

IMAGE PROCESSING METHOD, APPARATUS, READABLE STORAGE MEDIUM, AND ELECTRONIC DEVICE

TECHNICAL FIELD

[0001] The present disclosure relates to computer vision technology and deep learning technology, and in particular, to an image processing method, an apparatus, a readable storage medium, and an electronic device.

BACKGROUND

[0002] Time of Flight (TOF) data and image data are commonly used data in the field of machine vision. Both the TOF data and image data can be combined and applied to a multi-modal perception system to capture three-dimensional (3D) geometric structures and two-dimensional (2D) visual features of objects, so as to overcome limitations of a single sensor in complex environments.

[0003] In the related art, two separate neural network models are generally adopted to process the TOF data and the image data, so as to obtain depth maps and denoised images. Such a processing method increases hardware computational overhead and memory footprint, since it needs to adopt two separate neural networks with numerous parameters and high structural complexity. In addition, separate processing of the two types of data results in low computational efficiency and poor real-time performance.

SUMMARY

[0004] To solve the above problems, the present disclosure provides an image processing method, an apparatus, a readable storage medium, and an electronic device, in which feature

extraction can be performed on a scene image and a photon statistical graph by a shared encoding network, decoded data can more accurately represent actual scene features, effectively reduce parameters of a model, and improve processing efficiency of the model.

[0005] A first aspect of embodiments of the present disclosure provides an image processing method, which includes: obtaining a photon statistical graph and a scene image corresponding to a target scene, where, a pixel corresponds to a photon statistical data sequence in the photon statistical graph, photon statistical data in the photon statistical data sequence represents a number of reflected photons received from a target position at a corresponding moment, and the target position is a position corresponding to the pixel in the target scene; encoding the scene image and the photon statistical graph respectively by a shared encoding network included in a pre-trained image processing model to obtain image features and photon statistical features; decoding the photon statistical features by a depth decoder included in the image processing model, and generating a depth map corresponding to the target scene using the decoded photon statistical features; and decoding the image features by an image decoder included in the image processing model to obtain a denoised image.

[0006] A second aspect of embodiments of the present disclosure provides an image processing apparatus, which includes an obtaining model, an encoding module, a first decoding module, and a second decoding module. The obtaining model is configured to obtain a photon statistical graph and a scene image corresponding to a target scene, where, a pixel corresponds to a photon statistical data sequence in the photon statistical graph, photon statistical data in the photon statistical data sequence represents a number of reflected photons received from a target position at a corresponding moment, and the target position is a position corresponding to the pixel in the target scene. The encoding module is configured to encode the scene image and the photon

statistical graph, respectively, by a shared encoding network included in a pre-trained image processing model to obtain image features and photon statistical features. The first decoding module is configured to decode the photon statistical features by a depth decoder included in the image processing model, and generate a depth map corresponding to the target scene using the decoded photon statistical features. The second decoding module is configured to decode the image features by an image decoder included in the image processing model to obtain a denoised image.

[0007] A third aspect of embodiments of the present disclosure provides a computer-readable storage medium, which includes a computer program stored thereon, where the computer program, when executed, is configured to implement the image processing method according to any one of the embodiments of the present disclosure.

[0008] A fourth aspect of embodiments of the present disclosure provides an electronic device, which includes: a memory, configured to store instructions; and a processor, configured to read the instructions from the memory, and execute the instructions to implement the image processing method according to any one of the embodiments of the present disclosure.

[0009] A fifth aspect of embodiments of the present disclosure provides a computer program product, which includes computer program instructions, and when the computer program instructions are executed by a process, the processor is configured to perform the image processing method according to any one of the embodiments of the present disclosure.

[0010] Based on the embodiments of the present disclosure, the photon statistical graph and the scene image corresponding to the target scene are obtained, the scene image and the photon statistical graph are encoded respectively by the shared encoding network included in the pre-trained image processing model to obtain image features and photon statistical features, the

photon statistical features are decoded by the depth decoder, and the depth map corresponding to the target scene is generated using the decoded photon statistical features, and the image features are decoded by the image decoder to obtain the denoised image. Since the shared encoding network is used for encoding the scene image and the photon statistical graph, the shared encoding network can extract two-dimensional features and three-dimensional features contained in the data input thereto. The features output by the shared encoding network can reflect the complementarity between the two-dimensional features and the three-dimensional features. By such complementarity, the photon statistical features and the image features can be decoded more precisely, thereby improving the accuracy of the depth map decoded by the model and the quality of the denoised image decoded by the model. In addition, using the shared encoding network can avoid encoding the scene image and the photon statistical graph with separate encoding networks, thereby reducing parameter redundancy of the image processing model and improving the processing efficiency of the model.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] FIG. 1 is a schematic flowchart of an image processing method according to an exemplary embodiment of the present disclosure.

[0012] FIG. 2 is a schematic structural diagram of an image processing model according to an exemplary embodiment of the present disclosure.

[0013] FIG. 3 is a schematic flowchart of an image processing method according to another exemplary embodiment of the present disclosure.

[0014] FIG. 4 is a schematic structural diagram of an image processing model according to another exemplary embodiment of the present disclosure.

[0015] FIG. 5 is a schematic flowchart of an image processing method according to still another

exemplary embodiment of the present disclosure.

[0016] FIG. 6 is a schematic flowchart of an image processing method according to still another exemplary embodiment of the present disclosure.

[0017] FIG. 7 is a schematic flowchart of a training image processing model according to an exemplary embodiment of the present disclosure.

[0018] FIG. 8 is a schematic flowchart of a training image processing model according to another exemplary embodiment of the present disclosure.

[0019] FIG. 9 is a schematic flowchart of a training image processing model according to still another exemplary embodiment of the present disclosure.

[0020] FIG. 10 is a schematic flowchart of a training image processing model according to still another exemplary embodiment of the present disclosure.

[0021] FIG. 11 is a structural diagram of an image processing apparatus according to an exemplary embodiment of the present disclosure.

[0022] FIG. 12 is a structural diagram of an image processing apparatus according to another exemplary embodiment of the present disclosure.

[0023] FIG. 13 is a structural diagram of an electronic device according to an exemplary embodiment of the present disclosure.

DETAILED DESCRIPTION

[0024] Exemplary embodiments of the present disclosure are described below in detail with reference to the accompanying drawings to illustrate the present disclosure. Apparently, the described embodiments are merely a part, rather than all of the embodiments of the present disclosure. It should be understood that the present disclosure is not limited by the exemplary embodiments described herein.

[0025] It can be noted that, unless otherwise specifically stated, relative arrangements of components and steps, numerical expressions, and values described in these embodiments shall not limit the scope of the present disclosure.

[0026] Overview

[0027] In the related art, TOF data processing and image processing are usually implemented by separate neural networks, which present the following problems:

[0028] Waste of computing resources: the separate networks each require forward inference, increasing computational overhead and memory footprint.

[0029] Insufficient utilization of cross-modal information: TOF data and image data contain complementary information, that is, both TOF data and image data contain 2D and 3D features. Separate networks are unable to simultaneously extract both the 2D and 3D features from either TOF data or image data, making it impossible to fully exploit the complementarity between the 2D and 3D features, resulting in insufficient model processing accuracy.

[0030] Poor real-time performance: it is difficult to meet real-time requirements when the two types of data are processed separately on resource-constrained edge devices.

[0031] Parameter redundancy: the two separate networks have a large number of redundant parameters and computational structures, which also increases computational overhead and memory footprint.

[0032] To solve the above problems, the embodiments of the present disclosure use a shared encoding network to encode a scene image and a photon statistical graph; the shared encoding network can extract 2D features and 3D features contained in the data input thereto. The features output by the shared encoding network can reflect the complementarity between the 2D and 3D features. By such complementarity, photon statistical features and image features can be decoded more precisely, thereby improving the accuracy of a depth map decoded by an image processing

model and the quality of a denoised image decoded by the model. In addition, using a shared encoding network can avoid encoding the scene image and the photon statistical graph with separate encoding networks, thereby reducing parameter redundancy of the model and improving the processing efficiency of the model.

[0033] Exemplary Methods

[0034] FIG. 1 is a schematic flowchart of an image processing method according to an exemplary embodiment of the present disclosure. The embodiment can be applied to various types of electronic devices, such as vehicle controllers, personal computers, backend servers, etc. As shown in FIG. 1, the method includes the following steps 101 to 103.

[0035] In step 101, a photon statistical graph and a scene image corresponding to a target scene are obtained.

[0036] The target scene can be any type of scene, such as a scene of a vehicle driving, etc. The scene image may be a 2D image captured by a camera in the target scene. The photon statistical graph may be acquired by, for example, a dToF (direct Time-of-Flight) laser radar system. The above photon statistical graph may reflect the time of flight of photons corresponding to each pixel in the 2D image to which the target scene is mapped. The time of flight of a photon is the time taken for the photon to be emitted from a laser emitter, reflected by an object in the target scene, and received back by a photoelectric sensor.

[0037] The dToF laser radar system operates as follows. The laser emitter emits an extremely short light pulse. A Single Photon Avalanche Diode (SPAD) sensor detects photons reflected from various objects in the scene. For each detected photon, a Time to Digital Converter (TDC) precisely records the elapsed time from its emission to reception. The scene is mapped to the 2D image, and in that 2D image, the number of received photons is counted for the pixel corresponding to each scanned position, thereby obtaining the photon statistical graph.

[0038] For any pixel, the pixel corresponds to a photon statistical data sequence in the photon statistical graph. Photon statistical data in the photon statistical data sequence represents the number of reflected photons received from a target position at a corresponding moment. The target position is a position corresponding to the pixel in the target scene.

[0039] Generally, in a photon statistical graph to be denoised, for any pixel, the data sequence corresponding to the pixel may form a time-photon count histogram. For example, a histogram corresponding to one pixel has 512 time bins, covering a measurement range of 0–100 nanoseconds (corresponding to a distance of 0–15 meters). After repeatedly accumulating statistics of detected photon counts, the obtained data in the histogram are as follows:

[0040] Time bins: [1, 2, 3, 4, 5, ..., 510, 511, 512];

[0041] Photon counts: [2, 5, 21, 105, 203, 450, 301, 98, ..., 1, 0, 3].

[0042] A peak value of the histogram corresponds to the time of flight of the photons at the position corresponding to the pixel.

[0043] Additionally, in some optional implementations, the step 101 may also be executed by a Neural Processing Unit (NPU), or alternatively, by a processor invoking corresponding instructions stored in a memory, or alternatively, by an obtaining module 1101 running on a processor.

[0044] In step 102, the scene image and the photon statistical graph are encoded, respectively, by a shared encoding network included in a pre-trained image processing model to obtain image features and photon statistical features.

[0045] In the embodiment, the image processing model may process multi-modal image data, that is, the scene image and the photon statistical graph may be simultaneously input to and processed by the image processing model. In some implementations, the image processing model may be a neural network model with a Unet structure, in which the input image data is encoded

and then decoded to obtain processed image data.

[0046] As shown in FIG. 2, the image processing model may include the shared encoding network, a depth decoder, and an image decoder. The parameters of the image processing model may be obtained by pre-training. The shared encoding network may adopt the same parameters for encoding the scene image and the photon statistical graph, respectively. The shared encoding network may include at least one convolutional layer, at least one pooling layer, and the like, which performs layer-wise feature extraction on the input scene image and photon statistical graph and outputs the image features and the photon statistical features.

[0047] Additionally, in some optional implementations, the step 102 may also be executed by an NPU, or alternatively, by a processor invoking corresponding instructions stored in a memory, or alternatively, by an encoding module 1102 running on a processor.

[0048] In step 103, the photon statistical features are decoded by the depth decoder included in the image processing model, and a depth map corresponding to the target scene is generated using the decoded photon statistical features.

[0049] The depth decoder may include at least one convolutional layer, an upsampling layer, and the like. The depth decoder performs layer-wise decoding on the photon statistical features. The resolution of the decoded photon statistical features output by the last layer of the network may be restored to be the same as that of the input photon statistical graph. Then, the data corresponding to each pixel in the decoded photon statistical features may be mapped to a depth value to obtain the depth map. In the depth map, each pixel corresponds to a respective depth value, which represents a distance between a measured target and a sensor.

[0050] Additionally, in some optional implementations, step 103 may also be executed by an NPU, or alternatively, by a processor invoking corresponding instructions stored in a memory, or alternatively, by a first decoding module 1103 running on a processor.

[0051] In step 104, the image features are decoded by the image decoder included in the image processing model to obtain a denoised image.

[0052] The image decoder may include at least one convolutional layer, an upsampling layer, and the like. The image decoder performs layer-wise decoding on the image features. The resolution of the decoded image features output by the last layer of the network may be restored to be the same as that of the input scene image. Then, image reconstruction may be performed on each pixel contained in the decoded image features to obtain the denoised image.

[0053] Additionally, in some optional implementations, step 104 may also be executed by an NPU, or alternatively, by a processor invoking corresponding instructions stored in a memory, or alternatively, by a second decoding module 1104 running on a processor.

[0054] The image processing method provided by the embodiment of the present disclosure obtains the photon statistical graph and the scene image corresponding to the target scene, encodes the scene image and the photon statistical graph respectively by the shared encoding network included in the pre-trained image processing model to obtain the image features and the photon statistical features, decodes the photon statistical features by the depth decoder and generates the depth map corresponding to the target scene using the decoded photon statistical features, and decodes the image features by the image decoder to obtain the denoised image. Using the shared encoding network to encode the scene image and the photon statistical graph, the shared encoding network can extract 2D features and 3D features contained in the data input to the shared encoding network. The features output by the shared encoding network can reflect the complementarity between the 2D and 3D features. By such complementarity, the photon statistical features and the image features can be decoded more precisely, thereby improving the accuracy of the depth map decoded by the model and the quality of the denoised image decoded by the model. In addition, using the shared encoding network can avoid encoding the scene image and the photon statistical

graph with separate encoding networks, thereby reducing parameter redundancy of the image processing model and improving the processing efficiency of the model.

[0055] In some alternative implementations, as shown in FIG. 3, the step 102 may include steps 1021 to 1023.

[0056] In step 1021, the photon statistical graph is projected by a projection unit included in the shared encoding network to obtain projection data.

[0057] In the embodiment, as shown in FIG. 4, the shared encoding network includes the projection unit, an image pre-encoding unit, and a shared encoder. The projection unit is configured to perform dimension transformation on the photon statistical graph, so that dimensions of the transformed projection data are the same as dimensions of the image pre-encoding data. For example, an initial tensor size of the photon statistical graph is $96 \times 240 \times 256$, and after projection, the obtained projection data has a tensor size of $96 \times 240 \times 128$. An equation adopted by the projection unit may be an equation (1) shown below:

$$F_{tof}^{aligned} = ReLU(W_{proj} * X_{tof} + b_{proj}) \quad (1)$$

[0058] where W_{proj} and b_{proj} are parameters of the projection unit, and $ReLU$ is an activation function.

[0059] In step 1022, the scene image is pre-encoded by the image pre-encoding unit included in the shared encoding network to obtain image pre-encoding data.

[0060] The projection data and the image pre-encoding data are identical in dimension. The image pre-encoding unit may encode the input scene image, and the encoding procedure may be implemented by means of multiple convolutions, pooling, and the like.

[0061] For example, the size of the input scene image is $1080 \times 1920 \times 3$; Convolutional layer 1 structure: 3×3 convolution kernel, 64 channels, ReLU activation function; Convolutional layer 2

structure: 3×3 convolution kernel, 64 channels, ReLU activation function; Max pooling structure: 2×2 pooling kernel, the output tensor size of $540 \times 960 \times 64$; Convolutional layer 3 structure: 3×3 convolution kernel, 128 channels, ReLU activation function; Convolutional layer 4 structure: 3×3 convolution kernel, 128 channels, ReLU activation function; Adaptive pooling: the output tensor size of $96 \times 240 \times 128$. The dimensions of the output image pre-encoding data are the same as the dimensions of the projection data in the above example.

[0062] In step 1023, the projection data and the image pre-encoding data are encoded respectively by the shared encoder included in the shared encoding network to obtain the photon statistical features and the image features.

[0063] The shared encoder may include multiple stacked shared sub-networks. The input projection data and image pre-encoding data are encoded via the multiple shared sub-networks, respectively, to obtain downsampled photon statistical features and image features.

[0064] For example, in some implementations, the layer configuration of the shared encoder is as follows:

[0065] Layer 1: 128 input channels, 128 output channels, downsampled to 48×120 ;

[0066] Layer 2: 128 input channels, 256 output channels, downsampled to 24×60 ;

[0067] Layer 3: 256 input channels, 512 output channels, downsampled to 12×30 ;

[0068] Layer 4: 512 input channels, 1024 output channels, downsampled to 6×15 .

[0069] Each layer of the shared sub-network may include at least one convolutional layer and a downsampling layer. Following each convolution, the obtained data is computed by the ReLU activation function and then input into the subsequent layer for computation. Downsampling may be carried out to compute using the 2×2 pooling kernel.

[0070] In some implementations, for each layer of shared sub-network, two modalities of input data may be combined first for encoding the photon statistical features and the image features by

the layer of shared sub-network, and then the two modalities of encoded data are separated, and the encoded data after separation are input to the next layer of shared sub-network. The encoding is repeated following the same steps above until the last layer of shared sub-network outputs the photon statistical features and the image features.

[0071] In addition, in some optional implementations, the steps 1021 to 1023 may be executed respectively by a first transformation unit 11021, a second transformation unit 11022, and an encoding unit 11023 running on a processor.

[0072] In the embodiment, the shared encoding network includes the projection unit, the image pre-encoding unit, and the shared encoder, which achieves that, before using the shared encoder to encode the data of the two modalities, the scene image and the photon statistical graph are transformed to the same dimensions, and preliminary features are extracted from the scene image at the same time. This allows the shared encoder to subsequently encode the multi-modal data more accurately, so as to contribute to improving the accuracy of subsequent decoding.

[0073] In some alternative implementations, step 1023 may include: encoding the projection data by the shared encoder to obtain the photon statistical features containing two-dimensional planar features and three-dimensional depth features; and encoding the image pre-encoding data by the shared encoder to obtain the image features containing two-dimensional planar features and three-dimensional depth features.

[0074] In the embodiment, the shared encoder is obtained by training in advance based on a sample photon statistical graph and a sample scene image, where the sample photon statistical graph contains more 3D spatial features and the sample scene image contains more 2D planar features, thus, the shared encoder, after supervised learning with both 2D planar annotation information and 3D spatial annotation information, is capable of simultaneously extracting 2D

features and 3D features from input data. Accordingly, the obtained photon statistical features and image features may respectively reflect the complementarity between the 2D features and the 3D features, and may more accurately represent positions of various objects in an actual scene, thereby improving the accuracy of subsequent decoding.

[0075] In some alternative implementations, the shared encoder includes at least two layers of shared sub-networks arranged in sequence. Correspondingly, the step 1023 may be implemented in the following ways.

[0076] For any one of the at least two layers of shared sub-networks, data input to the layer of shared sub-network is encoded and downsampled using shared parameters of the layer of shared sub-network, to obtain the image features and the photon statistical features after downsampling, and image features and photon statistical features, which are after encoding and before downsampling, are backed up.

[0077] The image features and photon statistical features output by each layer of shared sub-network may be input to the next layer of shared sub-network, and the image features and photon statistical features output by the last layer of shared sub-network are respectively input to the image decoder and the depth decoder. The backup image features and photon statistical features are used for skip connection computation with corresponding features during decoding (in which skip connections are between decoders and encoders of a U-Net).

[0078] As an example, each layer of shared sub-network may include two convolutional layers. After a first convolution, output features are calculated by an activation function, and then a second convolution is performed; after that, another calculation is performed by the activation function to obtain the image features and photon statistical features output by the layer of shared sub-network, which are expressed as $A_{\text{img}}^{(1)}$ and $A_{\text{tof}}^{(1)}$; where the superscript “1” represents the current number

of the layer of shared sub-network. The $A_{\text{img}}^{(l)}$ and $A_{\text{tof}}^{(l)}$ are backed up as $S_{\text{img}}^{(l)} = A_{\text{img}}^{(l)}$ and $S_{\text{tof}}^{(l)} = A_{\text{tof}}^{(l)}$. Then the $A_{\text{img}}^{(l)}$ and $A_{\text{tof}}^{(l)}$ are performed downsampling calculation respectively according to a pooling kernel with a size of 2×2 to obtain the image features and photon statistical features output by the layer of shared sub-network, which are expressed as $F_{\text{img}}^{(l)}$ and $F_{\text{tof}}^{(l)}$.

[0079] In the embodiment, the two modalities of input data may be respectively encoded and performed the downsampling calculation by each layer of at least two layers of shared sub-networks in sequence, while the encoded image features and photon statistical features are backed up, which can achieve multi-level encoding and dimension transformation on the projection data and the image pre-encoding data using the shared parameters of the shared sub-networks. During decoding, the backup image features and photon statistical features may serve as a reference for the decoding procedure, contributing to avoiding the loss of spatial details of the image features and photon statistical features, and thereby improving the accuracy of the decoded data.

[0080] In some alternative implementations, the depth decoder includes at least two layers of depth decoding sub-networks. Correspondingly, as shown in FIG. 5, the step 103 may include steps 1031-1033 based on the above embodiments.

[0081] In step 1031, for any one of the at least two layers of depth decoding sub-networks, the photon statistical features input to the layer of depth decoding sub-network are upsampled, and the upsampled photon statistical features are combined with the backup photon statistical features corresponding to the layer of depth decoding sub-network; and the combined photon statistical features are decoded to obtain the decoded photon statistical features.

[0082] At least two layers of depth decoding sub-networks may adopt a sequentially stacked structure, where the output data of an upper layer of sub-network is transmitted to a lower layer of

sub-network for processing.

[0083] For example, in some implementations, the layer configuration of the depth decoder is as follows:

[0084] Layer 4: 1024 input channels, 512 output channels, upsampled to 12×30 ;

[0085] Layer 3: 512 input channels, 256 output channels, upsampled to 24×60 ;

[0086] Layer 2: 256 input channels, 128 output channels, upsampled to 48×120 ;

[0087] Layer 1: 128 input channels, 64 output channels, upsampled to 96×240 .

[0088] The layer serial numbers of the depth decoder correspond to those of the shared encoder.

For instance, the layer 4 of the depth decoder corresponds to the layer 4 of the shared encoder.

[0089] Each layer of depth decoding sub-network may include an upsampling layer, a skip connection layer, and at least one convolutional layer. Data obtained after upsampling every time is combined with (i.e., skip connected to) the corresponding backup photon statistical features, then the combined photon statistical features undergo convolution at least once and ReLU activation function calculation to obtain the decoded photon statistical features, and the decoded photon statistical features are input to the next layer of depth decoding sub-network for calculating.

[0090] An expression of combining the above upsampled photon statistical features with the corresponding backup photon statistical features may be shown by the following equation (2):

$$F_{depth,concnt}^{(l)} = Concat(F_{depth,up}^{(l)}, S_{tof}^{(l)}) \quad (2)$$

[0091] where $F_{depth,up}^{(l)}$ is the upsampled photon statistical feature, and $S_{tof}^{(l)}$ is the backup photon statistical feature.

[0092] As an example, each layer of depth decoding sub-network may include two convolutional layers. After the first convolution, the output features go through calculation by the activation

function, followed by the second convolution and another calculation by the activation function to obtain photon statistical features output by the layer of depth decoding sub-network.

[0093] In step 1032, if the layer of depth decoding sub-network is not the last layer, the decoded photon statistical features are input to a next layer of depth decoding sub-network.

[0094] In step 1033, if the layer of depth decoding sub-network is the last layer, depth calculation is performed on the decoded photon statistical features to obtain the depth map.

[0095] The resolution of the photon statistical features output by the last layer of depth decoding sub-network is the same as the resolution of the initial photon statistical graph. During depth calculation, a data set in the photon statistical features, which corresponds to each pixel of the photon statistical graph, may be mapped to a depth value. The specific calculation equations are shown by the following equations (3) and (4):

$$Z_{depth}^{out} = W_{depth}^{out} * F_{depth}^l + b_{depth}^{out} \quad (3)$$

$$D = d_{max} \cdot \sigma(Z_{depth}^{out}) \quad (4)$$

[0096] where F_{depth}^l is photon statistical features output by the last layer of depth decoding sub-network, W_{depth}^{out} and b_{depth}^{out} are parameters obtained by training, d_{max} is a preset maximum measurement distance, and $\sigma()$ is a sigma activation function.

[0097] In the embodiment, multiple layers of depth decoding sub-networks are set in the depth decoder; and during calculating in each layer of depth decoding sub-network, subsequence decoding is performed after combining with the corresponding backup photon statistical features, which may take the backup photon statistical features as a reference for the decoding procedure, contributing to avoid the loss of spatial details of the photon statistical features, and thereby

improving the accuracy of the depth calculation after decoding.

[0098] In some alternative implementations, the image decoder includes at least two layers of image decoding sub-networks. Correspondingly, as shown in FIG. 6, the step 104 may include steps 1041-1043 based on the above embodiments.

[0099] In step 1041, for any one of the at least two layers of image decoding sub-networks, the image features input to the layer of image decoding sub-network are upsampled, and the upsampled image features are combined with the backup image features corresponding to the layer of image decoding sub-network; and the combined image features are decoded to obtain the decoded image features.

[00100] At least two layers of image decoding sub-networks may adopt a sequentially stacked structure, where the output data of an upper layer of sub-network is transmitted to a lower layer of sub-network for processing.

[00101] Each layer of the image decoding sub-network may include an upsampling layer, a skip connection layer, and at least one convolutional layer. Data obtained after upsampling every time is combined with (i.e., skip connected to) the corresponding backup image features, then the combined image features undergo convolution at least once and ReLU activation function calculation to obtain the decoded image features, and the decoded image features are input to the next layer of image decoding sub-network for calculating.

[00102] An expression of combining the above upsampled image features with the corresponding backup image features may be shown by the following equation (5):

$$F_{img,concnt}^{(l)} = Concat(F_{img,up}^{(l)}, S_{img}^{(l)}) \quad (5)$$

[00103] where $F_{img,up}^{(l)}$ is the upsampled image features, and $S_{img}^{(l)}$ is the backup image features.

[00104] In step 1042, if the layer of image decoding sub-network is not the last layer, the decoded image features are input to a next layer of image decoding sub-network.

[00105] In step 1043, if the layer of the image decoding sub-network is the last layer, image reconstruction is performed on the decoded image features to obtain the denoised image.

[00106] The resolution of the image features output by the last layer of image decoding sub-network is the same as the resolution of the initial scene image. During image reconstruction, the adopted calculation equations are shown by the following equations (6) and (7):

$$Z_{img}^{out} = W_{img}^{out} * F_{img}^l + b_{img}^{out} \quad (6)$$

$$I_{denoised} = \sigma(Z_{img}^{out}) \quad (7)$$

[00107] where F_{img}^l is the image features output by the last layer of image decoding sub-network, W_{img}^{out} and b_{img}^{out} are parameters obtained by training, and $\sigma()$ is a sigma activation function.

[00108] In the embodiment, multiple layers of image decoding sub-networks are set in the image decoder; and during calculating in each layer of image decoding sub-network, subsequence decoding is performed after combining with the corresponding backup image features, which may take the backup image features as a reference for the decoding procedure, contributing to avoid the loss of spatial details of the image features, and thereby improving the accuracy of the image denoised.

[00109] In some alternative implementations, as shown in FIG. 7, the image processing model, in any one of the above embodiments, may be obtained by training in advance according to the following steps 701-705.

[00110] In step 701, a sample scene image and a corresponding sample denoised image are obtained, and a sample photon statistical graph and a corresponding sample depth map are obtained.

[00111] The sample denoised image is an image after the sample scene image is denoised, which serves as a reference for image denoising during training. The sample depth map is a depth map obtained after a distance measurement is performed on a spatial position, indicated by each pixel, in the sample photon statistical graph, which serves as a reference for depth value calculation during training.

[00112] In step 702, the sample scene image and the sample photon statistical graph are input to an initial image processing model to obtain a predicted denoised image and a predicted depth map, respectively.

[00113] In the embodiment of the present disclosure, the initial image processing model is an image processing model to be trained, or may be an image processing model that has not yet started training or is in the process of training.

[00114] A structure of the initial image processing model and a procedure of image processing may refer to the above embodiments, and details are not described herein.

[00115] The predicted depth map and the predicted denoised image are the depth map and the denoised image output by the initial image processing model, respectively.

[00116] In step 703, depth loss between the sample depth map and the predicted depth map is determined based on a preset depth loss function, and denoising loss between the sample denoised image and the predicted denoised image is determined based on a preset image denoising loss function.

[00117] The depth loss function and the image denoising loss function may be constructed by means of various loss functions. For example, the depth loss function may be constructed by means of an L1 loss function, and the image denoising loss function may be constructed by means of an

L1 loss function and a perceptual loss function.

[00118] In step 704, the shared encoding network, the depth decoder, and the image decoder, which are included in the initial image processing model, are trained based on the depth loss and the denoising loss.

[00119] In some alternative implementations, minimizing the depth loss and the denoising loss may be taken as a training objective, and parameters of various modules included in the initial image processing model are adjusted iteratively using backpropagation and gradient descent.

[00120] In step 705, in response to the initial image processing model meeting a preset training end condition, the current initial image processing model is determined as the image processing model after training.

[00121] In the embodiment of the present disclosure, the steps 701-705 may be performed iteratively for training the initial image processing model until the preset training end condition is met.

[00122] In some alternative examples, the training end condition may include, but is not limited to, at least one of the following: convergence of depth loss and denoising loss, the training number reaching a preset number, and the training duration reaching a preset duration, etc.

[00123] In the embodiment, using multi-modal sample data to train the image processing model, the trained image processing model can perform more accurate depth map prediction and image denoising for actual application scenarios.

[00124] In some alternative implementations, as shown in FIG. 8, the step 704 may include steps 7041-7042 based on the embodiment shown in FIG. 7.

[00125] In step 7041, in a first stage in accordance with a staged training strategy, the shared

encoding network is trained based on the depth loss and the denoising loss.

[00126] In some alternative examples, the initial image processing model may be trained using photon statistical data and image data, respectively, during each iteration of training. For example, parameters of the depth decoder and the image decoder may be frozen. The sample photon statistical graph is input to the initial image processing model to obtain the predicted depth map output by the initial image processing model; a depth loss representing an error between the predicted depth map and the sample depth map is calculated, and parameters of the shared encoding network are adjusted to gradually reduce the depth loss. Also, the sample scene image is input, and the predicted denoised image is output; a denoising loss representing an error between the predicted denoised image and the sample denoised image is calculated, and parameters of the shared encoding network are adjusted to gradually reduce the denoising loss.

[00127] It can be understood that, in the case of the shared encoding network including the projection unit, the image pre-encoding unit and the shared encoder, there is no need to adjust parameters of the image pre-encoding unit when training the shared encoding network using the sample photon statistical graph, and there is no need to adjust parameters of the projection unit when training the shared encoding network using the sample scene image.

[00128] In step 7042, in a second stage, the depth loss and the denoising loss are redetermined, and the depth decoder and the image decoder are trained based on redetermined depth loss and denoising loss.

[00129] In the second stage, the parameters of the trained shared encoding network may be frozen, and the depth decoder and the image decoder are trained using photon statistical data and image data, respectively. For example, the sample photon statistical graph may be input and the predicted depth map is output, a depth loss representing an error between the predicted depth map and the

sample depth map is calculated, and parameters of the depth decoder are adjusted to gradually reduce the depth loss; and the sample scene image is input and the predicted denoised image is output, a denoising loss representing an error between the predicted denoised image and the sample denoised image is calculated, and parameters of the image decoder are adjusted to gradually reduce the denoising loss.

[00130] In the embodiment, training the model in stages facilitates targeted training of various modules contained in the model and improves model training efficiency and the accuracy of data processed by the model. Moreover, in the first stage of training, using two modalities of data to train the same shared encoding network, the shared encoding network can learn the ability to extract 3D features from the sample photon statistical graph and learn the ability to extract 2D features from the sample scene image. Thus, 2D features and 3D features from data of a single modality can be simultaneously extracted using shared parameters in the shared encoding network, thereby achieving complementarity between 2D features and 3D features. With the above scheme, the photon statistical features and image features output by the shared encoding network can more accurately represent shapes and positions of objects in an actual 3D space, thereby improving the accuracy of depth estimation and image denoising.

[00131] In some alternative implementations, the above step 7041, based on the embodiment shown in FIG. 8, may be implemented in the following ways:

[00132] First, the projection unit included in the shared encoding network is trained with the objective of minimizing the depth loss.

[00133] Then, the image pre-encoding unit included in the shared encoding network is trained with the objective of minimizing the denoising loss.

[00134] Finally, the shared encoder included in the shared encoding network is trained with the

objective of minimizing a weighted sum result of the depth loss and the denoising loss.

[00135] In the embodiment, the shared encoding network includes the projection unit, the image pre-encoding unit, and the shared encoder. The parameters of the projection unit and the image pre-encoding unit are independent of each other. The parameters of the shared encoder may be shared when the shared encoder extracts the photon statistical features and the image features; that is, the shared encoder is capable of encoding both the projection data and the image pre-encoding data.

[00136] When training the projection unit, the parameters of the projection unit may be iteratively adjusted so that the depth loss gradually reduces. When training the image pre-encoding unit, the parameters of the image pre-encoding unit may be iteratively adjusted so that the denoising loss gradually reduces. When training the projection unit, it is not necessary to input the sample scene image to the image pre-encoding unit, and when training the image pre-encoding unit, it is not necessary to input the sample photon statistical graph to the projection unit.

[00137] The above weighted sum of the depth loss and denoising loss may be calculated with and shown in the following equation (8):

$$L_{total} = \alpha L_{depth} + \beta L_{img} \quad (8)$$

[00138] where L_{total} represents a total loss value, L_{depth} is a depth loss, L_{img} is a denoising loss, α and β are preset weights and are real numbers greater than 0, respectively.

[00139] When training the shared encoder, the sample photon statistical graph and the sample scene image may be used respectively for training, i.e., the parameters of the shared encoder are iteratively adjusted by calculating the above total loss value, to make the total loss value reduce gradually.

[00140] In the embodiment, the projection unit, the image pre-encoding unit, and the shared encoder are respectively trained with sample data, which realizes training the various units of the shared encoding network in stages, thereby contributing to simplifying the training procedure and improving the training efficiency.

[00141] In some alternative implementations, the above step 7042, based on the embodiment shown in FIG. 8, may be implemented in the following ways:

[00142] The depth decoder and the image decoder are trained with the objective of minimizing a weighted sum result of the redetermined depth loss and denoising loss.

[00143] In the embodiment, a method for calculating the weighted sum of the depth loss and denoising loss may refer to the above equation (8). During training of the depth decoder and the image decoder, the sample photon statistical graph and the sample scene image may be input, and the predicted depth map and the predicted denoised image are output. The total loss value expressed by equation (8) is calculated, and then the parameters of the depth decoder and the image decoder are adjusted to minimize the total loss value until the training end condition is met.

[00144] In the embodiment, the depth decoder and image decoder are trained by performing the weighted sum of the depth loss and denoising loss, which can make the training procedure reflect the correlation between multi-modal data and further enable the parameters of the decoder to achieve complementarity between the 2D features and 3D features of the photon statistical features and image features, thereby improving the accuracy of depth estimation and image denoising.

[00145] In some alternative implementations, the above step 703 may include steps 7031-7033, as shown in FIG. 9.

[00146] In step 7031, a signal-to-noise ratio of the photon statistical data sequence corresponding

to each pixel in the sample photon statistical graph is determined.

[00147] The signal-to-noise ratio may be obtained by: determining a peak value from the photon statistical data sequence corresponding to each pixel, and determining a ratio between the peak value and a baseline noise intensity (for example, a mean value of data other than the peak data).

[00148] In step 7032, a loss weight corresponding to each pixel in the sample photon statistical graph is determined based on the signal-to-noise ratio.

[00149] Specifically, a method for calculating the loss weight corresponding to each pixel may be expressed by the following equation (9):

$$w_{i,j} = \frac{SNR_{i,j}}{\frac{1}{H_{tof} W_{tof}} \sum_{k,l} SNR_{k,l}} \quad (9)$$

[00150] where $SNR_{i,j}$ represents a signal-to-noise ratio corresponding to a pixel at coordinates (i, j), $\sum_{k,l} SNR_{k,l}$ represents a sum of signal-to-noise ratios of all pixels in the sample photon statistical graph, H_{tof} and W_{tof} are the height and width of the sample photon statistical graph, respectively.

[00151] In step 7033, the depth loss between the sample depth map and the predicted depth map is determined by taking the loss weight as a parameter of the depth loss function.

[00152] In some alternative examples, the depth loss function may be constructed using a weighted L1 loss function. For example, the depth loss function is expressed by the following equation (10):

$$L_{depth} = \frac{1}{H_{tof} W_{tof}} \sum_{i=1}^{H_{tof}} \sum_{j=1}^{W_{tof}} w_{i,j} \cdot |D_{i,j} - D_{i,j}^{gt}| \quad (10)$$

[00153] where $D_{i,j}$ is a depth value of (i, j) pixel in the predicted depth map, and $D_{i,j}^{gt}$ is a corresponding annotation depth value.

[00154] In the embodiment, by calculating the loss weight corresponding to each pixel in the sample photon statistical graph, and using the loss weight to calculate the depth loss, the depth loss can fully reflect regions with higher signal quality in the sample photon statistical graph, and the accuracy of determining the depth loss can be improved according to the signal quality, thereby improving the accuracy of model training.

[00155] In some alternative implementations, the above step 703 may include steps 7034-7036, as shown in FIG. 10.

[00156] In step 7034, absolute error loss between the sample denoised image and the predicted denoised image is determined based on an absolute error loss function included in the image denoising loss function.

[00157] The absolute error loss function is an L1 loss function, and the calculation equation is expressed by the following equation (11):

$$L_{L1} = \frac{1}{H_{img}W_{img}C} \sum_{i=1}^{H_{img}} \sum_{j=1}^{W_{img}} \sum_{c=1}^C |I_{i,j,c}^{denoised} - I_{i,j,c}^{clean}| \quad (11)$$

[00158] where, H_{img} , W_{img} and C are height, width, and number of channels of the sample scene image, respectively, and $I_{i,j,c}^{denoised}$ is the predicted denoised image, and $I_{i,j,c}^{clean}$ is the sample denoised image.

[00159] In step 7035, perceptual loss between the sample denoised image and the predicted denoised image is determined based on a perceptual loss function included in the image denoising loss function.

[00160] The calculation equation of the perceptual loss is expressed by the following equation (12):

$$L_{pere} = \frac{1}{L} \sum_{l=1}^L \frac{1}{H_l W_l C_l} \|\phi_l(I^{denoised}) - \phi_l(I^{clean})\| \quad (12)$$

[00161] where H_l , W_l , C_l are height, width, and number of channels, respectively, of a feature map of layer l , and ϕ_l represents a feature extraction function of layer l .

[00162] In step 7036, a weighted sum is performed on the absolute error loss and the perceptual loss to obtain the denoising loss.

[00163] A weighted sum equation can be expressed by the following equation (13):

$$L_{img} = L_{L1} + \lambda L_{pere} \quad (13)$$

[00164] where λ is a weight. For example, $\lambda=0.1$, correspondingly in the above equation (8), $\alpha=0.1$ and $\beta=0.8$.

[00165] The embodiment configures the absolute error loss and the perceptual loss for the image data to calculate the denoising loss, which achieves that the error between the predicted denoised image and the sample denoised image is represented from multiple dimensions, and the accuracy of the denoising loss is improved.

[00166] In some alternative implementations, after step 7042, step 704 may further include: in a third stage, redetermining the depth loss and the denoising loss, and resetting a learning rate; and training jointly, based on the depth loss and denoising loss redetermined in the third stage, as well as the resetting learning rate, the shared encoding network, the depth decoder, and the image decoder.

[00167] Jointly training the shared encoding network, the depth decoder, and the image decoder

refers to simultaneously adjusting parameters for jointly training the shared encoding network, the depth decoder, and the image decoder using the input sample photon statistical graph and sample scene image. The above learning rate is a hyperparameter in deep learning, which controls a step size of parameter updating during model training. In the embodiment, after the completion of the first stage and the second stage, the third stage may be further entered, and the learning rate is set to be a smaller value, which can realize fine-tuning of the model, thereby further improving the accuracy of depth estimation and image denoising of the model. The implementation for determining the depth loss and the denoising loss, as well as the specific procedure for training the shared encoding network, the depth decoder, and the image decoder, may refer to the implementations of the above-described embodiments and will not be repeated here.

[00168] Exemplary Apparatuses

[00169] FIG. 11 is a structural diagram of an image processing apparatus according to an exemplary embodiment of the present disclosure. The image processing apparatus provided in the embodiments of the present disclosure can be used for implementing the image processing method of the embodiments of the present disclosure. The embodiment may be applied to an electronic device. As shown in FIG. 11, the image processing apparatus includes: an obtaining module 1101, an encoding module 1102, a first decoding module 1103, and a second decoding module 1104. The obtaining module 1101 is configured to obtain a photon statistical graph and a scene image corresponding to a target scene. A pixel corresponds to a photon statistical data sequence in the photon statistical graph, photon statistical data in the photon statistical data sequence represents a number of reflected photons received from a target position at a corresponding moment, and the target position is a position corresponding to the pixel in the target scene. The encoding module 1102 is configured to encode the scene image and the photon statistical graph, respectively, by a

shared encoding network included in a pre-trained image processing model to obtain image features and photon statistical features. The first decoding module 1103 is configured to decode the photon statistical features by a depth decoder included in the image processing model, and generate a depth map corresponding to the target scene using the decoded photon statistical features. The second decoding module 1104 is configured to decode the image features by an image decoder included in the image processing model to obtain a denoised image.

[00170] Refer to FIG. 12, FIG. 12 is a structural diagram of an image processing apparatus according to another exemplary embodiment of the present disclosure.

[00171] In some alternative implementations, the encoding module 1102 may include: a first transformation unit 11021, a second transformation unit 11022, and an encoding unit 11023. The first transformation unit 11021 is configured to project the photon statistical graph by a projection unit included in the shared encoding network to obtain projection data. The second transformation unit 11022 is configured to pre-encode the scene image by an image pre-encoding unit included in the shared encoding network to obtain image pre-encoding data. The projection data and the image pre-encoding data are identical in dimension. The encoding unit 11023 is configured to encode the projection data and the image pre-encoding data, respectively, by a shared encoder included in the shared encoding network to obtain the photon statistical features and the image features.

[00172] In some alternative implementations, the encoding unit 11023 may include: a first encoding sub-unit 110231 and a second encoding sub-unit 110232. The first encoding sub-unit 110231 is configured to encode the projection data by the shared encoder to obtain the photon statistical features including two-dimensional planar features and three-dimensional depth features. The second encoding sub-unit 110232 is configured to encode the image pre-encoding data by the shared encoder to obtain the image features including two-dimensional planar features and three-

dimensional depth features.

[00173] In some alternative implementations, the shared encoder includes at least two layers of shared sub-networks arranged in sequence. Correspondingly, the encoding unit 11023 may be further configured to, for any one of the at least two layers of shared sub-networks, encode and downsample, using shared parameters of the layer of shared sub-network, data input to the layer of shared sub-network, to obtain the image features and the photon statistical features after downsampling, and back up image features and photon statistical features after encoding and before downsampling.

[00174] In some alternative implementations, the depth decoder includes at least two layers of depth decoding sub-networks. Correspondingly, the first decoding module 1103 may be further configured to, for any one of the at least two layers of depth decoding sub-networks, upsample the photon statistical features input to the layer of depth decoding sub-network, combine the upsampled photon statistical features with the backup photon statistical features corresponding to the layer of depth decoding sub-network, and decode the combined photon statistical features to obtain the decoded photon statistical features; if the layer of depth decoding sub-network is not the last layer, input the decoded photon statistical features to a next layer of depth decoding sub-network; if the layer of depth decoding sub-network is the last layer, perform depth calculation on the decoded photon statistical features to obtain the depth map.

[00175] In some alternative implementations, the image decoder includes at least two layers of image decoding sub-networks. Correspondingly, the second decoding module 1104 may be further configured to, for any one of the at least two layers of image decoding sub-networks, upsample the image features input to the layer of image decoding sub-network, combine the upsampled image features with the backup image features corresponding to the layer of image decoding sub-

network, and decode the combined image features to obtain the decoded image features; if the layer of image decoding sub-network is not the last layer, inputting the decoded image features to a next layer of image decoding sub-network; and if the layer of image decoding sub-network is the last layer, performing image reconstruction on the decoded image features to obtain the denoised image.

[00176] In some alternative implementations, the image processing model may be pre-trained and obtained according to following steps: obtaining a sample scene image and a corresponding sample denoised image, as well as a sample photon statistical graph and a corresponding sample depth map; inputting the sample scene image and the sample photon statistical graph to an initial image processing model to obtain a predicted denoised image and a predicted depth map, respectively; determining, based on a preset depth loss function, depth loss between the sample depth map and the predicted depth map, and determining, based on a preset image denoising loss function, denoising loss between the sample denoised image and the predicted denoised image; training, based on the depth loss and the denoising loss, the shared encoding network, the depth decoder and the image decoder included in the initial image processing model; and in response to the initial image processing model meeting a preset training end condition, determining the current initial image processing model as the image processing model after training.

[00177] In some alternative implementations, the training, based on the depth loss and the denoising loss, the shared encoding network, the depth decoder and the image decoder included in the initial image processing model may include: in accordance with a staged training strategy, in a first stage, training, based on the depth loss and the denoising loss, the shared encoding network; and in a second stage, redetermining the depth loss and the denoising loss, and training, based on redetermined depth loss and denoising loss, the depth decoder and the image decoder.

[00178] In some alternative implementations, the training, based on the depth loss and the denoising loss, the shared encoding network may include: training a projection unit included in the shared encoding network with an objective of minimizing the depth loss; training an image pre-encoding unit included in the shared encoding network with an objective of minimizing the denoising loss; and training a shared encoder included in the shared encoding network with an objective of minimizing a weighted sum result of the depth loss and the denoising loss.

[00179] In some alternative implementations, the training, based on the redetermined depth loss and denoising loss, the depth decoder and the image decoder may include: training the depth decoder and the image decoder with an objective of minimizing a weighted sum result of the redetermined depth loss and denoising loss.

[00180] In some alternative implementations, the determining, based on a preset depth loss function, depth loss between the sample depth map and the predicted depth map may include: determining a signal-to-noise ratio of the photon statistical data sequence corresponding to each pixel in the sample photon statistical graph; determining, based on the signal-to-noise ratio, loss weight corresponding to each pixel in the sample photon statistical graph; and determining the depth loss between the sample depth map and the predicted depth map by taking the loss weight as a parameter of the depth loss function.

[00181] In some alternative implementations, the determining, based on a preset image denoising loss function, denoising loss between the sample denoised image and the predicted denoised image may include: determining, based on an absolute error loss function included in the image denoising loss function, absolute error loss between the sample denoised image and the predicted denoised image; determining, based on a perceptual loss function included in the image denoising loss function, perceptual loss between the sample denoised image and the predicted denoised image;

and performing weighted sum of the absolute error loss and the perceptual loss, to obtain the denoising loss.

[00182] In some alternative implementations, after redetermining, in the second stage, the depth loss and the denoising loss, and training, based on the redetermined depth loss and denoising loss, the depth decoder and the image decoder, the apparatus may be further configured to: in a third stage, redetermine the depth loss and the denoising loss, and reset a learning rate; and train jointly, based on the depth loss and denoising loss redetermined in the third stage, as well as the learning rate, the shared encoding network, the depth decoder and the image decoder.

[00183] The exemplary embodiments of the apparatus correspond to the above exemplary embodiments of the method in terms of implementation. For corresponding contents, the apparatus and the method may refer to, combine with, or cite each other, and thus, details are not repeated herein. The effects corresponding to the exemplary embodiments of the apparatus can refer to the corresponding effects of the above exemplary embodiments of the method, and are not repeated herein.

[00184] Exemplary Electronic Device

[00185] FIG. 13 is a structural diagram of an electronic device according to an exemplary embodiment of the present disclosure. The electronic device includes at least one processor 1301 and a memory 1302.

[00186] The processor 1301 may be a central processing unit (CPU) or another form of processing unit having a data processing capability and/or an instruction execution capability, and may control another component in the electronic device to perform a desired function.

[00187] The memory 1302 may include one or more computer program products. The computer program product may include various forms of computer-readable storage media, such as volatile

memory and/or non-volatile memory. The volatile memory may include, for example, a random access memory (RAM) and/or a cache, etc. The non-volatile memory may include, for example, a read-only memory (ROM), a hard disk, and a flash memory, etc. One or more computer program instructions may be stored on the computer-readable storage medium. The processor 1301 may execute one or more computer program instructions to implement the image processing method according to various embodiments of the present disclosure that are described above and/or other desired functions.

[00188] In an example, the electronic device may further include an input device 1303 and an output device 1304. These components are connected to each other through a bus system and/or another form of connection mechanism (not shown).

[00189] The input device 1303 may include, for example, a keyboard, a mouse, and the like.

[00190] The output device 1304 may output various information to the outside, which may include, for example, a display, a speaker, a printer, as well as a communication network and remote output devices connected thereto, and so on.

[00191] Certainly, for simplicity, FIG. 13 shows only some of the components in the electronic device that are related to the present disclosure, and components such as a bus and an input/output interface, etc., are omitted. In addition, the electronic device may further include any other appropriate components according to specific application situations.

[00192] Exemplary computer program product and computer-readable storage medium

[00193] In addition to the foregoing method and device, the embodiments of the present disclosure may also provide a computer program product, which includes computer program instructions. When the computer program instructions are run by a processor, the processor is enabled to perform the steps of the image processing method of the various embodiments of the present disclosure, as described in the above "exemplary method" part.

[00194] The computer program product may write program codes that are configured to perform the operations in the embodiments of the present disclosure, with one or any combination of one or more programming languages. The programming languages include an object-oriented programming language, such as Java or C++, etc., and further include a conventional procedural programming language, such as the "C" language or a similar programming language. The program codes may be entirely or partially executed on a user computing device, executed as an independent software package, partially executed on the user computing device and partially executed on a remote computing device, or entirely executed on the remote computing device or a server.

[00195] In addition, the embodiments of the present disclosure may further relate to a computer-readable storage medium, which stores computer program instructions. When the computer program instructions are run by the processor, the processor is enabled to perform the steps of the image processing method of the various embodiments of the present disclosure, as described in the above "exemplary method" part.

[00196] The computer-readable storage medium may be one or any combination of a plurality of readable media. The readable medium may be a readable signal medium or a readable storage medium. The readable storage medium may include, for example, but is not limited to, a system, an apparatus, or a device of electricity, magnetism, light, electromagnetism, infrared rays, or semiconductor, or any combination of the above. More specific examples (a non-exhaustive list) of the readable storage medium include: an electrical connection with one or more conducting wires, a portable disk, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or a flash memory), an optical fiber, a portable compact disk read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the above.

[00197] Basic principles of the present disclosure are described above in combination with specific embodiments. However, the advantages, superiorities, and effects mentioned in the present disclosure are merely examples but are not for limitation, and it cannot be considered that these advantages, superiorities, and effects are necessary for each embodiment of the present disclosure. In addition, specific details described above are merely for illustrative and ease-of-understanding purposes, rather than limiting. The above details do not restrict the present disclosure to be implemented using the foregoing specific details.

[00198] Those skilled in the art may make various modifications and variations to the present disclosure without departing from the spirit and scope of the present disclosure. Accordingly, provided that these modifications and variations of the present disclosure fall within the scope of the claims of the present disclosure and their equivalents, the present disclosure is also intended to encompass such modifications and variations.

CLAIMS

1. An image processing method, comprising:

obtaining a photon statistical graph and a scene image corresponding to a target scene, wherein, a pixel corresponds to a photon statistical data sequence in the photon statistical graph, photon statistical data in the photon statistical data sequence represents a number of reflected photons received from a target position at a corresponding moment, and the target position is a position corresponding to the pixel in the target scene;

encoding the scene image and the photon statistical graph, respectively, by a shared encoding network comprised in a pre-trained image processing model to obtain image features and photon statistical features;

decoding the photon statistical features by a depth decoder comprised in the image processing model, and generating a depth map corresponding to the target scene using the decoded photon statistical features; and

decoding the image features by an image decoder comprised in the image processing model to obtain a denoised image.

2. The method according to claim 1, wherein the encoding the scene image and the photon statistical graph, respectively, by a shared encoding network comprised in a pre-trained image processing model to obtain image features and photon statistical features comprises:

projecting the photon statistical graph by a projection unit comprised in the shared encoding network to obtain projection data;

pre-encoding the scene image by an image pre-encoding unit comprised in the shared encoding network to obtain image pre-encoding data, wherein the projection data and the image pre-encoding data are identical in dimension; and

encoding the projection data and the image pre-encoding data, respectively, by a shared encoder comprised in the shared encoding network to obtain the photon statistical features and the image features.

3. The method according to claim 2, wherein the encoding the projection data and the image pre-encoding data respectively by a shared encoder comprised in the shared encoding network to

obtain the photon statistical features and the image features comprises:

encoding the projection data by the shared encoder to obtain the photon statistical features comprising two-dimensional planar features and three-dimensional depth features; and

encoding the image pre-encoding data by the shared encoder to obtain the image features comprising two-dimensional planar features and three-dimensional depth features.

4. The method according to claim 2, wherein the shared encoder comprises at least two layers of shared sub-networks arranged in sequence; and

the encoding the projection data and the image pre-encoding data, respectively, by a shared encoder comprised in the shared encoding network to obtain the photon statistical features and the image features comprises:

encoding and downsampling, using shared parameters of a layer of the at least two layers of shared sub-networks, data input to the layer of shared sub-network, to obtain the image features and the photon statistical features after downsampling, and backing up image features and photon statistical features after encoding and before downsampling.

5. The method according to claim 4, wherein the depth decoder comprises at least two layers of depth decoding sub-networks; and

the decoding the photon statistical features by a depth decoder comprised in the image processing model, and generating a depth map corresponding to the target scene using the decoded photon statistical features comprises:

upsampling the photon statistical features input to a layer of the at least two layers of depth decoding sub-networks, and combining the upsampled photon statistical features with the backup photon statistical features corresponding to the layer of depth decoding sub-network; and decoding the combined photon statistical features to obtain the decoded photon statistical features;

if the layer of depth decoding sub-network is not the last layer, inputting the decoded photon statistical features to a next layer of depth decoding sub-network;

if the layer of depth decoding sub-network is the last layer, performing depth calculation on the decoded photon statistical features to obtain the depth map.

6. The method according to claim 4, wherein the image decoder comprises at least two layers of image decoding sub-networks; and

the decoding the image features by an image decoder comprised in the image processing model to obtain a denoised image comprises:

upsampling the image features input to a layer of the at least two layers of image decoding sub-networks, and combining the upsampled image features with the backup image features corresponding to the layer of image decoding sub-network; and decoding the combined image features to obtain the decoded image features;

if the layer of image decoding sub-network is not the last layer, inputting the decoded image features to a next layer of image decoding sub-network;

if the layer of image decoding sub-network is the last layer, performing image reconstruction on the decoded image features to obtain the denoised image.

7. The method according to claim 1, wherein the image processing model is pre-trained in the following way:

obtaining a sample scene image and a corresponding sample denoised image, as well as a sample photon statistical graph and a corresponding sample depth map;

inputting the sample scene image and the sample photon statistical graph to an initial image processing model to obtain a predicted denoised image and a predicted depth map, respectively;

determining, based on a preset depth loss function, depth loss between the sample depth map and the predicted depth map, and determining, based on a preset image denoising loss function, denoising loss between the sample denoised image and the predicted denoised image;

training, based on the depth loss and the denoising loss, the shared encoding network, the depth decoder, and the image decoder comprised in the initial image processing model; and

in response to the initial image processing model meeting a preset training end condition, determining the current initial image processing model as the image processing model after training.

8. The method according to claim 7, wherein the training, based on the depth loss and the

denoising loss, the shared encoding network, the depth decoder, and the image decoder comprised in the initial image processing model comprises:

in a first stage, in accordance with a staged training strategy, training, based on the depth loss and the denoising loss, the shared encoding network; and

in a second stage, redetermining the depth loss and the denoising loss, and training, based on redetermined depth loss and denoising loss, the depth decoder and the image decoder.

9. The method according to claim 8, wherein the training, based on the depth loss and the denoising loss, the shared encoding network comprises:

training a projection unit comprised in the shared encoding network with an objective of minimizing the depth loss;

training an image pre-encoding unit comprised in the shared encoding network with an objective of minimizing the denoising loss; and

training a shared encoder comprised in the shared encoding network with an objective of minimizing a weighted sum result of the depth loss and the denoising loss.

10. The method according to claim 8, wherein the training, based on redetermined depth loss and denoising loss, the depth decoder and the image decoder comprises:

training the depth decoder and the image decoder with an objective of minimizing a weighted sum result of the redetermined depth loss and denoising loss.

11. The method according to claim 7, wherein the determining, based on a preset depth loss function, depth loss between the sample depth map and the predicted depth map comprises:

determining a signal-to-noise ratio of the photon statistical data sequence corresponding to each pixel in the sample photon statistical graph;

determining, based on the signal-to-noise ratio, a loss weight corresponding to each pixel in the sample photon statistical graph; and

determining the depth loss between the sample depth map and the predicted depth map by taking the loss weight as a parameter of the depth loss function.

12. The method according to claim 7, wherein the determining, based on a preset image

denoising loss function, denoising loss between the sample denoised image and the predicted denoised image comprises:

determining, based on an absolute error loss function comprised in the image denoising loss function, absolute error loss between the sample denoised image and the predicted denoised image;

determining, based on a perceptual loss function comprised in the image denoising loss function, perceptual loss between the sample denoised image and the predicted denoised image;
and

performing a weighted sum of the absolute error loss and the perceptual loss, to obtain the denoising loss.

13. The method according to claim 8, wherein, after redetermining, in the second stage, the depth loss and the denoising loss, and training, based on the redetermined depth loss and denoising loss, the depth decoder and the image decoder, the method further comprises:

in a third stage, redetermining the depth loss and the denoising loss, and resetting a learning rate; and

training jointly, based on the depth loss and denoising loss redetermined in the third stage, as well as the learning rate, the shared encoding network, the depth decoder, and the image decoder.

14. An image processing apparatus, comprising:

an obtaining model, configured to obtain a photon statistical graph and a scene image corresponding to a target scene, wherein, a pixel corresponds to a photon statistical data sequence in the photon statistical graph, photon statistical data in the photon statistical data sequence represents a number of reflected photons received from a target position at a corresponding moment, and the target position is a position corresponding to the pixel in the target scene;

an encoding module, configured to encode the scene image and the photon statistical graph respectively by a shared encoding network comprised in a pre-trained image processing model to obtain image features and photon statistical features;

a first decoding module, configured to decode the photon statistical features by a depth decoder comprised in the image processing model, and generate a depth map corresponding to the target scene using the decoded photon statistical features; and

a second decoding module, configured to decode the image features by an image decoder comprised in the image processing model to obtain a denoised image.

15. A computer-readable storage medium comprising a computer program stored thereon, wherein the computer program, when executed, is configured to implement the image processing method according to any one of claims 1 to 13.

16. An electronic device, comprising:

a memory, configured to store instructions; and

a processor, configured to read the instructions from the memory, and execute the instructions to implement the image processing method according to any one of claims 1 to 13.