantics can be generated without using text panorama paired training data. In the second stage, a panorama with ultrahigh resolution (4K+) and high quality is generated. Through the proposed two-stage framework, a panorama with high quality and HDR which has a resolution of at least 4K can be generated based on text description. FIG. 5 is a panorama generated based on a panorama generation method provided by an embodiment of the application. In FIG. 5, 501 is an image with more realistic lighting effects and immersive environment textures, which is generated based on this scheme; 502, 503 and 504 are panoramas applied to real scenes, such as a 3D scene and an immersive virtual reality scene.

[00195] In the implementation process of generating a panorama based on an input text, the SPE is introduced into the application to represent 360-degree scene information in the panorama.

[00196] Here, considering the spherical structure of the panorama, the spherical positional encoding is taken as a condition in the process of generating a panorama based on text drive; for example, a preset panorama is $I \in R^{H \times W \times 3}$; here, each pixel in the panorama may be represented as a 3D point S on a unit sphere, and H , W and 3 are the length of image, the width of image, and the number of image channels, respectively.

$I(i, j)$ is any pixel in the preset panorama, and in $I(i, j)$, $i \in \{1, \dots, H\}$ and $j \in \{1, \dots, W\}$; here, $I(i, j)$ may be converted into the pixel $S(\theta, \phi, r)$ in the panorama, as shown in formula (1):

$$\begin{cases} S(\theta, \phi, r) = I(i, j), \\ \theta = (2i/H - 1)\pi, \\ \phi = (2j/W - 1)\pi/2, \\ r = 1 \end{cases} \quad \text{formula (1);}$$

where formula (1) represents the implementation process of projecting any pixel in a 2D image to a 360-degree panorama, that is, the center point of the sphere in the spherical structure corresponding to the 360-degree panorama is the position where the 360-degree panorama is shot in the 3D world.

[00197] In addition, any pixel S in the preset panorama may also be mapped to an image corresponding to a higher dimension based on the Fourier transform, as shown in formula (2):

$$\begin{cases} \gamma(\theta) = [\sin(2^0 \pi \theta), \cos(2^0 \pi \theta), \dots, \sin(2^{L-1} \pi \theta), \cos(2^{L-1} \pi \theta)], \\ \gamma(\phi) = [\sin(2^0 \pi \phi), \cos(2^0 \pi \phi), \dots, \sin(2^{L-1} \pi \phi), \cos(2^{L-1} \pi \phi)], \end{cases} \quad \text{formula (2);}$$

where $\gamma(\cdot)$ is any pixel in the image corresponding to a higher dimension than 3D, and L is the number of dimensions.

[00198] Here, the training process of the panorama generation network provided by the embodiment of the application is the training process of the panorama generation network which generates a panorama based on text drive; referring to the implementation process of generating a panorama based on text drive, the training process may be divided into two stages.

[00199] The first stage is the training process corresponding to using a natural language description text to generate a panorama with LDR and LR; here, the LR may refer to 1024×512. The first stage may be further divided into the following three parts.

[00200] 1) Representation by dual codebooks: dual codebooks are constructed to represent the overall appearance of an image and local fine granularity details of an image.

[00201] 2) Global sampler conditional on a text: global semantics corresponding to the images is extracted from a global codebook conditional on an input text.

[00202] 3) Structure-aware local sampler: based on the overall semantics corresponding to the images, the local fine-grained details of an image are acquired and feature decoding and feature synthesis are performed based on the spherical position encoding, so as to obtain a panorama with LDR and LR.

[00203] FIG. 6 is a schematic diagram of an implementation framework for generating a panorama with LDR and LR based on an input text provided by an embodiment of the application. In FIG. 6, 601 is using the discrete features corresponding to the dual codebooks to represent the overall semantics and the local fine granularity information of

an image respectively; 602 is the training process and application process of the global sampler conditional on a text; 603 is the training process and application process of the structure-aware local sampler.

[00204] 1) Representation by dual codebooks: the VQVAE is adopted, and the image information is represented by the discrete features Z .

[00205] Firstly, image A is input into the VQVAE to obtain the continuous feature $\hat{z} = E(A) \in R^{h \times w \times c_z}$ corresponding to the image A , where h , w and c_z are the length of image, the width of image, and the number of image channels, respectively.

[00206] Next, the continuous feature $\hat{z}$ is represented discretely to obtain the discrete feature z_q , as shown in formula (3):

$$z_q = \arg \min_{z_k \in Z} \|\hat{z} - z_k\|_2 \quad \text{formula (3)}.$$

[00207] Thirdly, feature decoding is performed based on the discrete feature z_q to obtain a reconstructed image $\hat{A}$, where $\hat{A} = G(z_q)$.

[00208] Finally, the VQVAE may be trained based on the reconstructed image $\hat{A}$, the image A , etc. The purpose of training is that L satisfies the convergence condition. L may refer to the following formula (4):

$$L = L_{rec} + \left\| sg(z_q) - \hat{z} \right\|_2^2 + \left\| sg(\hat{z}) - z_q \right\|_2^2 + L_{GAN} \quad \text{formula (4)}.$$

where, in formula (4), L_{rec} is the loss of the similarity between the reconstructed image $\hat{A}$ and the image A , and the quantization loss between the continuous feature $\hat{z}$ and the discrete feature z_q is $\left\| sg(z_q) - \hat{z} \right\|_2^2 + \left\| sg(\hat{z}) - z_q \right\|_2^2$; $sg(\cdot)$ means stopping a gradient operation, and L_{GAN} represents the adversarial loss, referring to formula (5):

$$L_{GAN} = \log D(I) + \log(1 - D(\hat{I})) \quad \text{formula (5)}.$$

[00209] Based on this, in the application, two VQVAEs are used to represent the overall semantics, namely the global semantics, of the image and the fine granularity semantics of the image respectively, which may refer to the training process about the VQVAE. Here, as shown in 601 in FIG. 6, the global codebook Z_g is established first, and a down-sampled panorama $A_g \in R^{128 \times 256 \times 3}$ is reconstructed by training a VQVAE. E_g and G_g are the corresponding encoder and decoder respectively. The encoder E_g converts A_g into the continuous feature $\hat{z}_g = E_g(A_g) \in R^{8 \times 16 \times c_z}$. Then, in order to represent enough details and different textures of the image to correspond to different

scenes, another local codebook Z_l is trained. The VQVAE is used to reconstruct the HR $A_l \in R^{128 \times 256 \times 3}$. E_l and G_l are the corresponding encoder and decoder, and a fine granularity feature $Z_l \in R^{16 \times 16 \times c_z}$ may be obtained. During training, A_l is randomly clipped from a complete panorama with LDR, $A \in R^{512 \times 1024 \times 3}$, and C_z is set to 256.

[00210] 2) Global sampler conditional on a text: feature sampling is performed in the global codebook which is obtained by performing feature coding on the preset scene image set based on the trained global encoder, so as to align the input text with the scene of the image in text language training.

[00211] In the process of training the global sampler conditional on a text, the feature obtained from sampling by the global sampler are aligned with the input text through an unsupervised learning technology, namely K-nearest neighbor and contrast learning. In the training process, based on a text condition C_{txt} , the aim is to approximate the real semantics of the target panorama without the text panorama paired training data. E_{txt} and E_{img} serve as a text encoder and an image encoder of CLIP, respectively. Feature coding is first performed on an input sample image based on the image encoder E_{img} , and then the obtained features are perturbed to generate the pseudo text feature $\hat{C}_{\text{txt}}$, as shown in FIG. (6):

$$\hat{C}_{\text{txt}} = (1 - \alpha)E_{\text{img}}(I) + \alpha \varepsilon \left\| E_{\text{img}}(I) \right\|_2 / \left\| \varepsilon \right\|_2 \quad \text{formula (6);}$$

where α is a fixed hyper-parameter, and ε is Gaussian noise $\varepsilon \sim N(0, I)$. At the same time, K images that are most similar to image I are acquired, and feature coding is performed to obtain $C_{\text{knn}} = \left\{ E_{\text{img}}(I_k) \right\}_{k=1}^K$, which is taken as an additional text of $\hat{C}_{\text{txt}}$, that is, the text condition C_{txt} is obtained, as shown in 602 in FIG. 6 and formula (7):

$$C_{\text{txt}} = \left\{ \hat{C}_{\text{txt}}, C_{\text{knn}} \right\} \quad \text{formula (7).}$$

[00212] Here, the global sampler conditional on a text may sample the overall feature z_g in the global codebook Z_g based on the input text condition C_{txt} in the autoregressive way, where z_g is a discrete feature sequence which is determined based on $s \in \left\{ 0, \dots, |Z_g| - 1 \right\}^{8 \times 16}$. When the global feature is sampled, the first feature in the sequence is determined based on C_{txt} , and the feature in a subsequent prediction sequence is determined based on the previous determined feature and the text condition C_{txt} . The feature sampled by the global sampler conditional on a text accords to an exponential distribution, namely $p(s_i | s < i, C_{\text{txt}})$, and the feature is fully represented as

$p(s) = \prod_i p(s_i | s < i, C_{\text{txt}})$. The global sampler is trained based on the following purposes, as shown in formula (8):

$$L_{\text{text}} = E_{I \sim p(I)} [-\log p(s)] + L_{\text{con}} \quad \text{formula (8);}$$

where L_{con} is contrast regularization, representing that based on the similarity between the images obtained by performing feature decoding to the input images and the predicted global feature, the global sampler is trained, so that the loss output by the trained global sampler converges, as shown in formula (9):

$$L_{\text{con}} = -\tau \sum_{i=1}^n \log \frac{\exp((E_{\text{img}}(I_i) \cdot \hat{C}_{\text{txt}}) / \tau)}{\sum_{j=1}^n \exp((E_{\text{img}}(I_j) \cdot \hat{C}_{\text{txt}}) / \tau)} \quad \text{formula (9).}$$

[00213] Here, the cosine similarity $\tau \in (0, 1]$ between the real image feature and the pseudo text feature is maximized as a hyper-parameter. The global sampler is trained through formula (8) to align the scene semantics of the image output by the global sampler with the text semantics corresponding to the input text feature.

[00214] 3) Structure-aware local sampler: given an overall condition Z_g obtained, the use of the structure-aware local sampler focuses on synthesizing the local features with sufficient fine details; in addition, the structure-aware local sampler also needs to use the internal structure of the panorama, namely the spherical structure information, to obtain more accurate local information of an image.

[00215] In the process of training the structure-aware local sampler, mainly through the unsupervised learning technology, another feature converter is used as the structure-aware local sampler to sample the local feature from the local codebook Z_l . Here, considering the overall situation Z_g , the purpose is synthesizing the image block $I_l \in R^{256 \times 256 \times 3}$ later. The spherical positional encoding represents any pixel in the panorama; here, the coordinates of any pixel corresponding to the spherical positional encoding may be denoted as $S(\theta_l, \phi_l, 1)$. Based on formula (1), $\theta_l \in R^{256 \times 256}$, and $\phi_l \in R^{256 \times 256}$, as shown in 603 in FIG. 6.

[00216] Here, Z_l is a discrete feature sequence, which is determined based on $s \in \{0, \dots, |Z_l| - 1\}^{16 \times 16}$, and the feature in the subsequent prediction sequence is determined based on the previous determined feature and the current positional code. The feature sampled by the global sampler conditional on a text accords to an exponential distribution, namely $p(s_i | s < i, z_g, \gamma(\theta_l), \gamma(\phi_l))$. The structure-aware local sampler is trained based on the following purposes, as shown in formula (10):

$$L_{\text{local}} = E_{I_t \sim p(I_t)} [-\log p(s)] \quad \text{formula (10);}$$

where $p(s) = \prod_i p(s_i | s < i, z_g, \gamma(\theta_i), \gamma(\phi_i))$.

[00217] Here, in an inference process, a complete panorama may be generated with a sliding window based on the structure-aware local sampler.

[00218] In the second stage, the panorama generated in the first stage is amplified in terms of spatial resolution and dynamic range, and a panorama with high quality and HDR which has a 4K+ resolution is output.

[00219] Firstly, the pixel corresponding to any position in the panorama may be expressed by the following formula (11):

$$S(\theta, \phi, 1) = f_c(z_c, \theta, \phi) \quad \text{formula (11).}$$

[00220] Here, the panorama with LDR 408 generated in the first stage as shown in FIG. 4 may be randomly clipped to obtain the image block $I' \in R^{128 \times 128 \times 3}$. FIG. 7A is a schematic diagram of an implementation framework for adjusting the resolution and dynamic range of a generated panorama provided by an embodiment of the application, in which 701 is the image block $I' \in R^{128 \times 128 \times 3}$, its corresponding pixel in the spherical structure is $S'(\theta', \phi', 1)$, and the potential code corresponding to any pixel in the image block I' is determined. Here, the encoder 702 shown in FIG. 7A may perform feature coding on the image block $I' \in R^{128 \times 128 \times 3}$ 701, and then the feature code of the pixel in 703 and its corresponding potential code 704 are obtained; then, the continuous feature Z_c , namely 705, corresponding to the pixel is determined based on the potential code 704. The process may be determined based on area interpolation corresponding to the potential code. As shown in 704 in FIG. 7A, the potential code is $\{z_{00}, z_{01}, z_{11}, z_{10}\}$, as shown in formula (12):

$$z_c = \sum_{i \in \{00, 01, 10, 11\}} \frac{A_i}{A} z_i, z_i \in Z_c \quad \text{formula (12);}$$

where z_i is the nearest potential code at (θ, ϕ) , and A_i is its corresponding area.

[00221] Then, based on the continuous feature Z_c , an SR-ITMO module is used to map each pixel in the image block $I' \in R^{128 \times 128 \times 3}$ to the image block with HR and HDR. The SR-ITMO may be composed of two different Multilayer Perceptrons (MLP), namely f_{sr} 706 and f_{itmo} 708 in FIG. 7A, which correspond to amplifying the spatial resolution and the dynamic range of the image in turn. In the practical application process, the MLP corresponding to f_{sr} has a four-layer structure, and the MLP corresponding to f_{itmo} has a two-layer structure. As shown in formula (13), the resolution of the image block $I' \in R^{128 \times 128 \times 3}$ is improved to obtain the first image block 707 shown in FIG. 7A:

$$I_{hr}^{ldr}, c_{hr} = f_{sr}(z_c) \quad \text{formula (13);}$$

where C_{hr} is the intermediate feature output by the second layer of f_{sr} . At the same time, as shown in formula (14), the dynamic range of the image block $I' \in R^{128 \times 128 \times 3}$ is

improved to obtain the second image block 709 shown in FIG. 7A:

$$I_{hr}^{hdr} = f_{itmo}(chr, \theta, \phi) \quad \text{formula (14).}$$

[00222] Here, in the process of training the SR-ITMO module, a corresponding training sample may be obtained as shown in FIG. 7B. FIG. 7B is a schematic diagram of a framework for generating a training sample of an image resolution and dynamic range adjustment part in a panorama generation network provided by an embodiment of the application. The training sample is a sample data pair $\hat{I}_{lr}^{ldr} - \hat{I}_{hr}^{hdr}$.

[00223] The sample data pair $\hat{I}_{lr}^{ldr} - \hat{I}_{hr}^{hdr}$ may be generated based on a panorama with HDR which includes a 4K-8K HR. The acquisition of the sample may include three steps: tone mapping, calibration and pixel sample acquisition.

[00224] At the first step, tone mapping is performed based on the panorama with HDR which has a 4K+ resolution to obtain the panorama with LDR. As shown in FIG. 7B, tone mapping is performed to a first panorama 711 to obtain a second panorama 716.

[00225] At the second step, brightness calibration is performed on the first panorama 711, that is, based on the two images I_1^{hdr} and I_2^{hdr} with different brightness, where $I_1^{hdr} = kI_2^{hdr}$, and $k \in R^+$, each panorama with HDR is degraded to $\hat{I}^{hdr}$ with the same brightness ratio, as shown in formula (15):

$$\hat{I}^{hdr} = \frac{\sum (M \square I^{ldr})}{\sum (M \square I^{hdr})} \cdot I^{hdr} \quad \text{formula (15);}$$

where $\square$ represents element multiplication, and M is the position to be calibrated which is determined according to $I^{ldr} \in R^{h \times w \times 3}$, as shown in formula (16):

$$M(i, j) = \begin{cases} 1, & \sum_{k=1}^3 I^{ldr}(i, j, k) < 3\sigma, \\ 0, & \text{otherwise,} \end{cases} \quad \text{formula (16);}$$

where $i=0,1,\dots,h-1$, $j=0,1,\dots,w-1$, and it is determined that $\sigma=0.83$ by giving experimental data.

[00226] At the third step, a pixel sample is obtained, that is, a first pixel sample 715 and a second pixel sample 718 shown in FIG. 7B as well as an input sample image 720 are obtained. The second panorama 716 is clipped to obtain the second panoramic image block 717, and then the second pixel sample 718 is obtained based on the second panoramic image block 717. In addition, down-sampling is performed to the second panoramic image block 717 to obtain the third panoramic image block 719, and then the third panoramic image block 719 is taken as the input sample in the process of training the SR-ITMO module.

[00227] Moreover, the first target panorama 713 obtained by performing brightness calibration by a brightness calibration unit 712 is simultaneously clipped to obtain the first panoramic image block 714, and then the first pixel sample 715 is obtained based on

the first panoramic image block 714.

[00228] In the second stage, the encoder E and the SR-ITMO may be taken as a whole. For example, in the process of performing end-to-end training to an image adjustment network to be trained, the sample image 720 is input into the image adjustment network to be trained, and two images, namely 707 and 709, may be obtained referring to FIG. 7A, the pixel parameter included in the first image block 707 is compared with the second pixel sample 718 in terms of the similarity, and the pixel parameter included in the second image block 709 is compared with the first pixel sample 715 in terms of the similarity, and then, based on the loss of supervising the similarity, the network of the encoder E and the SR-ITMO is adjusted, so that the loss output by the trained encoder E and SR-ITMO converges.

[00229] In the training process, the encoder E and the SR-ITMO may be trained based on the following formula (17):

$$L_{sr} = \frac{1}{n} \sum \left\| I_{hr}^{ldr} - \hat{I}_{hr}^{ldr} \right\|_1 \quad \text{formula (17);}$$

where n is the number of pixels sampled, representing the number of pixels sampled in the first image block 707 or the second image block 709, and I_{hr}^{ldr} is obtained by inputting the sample image to the network for prediction, that is, by referring to formula (13).

[00230] Moreover, the model learns to extend the dynamic range by minimizing the following scaled target variable, as shown in formula (18):

$$L_{itmo} = \frac{1}{n} \sum D^2 \left(I_{hr}^{hdr}, \hat{I}_{hr}^{hdr} \right) - \left[\frac{1}{n} \sum D \left(I_{hr}^{hdr}, \hat{I}_{hr}^{hdr} \right) \right]^2 \quad \text{formula (18);}$$

where $D(I_1, I_2) = \log(I_1) - \log(I_2)$ represents a distance parameter. The overall loss in the second stage may refer to formula (19):

$$L_{full} = L_{sr} + L_{itmo} \quad \text{formula (19).}$$

[00231] Based on this, the panorama generation network provided by the embodiment of the application may be divided into two parts. In the first part, a hierarchical framework is proposed to generate a text-driven panorama with LDR, which is discretely represented based on dual codebooks. In the second part, the SR-ITMO is proposed, which amplifies the spatial resolution and dynamic range of the results obtained from the first part.

[00232] Thus, in the embodiments of the application, firstly, the rough global feature of the subsequent constructed image can be determined according to the input scene description text; according to the rough global feature obtained by sampling, the local image feature that represents the local fine granularity of an image is sampled at the corresponding position of the panorama based on the spherical positional encoding, and then by performing feature decoding and feature synthesis on the acquired feature, the panorama that has a spherical structure and matches text semantics of the scene description text can be obtained. In this way, based on the input text information, it is

relatively more convenient to obtain the panorama that matches the text semantics of the text information. Finally, based on the super-resolution inverse tone mapping, a panorama with higher resolution and higher dynamic range which matches the text semantics of the text information is output.

- [00233] An embodiment of the application provides a panorama generation apparatus. FIG. 8A is a structural schematic diagram of a panorama generation apparatus provided by an embodiment of the application. As shown in FIG. 8A, the panorama generation apparatus 800 may include: a first acquisition module 801, a first sampling module 802, and a second sampling module 803.
- [00234] The first acquisition module 801 is configured to acquire a scene description text.
- [00235] The first sampling module 802 is configured to, based on the scene description text, perform feature sampling in a global feature book corresponding to a preset scene image set to obtain a global image feature.
- [00236] The second sampling module 803 is configured to, based on the global image feature and a spherical positional encoding corresponding to a preset panorama, perform feature sampling in a local feature book corresponding to the preset scene image set to obtain a target panorama corresponding to the scene description text.
- [00237] In some embodiments, the first sampling module 802 is further configured to: determine a text feature corresponding to the scene description text; and sample, in the global feature book, a feature matching the text feature on the scene semantics to obtain the global image feature.
- [00238] In some embodiments, the second sampling module 803 is further configured to: determine, in the local feature book, the local image feature corresponding to each positional code in the spherical positional encoding based on the global image feature; and perform feature decoding and feature synthesis on the local image feature corresponding to each positional code to obtain the target panorama.
- [00239] In some embodiments, the panorama generation apparatus 800 may further include: a second determination module, configured to determine a continuous feature corresponding to each pixel in the target panorama; a first adjustment module, configured to adjust the continuous feature corresponding to each pixel in terms of resolution and dynamic range to obtain an adjusted target panorama.
- [00240] In some embodiments, the second determination module is further configured to: perform region division on the target panorama to obtain the image block set; for each pixel in each image block in the image block set, perform feature coding on the pixel to obtain at least one potential feature surrounding the pixel; and obtain the continuous feature corresponding to the pixel based on each potential feature surrounding the pixel and the ratio between the area of each potential feature surrounding the pixel and the preset area. The preset area is the sum of the areas of potential codes surrounding the pixel.
- [00241] In some embodiments, the first adjustment module is further configured to perform up-sampling on the continuous feature corresponding to each pixel in terms of resolution to obtain an intermediate feature corresponding to each pixel and an intermediate panorama; and based on the positional code of each pixel in the target

panorama, perform up-sampling on the intermediate feature corresponding to each pixel in terms of dynamic range to obtain the adjusted target panorama.

- [00242] An embodiment of the application further provides an apparatus for training a panorama generation network. FIG. 8B is a structural schematic diagram of an apparatus for training a panorama generation network provided by an embodiment of the application. As shown in FIG. 8B, the apparatus 810 for training a panorama generation network may include: a second acquisition module 811, a first training module 812, a first determination module 813, and a second training module 814.
- [00243] The second acquisition module 811 is configured to acquire a sample text feature. The sample text feature includes a pseudo text feature and a text-like feature. The pseudo text feature is obtained by perturbing the text feature corresponding to the sample image. The text-like feature is the text feature corresponding to an image whose similarity with the sample image satisfies a preset condition.
- [00244] The first training module 812 is configured to, according to the sample text feature and a global feature book, train a global sampler to be trained in a panorama generation network to be trained to obtain a trained global sampler. The global feature book is obtained by performing feature coding on a preset scene image set based on a trained global encoder.
- [00245] The first determination module 813 is configured to determine, in the global feature book, global sample feature matching the sample text feature on scene semantics using the trained global sampler.
- [00246] The second training module 814 is configured to, according to the global sample feature, spherical positional encoding corresponding to a preset panorama and a local feature book, train a local sampler to be trained in the panorama generation network to be trained to obtain a trained local sampler. The local feature book is obtained by performing feature coding on the preset scene image set based on a trained local encoder.
- [00247] In some embodiments, the first training module 812 is further configured to: determine, in the global feature book, the initial global feature matching the sample text feature using the global sampler to be trained; perform feature decoding on the initial global feature to obtain the first reconstructed image; determine a first reconstruction loss of the cosine similarity between the first reconstructed image and the sample image; and based on the first reconstruction loss, adjust a network parameter of the global sampler to be trained, so that the first reconstruction loss output by the trained global sampler satisfies a convergence condition.
- [00248] In some embodiments, the first training module 812 is further configured to: determine the first feature in the first sequence according to the sample text feature using the global sampler to be trained, the first sequence representing the empty sequence associated with the index sequence of the global feature book; based on the i -th feature in the first sequence and the sample text feature, predicting an $(i+1)$ -th feature in the first sequence in the global feature book until the first feature sequence containing the same number of feature as the first sequence is obtained, i being a positive integer greater than or equal to 1; and determine the first feature sequence as the initial global feature.
- [00249] In some embodiments, the second training module 814 is further configured to determine, in the local feature book, the local sample feature corresponding to the preset

positional code in the spherical positional encoding according to the global sample feature using the local encoder to be trained; perform feature coding on the global sample feature to obtain the first reconstructed image block; determine a second reconstruction loss of the cosine similarity between the first reconstructed image block and the sample image block, the sample image block being the image region at the preset position in the sample image, and there being a mapping relationship between the preset position and the preset positional code; and based on the second reconstruction loss, adjust a network parameter of the local sampler to be trained, so that the second reconstruction loss output by the trained local sampler satisfies a convergence condition.

[00250] In some embodiments, the second training module 814 is further configured to: determine the first feature in the second sequence according to the global sample feature and the preset positional code using the local sampler to be trained, the second sequence representing the empty sequence associated with the index sequence of the local feature book; based on the i -th feature in the second sequence, the global sample feature and the preset positional code, predict the $(i+1)$ -th feature in the second sequence in the local feature book until the second feature sequence containing the same number of features as the second sequence is obtained, i being a positive integer greater than or equal to 1; and determine the second feature sequence as the local sample feature.

[00251] In some embodiments, the apparatus 810 for training a panorama generation network may further include: a third training module, configured to: perform feature coding on the image to be trained using the global encoder to be trained, so as to obtain the global continuous sample feature and the global discrete sample feature corresponding to the global continuous sample feature; perform feature decoding on the global discrete sample feature to obtain the second reconstructed image; determine a third reconstruction loss of the similarity between the second reconstructed image and the image to be trained, as well as a first adversarial loss; determine a first quantization loss between the global discrete sample feature and the global continuous sample feature; fuse the third reconstruction loss, the first adversarial loss and the first quantization loss to obtain a first overall loss; based on the first overall loss, adjust the network parameter of the global encoder to be trained, so that the loss of the feature output by the trained global encoder satisfies a convergence condition.

[00252] In some embodiments, the apparatus 810 for training a panorama generation network may further include: a third training module, configured to: perform feature coding on a preset region in the image to be trained using the local encoder to be trained, so as to obtain a local continuous sample feature and a local discrete sample feature corresponding to the local continuous sample feature; perform feature decoding on the local discrete sample feature to obtain the second reconstructed image block; determine a fourth reconstruction loss of the similarity between the second reconstructed image block and the preset region in the image to be trained, as well as a second adversarial loss; determine the second quantization loss between the local discrete sample feature and the local continuous sample feature; fuse the fourth reconstruction loss, the second adversarial loss and the second quantization loss to obtain the second overall loss; based on the second overall loss, adjust a network parameter of the local encoder to be trained, so that the loss of the feature output by the trained local encoder satisfies convergence conditions.

[00253] In some embodiments, the panorama generation network to be trained includes

an image adjustment sub-network to be trained. The apparatus 810 for training a panorama generation network may further include: a fifth training module, configured to: acquire a panoramic sample image block, as well as the first panoramic image block and the second panoramic image block that correspond to the panoramic sample image block and have different dynamic ranges; perform feature coding and tone adjustment on each pixel in the panoramic sample image block using the image adjustment sub-network to be trained, so as to obtain the third panoramic image block and the fourth panoramic image block; determine the first loss of the similarity between the first panoramic image block and the third panoramic image block is determined, and determine the second loss of the similarity between the second panoramic image block and the fourth panoramic image block; based on the first loss and the second loss, adjust a network parameter of the image adjustment sub-network to be trained, so that the loss of the image output by the trained image adjustment sub-network satisfies a convergence condition.

[00254] In some embodiments, the fifth training module is further configured to: acquire the panoramic sample image; perform tone mapping on the panoramic sample image and select any image block to obtain the first panoramic image block; perform down-sampling on the first panoramic image block to obtain the panoramic sample image block; and perform brightness calibration on the panoramic sample image, and select the second panoramic image block corresponding to the first panoramic image block.

[00255] In some embodiments, the fifth training module is further configured to: perform feature coding on each pixel in the panoramic sample image block using the image adjustment sub-network to be trained, so as to obtain the continuous sample feature corresponding to each pixel; perform up-sampling on the continuous sample feature corresponding to each pixel in terms of resolution, so as to obtain an intermediate sample feature corresponding to each pixel and the corresponding third panoramic image block; based on a spherical code corresponding to each pixel, perform up-sampling on the dynamic range of the intermediate sample feature corresponding to each pixel to obtain the fourth panoramic image block.

[00256] It is to be noted that the above descriptions about the apparatus embodiments are similar to descriptions about the method embodiments and beneficial effects similar to those of the method embodiments are achieved. Technical details undisclosed in the apparatus embodiments of the application may be understood with reference to the descriptions about the method embodiments of the application.

[00257] It is to be noted that, in the embodiments of the application, when being implemented in form of software function module and sold or used as an independent product, the panorama generation method or the method for training a panorama generation network may also be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the embodiments of the application substantially or parts making contributions to the conventional art may be embodied in form of software product, and the computer software product is stored in a storage medium, including multiple instructions configured to enable a computer device (which may be a terminal, a server, and the like) to execute all or part of the method in each embodiment of the application. The storage medium includes: various media capable of storing program codes such as a U disk, a mobile hard disk, a Read Only Memory (ROM), a magnetic disk or an optical disk. Therefore, the application is not limited to

any specific combination of hardware and software.

[00258] Correspondingly, the embodiments of the application also provide a computer program product, which may include a computer-executable instruction. After executed, the computer-executable instruction can implement the panorama generation method or the method for training a panorama generation network provided by the embodiments of the application.

[00259] Accordingly, the embodiments of the application provide a computer device. FIG. 9 illustrates a composition structure diagram of a computer device according to an embodiment of the application. As illustrated in FIG. 9, the computer device 900 may include: a processor 901, at least one communication bus 904, a communication interface 902, at least one external communication interface and a memory 903. The communication interface 902 is configured to implement connection communication between these components. The communication interface 902 may include a display screen, and the external communication interface may include a standard wired interface and a standard wireless interface. The processor 901 is configured to execute the program in the memory to implement the panorama generation method or the method for training a panorama generation network provided in the above embodiments.

[00260] Correspondingly, an embodiment of the application also provides a computer storage medium, in which a computer-executable instruction is stored. The computer-executable instruction can be executed by the processor to implement the panorama generation method or the method for training a panorama generation network provided in the above embodiments.

[00261] The above descriptions about the embodiments of the panorama generation apparatus, the apparatus for training a panorama generation network, the computer device and the storage medium are similar to the descriptions about the method embodiments, and technical descriptions and beneficial effects are similar to those of the corresponding method embodiments. Due to the space limitation, references can be made to the records in the method embodiments. Technical details undisclosed in the embodiments of the panorama generation apparatus, the apparatus for training a panorama generation network, the computer device and the storage medium of the application may be understood with reference to the descriptions about the method embodiments of the application.

[00262] It is to be understood that "one embodiment" and "an embodiment" mentioned in the whole specification mean that specific features, structures or characteristics related to the embodiment is included in at least one embodiment of the application. Therefore, "in one embodiment" or "in an embodiment" appearing at any place of the whole specification does not always refer to the same embodiment. In addition, these specific features, structures or characteristics may be combined in one or more embodiments in any proper manner. It should be understood that, in various embodiments of the application, a magnitude of a sequence number of each process does not mean an execution sequence and the execution sequence of each process should be determined by its function and an internal logic and should not form any limit to an implementation process of the embodiments of the application. The sequence numbers of the embodiments of the application are just for describing, instead of representing superiority-inferiority of the embodiments. It is to be noted that terms "include" and "contain" or any other variant thereof is intended to cover nonexclusive inclusions herein,

so that a process, method, object or device including a series of components not only includes those components but also includes other components which are not clearly listed or further includes components intrinsic to the process, the method, the object or the device. Under the condition of no more limitations, an component defined by the statement "including a/an....." does not exclude existence of the same other components in a process, method, object or device including the component.

[00263] In several embodiments of the application, it should be understood that the disclosed device and method may be implemented in another manner. The device embodiment described above is only schematic, and for example, division of the units is only logic function division, and other division manners may be adopted during practical implementation. For example, multiple units or components may be combined or integrated into another system, or some characteristics may be neglected or not executed. In addition, coupling or direct coupling or communication connection between each displayed or discussed component may be indirect coupling or communication connection, implemented through some interfaces, of the device or the units, and may be electrical and mechanical or adopt other forms.

[00264] The units described as separate parts may or may not be physically separated, and parts displayed as units may or may not be physical units, and namely may be located in the same place, or may also be distributed to multiple network units. Part of all of the units may be selected according to a practical requirement to achieve the purposes of the solutions of the embodiments.

[00265] In addition, each functional unit in each embodiment of the application may be integrated into a processing unit, each unit may also serve as an independent unit and two or more than two units may also be integrated into a unit. The integrated unit may be implemented in a hardware form and may also be implemented in form of hardware and software functional unit. Those of ordinary skill in the art should know that all or part of the steps of the method embodiment may be implemented by related hardware instructed through a program, the program may be stored in a computer-readable storage medium, and the program is executed to execute the steps of the method embodiment. The storage medium includes: various media capable of storing program codes such as a mobile storage device, an ROM, a magnetic disk or a compact disc.

[00266] Or, when being implemented in form of software function module and sold or used as an independent product, the integrated unit in the embodiments of the application may also be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the embodiments of the application substantially or parts making contributions to the conventional art may be embodied in form of software product, and the computer software product is stored in a storage medium, including a plurality of instructions configured to enable a computer device (which may be a personal computer, a server, a network device or the like) to execute all or part of the method in each embodiment of the application. The storage medium includes: various media capable of storing program codes such as a mobile hard disk, a ROM, a magnetic disk or a compact disc. The above is only the specific implementation mode of the embodiments of the application and not intended to limit the scope of protection of the embodiments of the application. Any variations or replacements apparent to those skilled in the art within the technical scope disclosed by the embodiments of the application shall fall within the scope of protection of the embodiments of the application. Therefore,

the scope of protection of the application shall be subject to the scope of protection of the claims.

CLAIMS

1. A panorama generation method, comprising:

- acquiring a scene description text;
- performing, based on the scene description text, feature sampling in a global feature book corresponding to a preset scene image set to obtain a global image feature; and
- performing, based on the global image feature and spherical positional encoding corresponding to a preset panorama, feature sampling in a local feature book corresponding to the preset scene image set to obtain a target panorama corresponding to the scene description text.

2. The method of claim 1, wherein performing, based on the scene description text, feature sampling in the global feature book corresponding to the preset scene image set to obtain the global image feature comprises:

- determining a text feature corresponding to the scene description text; and
- sampling, in the global feature book, a feature matching the text feature on scene semantics to obtain the global image feature.

3. The method of claim 1 or 2, wherein performing, based on the global image feature and the spherical positional encoding corresponding to the preset panorama, feature sampling in the local feature book corresponding to the preset scene image set to obtain the target panorama corresponding to the scene description text comprises:

- determining, based on the global image feature, in the local feature book, a local image feature corresponding to each positional code in the spherical positional encoding; and
- performing feature decoding and feature synthesis on the local image feature corresponding to each positional code to obtain the target panorama.

4. The method of any one of claims 1 to 3, further comprising:

- determining a continuous feature corresponding to each pixel in the target panorama; and
- adjusting the continuous feature corresponding to each pixel in terms of resolution and dynamic range to obtain an adjusted target panorama.

5. The method of claim 4, wherein determining the continuous feature corresponding to each pixel in the target panorama comprises:

- performing region division on the target panorama to obtain an image block set;
- performing, for each pixel in each image block in the image block set, feature coding on the pixel to obtain at least one potential feature surrounding the pixel;

obtaining the continuous feature corresponding to the pixel based on each potential feature surrounding the pixel and a ratio between an area of each potential feature surrounding the pixel and a preset area, wherein the preset area is a sum of an area of at least one potential code surrounding the pixel.

6. The method of claim 4 or 5, wherein adjusting the continuous feature corresponding to each pixel in terms of resolution and dynamic range to obtain the adjusted target panorama comprises:

performing up-sampling on the continuous feature corresponding to each pixel in terms of resolution to obtain an intermediate feature corresponding to each pixel and an intermediate panorama;

performing, based on the positional code of each pixel in the target panorama, up-sampling on the intermediate feature corresponding to each pixel in terms of dynamic range to obtain the adjusted target panorama.

7. A method for training a panorama generation network, comprising:

acquiring a sample text feature, wherein the sample text feature comprises a pseudo text feature and a text-like feature; the pseudo text feature is obtained by perturbing a text feature corresponding to a sample image; the text-like feature is a text feature corresponding to an image whose similarity with the sample image satisfies a preset condition;

training, according to the sample text feature and a global feature book, a global sampler to be trained in a panorama generation network to be trained to obtain a trained global sampler, wherein the global feature book is obtained by performing feature coding on a preset scene image set based on a trained global encoder;

determining, in the global feature book, a global sample feature matching the sample text feature on scene semantics using the trained global sampler;

training a local sampler to be trained in the panorama generation network to be trained according to the global sample feature, a spherical positional encoding corresponding to a preset panorama and a local feature book to obtain a trained local sampler, wherein the local feature book is obtained by performing feature coding on the preset scene image set based on a trained local encoder.

8. The method of claim 7, wherein training, according to the sample text feature and the global feature book, the global sampler to be trained in the panorama generation network to be trained to obtain the trained global sampler comprises:

determining, in the global feature book, an initial global feature matching the sample text feature using the global sampler to be trained;

performing feature decoding on the initial global feature to obtain a first reconstructed image;

determining a first reconstruction loss of a cosine similarity between the first reconstructed image and the sample image; and

adjusting, based on the first reconstruction loss, a network parameter of the global sampler to be trained, so that the first reconstruction loss output by the trained global sampler satisfies a convergence condition.

9. The method of claim 8, wherein determining, in the global feature book, the initial global feature matching the sample text feature using the global sampler to be trained comprises:

determining a first feature in a first sequence according to the sample text feature using the global sampler to be trained, wherein the first sequence represents an empty sequence associated with an index sequence of the global feature book;

predicting, in the global feature book, an (i+1)-th feature in the first sequence based on an i-th feature in the first sequence and the sample text feature, until a first feature sequence containing the same number of features as the first sequence is obtained, wherein i is a positive integer greater than or equal to 1; and

determining the first feature sequence as the initial global feature.

10. The method of any one of claims 7 to 9, wherein training the local sampler to be trained in the panorama generation network to be trained according to the global sample feature, the spherical positional encoding corresponding to the preset panorama and the local feature book to obtain the trained local sampler comprises:

determining, in the local feature book, a local sample feature corresponding to a preset positional code in the spherical positional encoding according to the global sample feature using the local sampler to be trained;

performing feature decoding on the local sample feature to obtain a first reconstructed image block;

determining a second reconstruction loss of a cosine similarity between the first reconstructed image block and a sample image block, wherein the sample image block is an image region at a preset position in the sample image; there is a mapping relationship between the preset position and the preset positional code; and

adjusting, based on the second reconstruction loss, a network parameter of the local sampler to be trained, so that the second reconstruction loss output by the trained local sampler satisfies a convergence condition.

11. The method of claim 10, wherein determining, in the local feature book, the local sample feature corresponding to the preset positional code in the spherical positional encoding according to the global sample feature using the local sampler to be trained comprises:

determining a first feature in a second sequence according to the global sample feature and the preset positional code using the local sampler to be trained, wherein the second sequence represents an empty sequence associated with an index sequence of the local feature book;

predicting, in the local feature book, an (i+1)-th feature in the second sequence based on an i-th feature in the second sequence, the global sample feature and the preset positional

code, until a second feature sequence containing the same number of features as the second sequence is obtained, wherein i is a positive integer greater than or equal to 1; and

determining the second feature sequence as the local sample feature.

12. The method of any one of claims 7 to 11, wherein a training process of the trained global encoder comprises:

performing feature coding on an image to be trained using the global encoder to be trained, to obtain a global continuous sample feature and a global discrete sample feature corresponding to the global continuous sample feature;

performing feature decoding on the global discrete sample feature to obtain a second reconstructed image;

determining a third reconstruction loss of a similarity between the second reconstructed image and the image to be trained, as well as a first adversarial loss;

determining a first quantization loss between the global discrete sample feature and the global continuous sample feature;

fusing the third reconstruction loss, the first adversarial loss and the first quantization loss to obtain a first overall loss; and

adjusting, based on the first overall loss, a network parameter of the global encoder to be trained, so that a loss of a feature output by the trained global encoder satisfies a convergence condition.

13. The method of any one of claims 7 to 12, wherein a training process of the trained local encoder comprises:

performing feature coding on a preset region in the image to be trained using the local encoder to be trained, to obtain a local continuous sample feature and a local discrete sample feature corresponding to the local continuous sample feature;

performing feature decoding on the local discrete sample feature to obtain a second reconstructed image block;

determining a fourth reconstruction loss of a similarity between the second reconstructed image block and the preset region in the image to be trained, as well as a second adversarial loss;

determining a second quantization loss between the local discrete sample feature and the local continuous sample feature;

fusing the fourth reconstruction loss, the second adversarial loss and the second quantization loss to obtain a second overall loss; and

adjusting, based on the second overall loss, a network parameter of the local encoder to be trained, so that a loss of a feature output by the trained local encoder satisfies the a convergence condition.

14. The method of any one of claims 7 to 13, wherein the panorama generation network to be

trained comprises an image adjustment sub-network to be trained; the method further comprises:

- acquiring a panoramic sample image block, as well as a first panoramic image block and a second panoramic image block that correspond to the panoramic sample image block and have different dynamic ranges;

- performing feature coding and tone adjustment on each pixel in the panoramic sample image block using the image adjustment sub-network to be trained, to obtain a third panoramic image block and a fourth panoramic image block;

- determining a first loss of a similarity between the first panoramic image block and the third panoramic image block, and determining a second loss of a similarity between the second panoramic image block and the fourth panoramic image block;

- adjusting, based on the first loss and the second loss, a network parameter of the image adjustment sub-network to be trained, so that a loss of an image output by a trained image adjustment sub-network satisfies a convergence condition.

15. The method of claim 14, wherein acquiring the panoramic sample image block, as well as the first panoramic image block and the second panoramic image block that correspond to the panoramic sample image block and have different dynamic ranges comprises:

- acquiring a panoramic sample image;

- performing tone mapping on the panoramic sample image and selecting any image block to obtain the first panoramic image block;

- performing down-sampling on the first panoramic image block to obtain the panoramic sample image block;

- performing brightness calibration on the panoramic sample image, and selecting the second panoramic image block corresponding to the first panoramic image block.

16. The method of claim 14 or 15, wherein performing feature coding and tone adjustment on each pixel in the panoramic sample image block using the image adjustment sub-network to be trained, to obtain the third panoramic image block and the fourth panoramic image block comprises:

- performing feature coding on each pixel in the panoramic sample image block using the image adjustment sub-network to be trained, to obtain a continuous sample feature corresponding to each pixel;

- performing up-sampling on the continuous sample feature corresponding to each pixel in terms of resolution, to obtain an intermediate sample feature corresponding to each pixel and the corresponding third panoramic image block;

- performing, based on a spherical code corresponding to each pixel, up-sampling on the intermediate sample feature corresponding to each pixel in terms of dynamic range to obtain the fourth panoramic image block.

17. A computer device, comprising a memory and a processor, wherein a computer-executable instruction is stored in the memory; and when executing the computer-executable instruction in the memory, the processor is configured to:

acquire a scene description text;

perform, based on the scene description text, feature sampling in a global feature book corresponding to a preset scene image set to obtain a global image feature; and

perform, based on the global image feature and spherical positional encoding corresponding to a preset panorama, feature sampling in a local feature book corresponding to the preset scene image set to obtain a target panorama corresponding to the scene description text.

18. A computer device, comprising a memory and a processor, wherein a computer-executable instruction is stored in the memory; and when executing the computer-executable instruction in the memory, the processor is configured to:

acquire a sample text feature, wherein the sample text feature comprises a pseudo text feature and a text-like feature; the pseudo text feature is obtained by perturbing a text feature corresponding to a sample image; the text-like feature is a text feature corresponding to an image whose similarity with the sample image satisfies a preset condition;

train, according to the sample text feature and a global feature book, a global sampler to be trained in a panorama generation network to be trained to obtain a trained global sampler, wherein the global feature book is obtained by performing feature coding on a preset scene image set based on a trained global encoder;

determine, in the global feature book, a global sample feature matching the sample text feature on scene semantics using the trained global sampler;

train a local sampler to be trained in the panorama generation network to be trained according to the global sample feature, a spherical positional encoding corresponding to a preset panorama and a local feature book to obtain a trained local sampler, wherein the local feature book is obtained by performing feature coding on the preset scene image set based on a trained local encoder.

19. A computer storage medium, having a computer-executable instruction stored thereon, wherein when executed, the computer-executable instruction implements the panorama generation method of any one of claims 1 to 6, or, when executed, the computer-executable instruction implements the method for training a panorama generation network of any one of claims 7 to 16.

20. A computer program product, comprising a non-transitory computer-readable storage medium storing a computer program, wherein the computer program, when read and executed by a computer, implements the panorama generation method of any one of claims 1 to 6, or, wherein the computer program, when executed, the computer-executable instruction implements the method for training a panorama generation network of any one of claims 7 to 16.