

# **MONOCULAR IMAGE DEPTH ESTIMATION METHOD AND APPARATUS, DEVICE, PROGRAM AND STORAGE MEDIUM**

5

## **CROSS-REFERENCE TO RELATED APPLICATION**

The present application is based on and claims priority to Chinese Patent Application No. 201810845040.4 filed on July 27, 2018, the disclosure of which is  
10 hereby incorporated by reference in its entirety.

## **TECHNICAL FIELD**

15 The present disclosure relates to computer vision technologies, and in particular, to a method and apparatus for the depth estimation of a monocular image, an electronic device, a computer program, and a storage medium.

20

## **BACKGROUND**

The depth estimation of images is an important problem in the field of computer vision. The current depth estimation of images mainly includes: the depth estimation of monocular images and the depth estimation of binocular images. The  
25 depth estimation of monocular images is mainly to estimate depth information of images on the basis of monocular images. However, because the depth estimation of monocular images is a very challenging problem, at present, the accuracy of the depth information of monocular images predicted by many existing methods for the depth estimation of monocular images is still poor.

30

## **SUMMARY**

Embodiments of the present disclosure provide a technical solution for the  
35 depth estimation of a monocular image.

According to one aspect of the embodiments of the present disclosure, provided is a method for the depth estimation of a monocular image, including:

on the basis of a depth estimation neural network, obtaining a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image; and

obtaining a predicted depth map of the monocular image according to the  
5 global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image.

According to one or more disclosure of the present disclosure, in the method embodiments of the present disclosure, before the obtaining a global feature of a monocular image according to absolute features of preset regions and relative features  
10 among the preset regions in the monocular image, the method further includes:

performing feature extraction on the monocular image by means of a first neural network to obtain features of preset regions in the monocular image, and taking the features of the preset regions as absolute features of the preset regions in the monocular image; and

15 obtaining relative features among the preset regions in the monocular image according to the absolute features of the preset regions in the monocular image.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, the obtaining relative features among the preset regions in the monocular image according to the absolute features of the  
20 preset regions in the monocular image includes:

performing vector operation on the absolute features of the preset regions in the monocular image by means of an association layer to obtain the relative features among the preset regions in the monocular image.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, before performing feature extraction on the monocular image by means of a first neural network, the method further includes:

performing downsampling on the monocular image to obtain a monocular image having a preset dimension, where the dimension of the monocular image is a  
30 multiple of the preset dimension.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, the obtaining a global feature of a

monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image includes:

obtaining a global feature of a monocular image by combining absolute features of preset regions and relative features among the preset regions in the monocular image by means of a full connection layer.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, the obtaining a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image includes:

performing depth estimation by means of a depth estimator according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image to obtain a predicted depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, after the obtaining a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image includes:

performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, the performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image includes:

performing residual estimation on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map; and

performing optimization on the predicted depth map according to the residual plot to obtain a target depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, the performing residual estimation on

the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map includes

performing residual estimation on the predicted depth map by means of a residual estimation network according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map; and

the performing optimization on the predicted depth map according to the residual plot to obtain a target depth map of the monocular image includes

performing pixel-by-pixel superposition operation on the residual plot and the predicted depth map to obtain a target depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, before the performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image, the method further includes:

obtaining a longitudinal variation law of depth information of the monocular image according to the predicted depth map.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, the obtaining a longitudinal variation law of depth information of the monocular image according to the predicted depth map includes:

performing processing on the predicted depth map by means of a longitudinal pooling layer to obtain a longitudinal variation law of depth information of the monocular image.

According to one or more disclosure of the present disclosure, in any of the method embodiments of the present disclosure, the performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image includes:

performing a preset number of upsamplings on the predicted depth map, obtaining a longitudinal variation law of depth information according to the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple, and performing optimization on the predicted depth map, obtained by

each upsampling, having a dimension sequentially increased by a multiple according to the longitudinal variation law of the depth information thereof to obtain an optimized target depth map;

where the optimized target depth map obtained by each of the upsamplings  
5 other than the last upsampling is taken as the predicted depth map of the next upsampling, the optimized target depth map obtained by the last upsampling is taken as the target depth map of the monocular image, and the target depth map has the same dimension as the monocular image.

According to one or more disclosure of the present disclosure, in any of the  
10 method embodiments of the present disclosure, the depth estimation neural network includes an association layer, a full connection layer, and a depth estimator, and is obtained by training the depth estimation neural network using a sparse depth map and a dense depth map obtained by means of stereoscopic binocular image matching as marking data.

15 According to another aspect of the embodiments of the present disclosure, provided is an apparatus for the depth estimation of a binocular image, including:

a depth estimation neural network, configured to obtain a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image, and obtain a predicted depth map  
20 of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image.

According to one or more disclosure of the present disclosure, in the apparatus embodiments of the present disclosure, the apparatus further includes:

a first neural network, configured to perform feature extraction on the  
25 monocular image to obtain features of preset regions in the monocular image, and take the features of the preset regions as absolute features of the preset regions in the monocular image; and

the depth estimation neural network, further configured to obtain relative features among the preset regions in the monocular image according to the absolute  
30 features of the preset regions in the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the depth estimation neural network includes:

an association layer, configured to perform vector operation on the absolute features of the preset regions in the monocular image to obtain the relative features among the preset regions in the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the apparatus further includes:

a downsampling layer, configured to perform downsampling on the monocular image before performing feature extraction on the monocular image to obtain a monocular image having a preset dimension, where the dimension of the monocular image is a multiple of the preset dimension.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the depth estimation neural network includes:

a full connection layer, configured to obtain a global feature of a monocular image by combining absolute features of preset regions and relative features among the preset regions in the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the depth estimation neural network includes:

a depth estimator, configured to perform depth estimation according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image to obtain a predicted depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the apparatus further includes:

a second neural network, configured to perform optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the second neural network is configured to perform residual estimation on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map, and perform optimization on the predicted

depth map according to the residual plot to obtain a target depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the second neural network includes:

5        a residual estimation network, configured to perform residual estimation on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map; and

10        an additive operation unit, configured to perform pixel-by-pixel superposition operation on the residual plot and the predicted depth map to obtain a target depth map of the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the second neural network is further configured to obtain a longitudinal variation law of depth information of the monocular image according to the predicted depth map.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the second neural network includes:

20        a longitudinal pooling layer, configured to perform processing on the predicted depth map to obtain a longitudinal variation law of depth information of the monocular image.

According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the apparatus further includes:

an upsampling layer, configured to perform a preset number of upsamplings on the predicted depth map;

25        a longitudinal pooling layer, configured to obtain a longitudinal variation law of depth information according to the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple; and

30        the second neural network, configured to perform optimization on the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple according to the longitudinal variation law of the depth information thereof to obtain an optimized target depth map;

where the optimized target depth map obtained by each of the upsamplings other than the last upsampling is taken as the predicted depth map of the next

upsampling, the optimized target depth map obtained by the last upsampling is taken as the target depth map of the monocular image, and the target depth map has the same dimension as the monocular image.

5 According to one or more disclosure of the present disclosure, in any of the apparatus embodiments of the present disclosure, the depth estimation neural network includes an association layer, a full connection layer, and a depth estimator, and is obtained by training the depth estimation neural network using a sparse depth map and a dense depth map obtained by means of stereoscopic binocular image matching as marking data.

10 According to still another aspect of the embodiments of the present disclosure, provided is an electronic device, including the apparatus according to any one of the embodiments.

According to yet another aspect of the embodiments of the present disclosure, provided is an electronic device, including:

15 a memory, configured to store executable instructions; and  
a processor configured to execute the executable instructions to implement the method according to any one of the embodiments.

According to yet another aspect of the embodiments of the present disclosure, provided is a computer program, including a computer readable code that, when being  
20 executed in a device, causes a processor in the device to execute instructions for implementing the method according to any one of the embodiments.

According to yet another aspect of the embodiments of the present disclosure, provided is a computer storage medium configured to store computer-readable instructions that, when being executed, implement the method according to any one of  
25 the embodiments.

On the basis of the method and apparatus for the depth estimation of a monocular image, the electronic device, the computer program, and the storage medium provided in the embodiments of the present disclosure, a global feature of a monocular image is obtained according to absolute features of preset regions and  
30 relative features among the preset regions in the monocular image on the basis of a depth estimation neural network, and a predicted depth map of the monocular image is obtained according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image. By means of

the mutual supplementation between relative features and absolute features of preset regions in an image in the depth estimation of monocular images, the accuracy of the relative distance prediction in the depth estimation is improved, such that the accuracy of the depth estimation of monocular images can be improved.

5           The technical solutions of the present disclosure are further described in detail with reference to the accompanying drawings and embodiments as follows.

## BRIEF DESCRIPTION OF DRAWINGS

10           The accompanying drawings constituting a part of the specification describe the embodiments of the present disclosure and are intended to explain the principles of the present disclosure together with the descriptions.

          According to the following detailed descriptions, the present disclosure can be  
15   understood more clearly with reference to the accompanying drawings.

          FIG. 1 is a flowchart of a method for the depth estimation of a monocular image according to some embodiments of the present disclosure;

          FIG. 2 is a flowchart of a method for the depth estimation of a monocular image according to other embodiments of the present disclosure;

20           FIG. 3 is a flowchart of each scale optimization during performing optimization using multi-scale learning of the embodiments of the present disclosure.

          FIGs. 4A to 4C are structural schematic diagrams of networks for implementing the method for the depth estimation of a monocular image according to some embodiments of the present disclosure;

25           FIG. 5 is a schematic structural diagram of an apparatus for the depth estimation of a monocular image according to some embodiments of the present disclosure;

          FIG. 6 is a schematic structural diagram of an apparatus for the depth estimation of a monocular image according to other embodiments of the present  
30   disclosure;

          FIG. 7 is a schematic structural diagram of an apparatus for the depth estimation of a monocular image according to some other embodiments of the present disclosure; and

FIG. 8 is a schematic structural diagram of an electronic device provided by some embodiments of the present disclosure.

5

## DETAILED DESCRIPTION

Various exemplary embodiments of the present disclosure are now described in detail with reference to the accompanying drawings. It should be noted that, unless otherwise stated specifically, relative arrangement of the components and operations,  
10 numerical expressions, and values set forth in the embodiments are not intended to limit the scope of the present disclosure.

In addition, it should be understood that, for ease of description, the size of each part shown in the accompanying drawings is not drawn in actual proportion.

The following descriptions of at least one exemplary embodiment are merely  
15 illustrative actually, and are not intended to limit the present disclosure and applications or uses thereof.

Technologies, methods and devices known to a person of ordinary skill in the related art may not be discussed in detail, but such technologies, methods and devices should be considered as a part of the specification in appropriate situations.

20 It should be noted that similar reference numerals and letters in the following accompanying drawings represent similar items. Therefore, once an item is defined in an accompanying drawing, the item does not need to be further discussed in the subsequent accompanying drawings.

The embodiments of the present disclosure may be applied to a computer  
25 system/server, which may operate with numerous other general-purpose or special-purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations suitable for use together with the computer system/server include, but are not limited to, personal computer systems, server computer systems, thin clients, thick clients, handheld or laptop  
30 devices, microprocessor-based systems, set top boxes, programmable consumer electronics, network personal computers, small computer systems, large computer systems, distributed cloud computing environments that include any of the foregoing systems, and the like.

The computer system/server may be described in the general context of computer system executable instructions (for example, program modules) executed by the computer system. Generally, the program modules may include routines, programs, target programs, components, logics, data structures, and the like for performing specific tasks or implementing specific abstract data types. The computer systems/servers may be practiced in the distributed cloud computing environments in which tasks are performed by remote processing devices that are linked through a communications network. In the distributed computing environments, the program modules may be located in local or remote computing system storage mediums including storage devices.

FIG. 1 is a flowchart of a method for the depth estimation of a monocular image according to some embodiments of the present disclosure.

As shown in FIG. 1, the method includes the following operations.

At 102, on the basis of a depth estimation neural network, a global feature of a monocular image is obtained according to absolute features of preset regions and relative features among the preset regions in the monocular image.

In the embodiments, the monocular image may be an image obtained from an image collection device, and may also be an image obtained from a storage apparatus. For example, the image collection device may be a still camera, a video camera, a scanner, or the like, and the storage apparatus may be a USB flash drive, an optical disk, a hard disk drive, or the like. The embodiments do not limit the obtaining mode of the monocular image. The absolute features of the preset regions in the monocular image are used for representing local appearances of the preset regions in the monocular image. For example, the absolute features include texture features, geometric features, or the like. The relative features among the preset regions in the monocular image are used for representing the differences between the local appearances of the preset regions in the monocular image. For example, the relative features include texture differences, geometric differences, or the like. The preset regions in the monocular image are set according to the features of the image. The depth map of the embodiments is a map that represents the distances between pixels in an image and the image collection device using the pixels values of the pixels in the image.

In an optional example, a global feature of a monocular image is obtained by combining absolute features of preset regions and relative features among the preset regions in the monocular image by means of a full connection layer.

At 104, a predicted depth map of the monocular image is obtained according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image.

In an optional example, depth estimation can be performed by means of a depth estimator according to the global feature of the monocular image and the absolute features of preset regions and relative features among the preset regions in the monocular image to obtain a predicted depth map of the monocular image. For example, the depth estimator may be a fully convolutional network which mainly consists of a convolution layer and a deconvolution layer. The depth estimator is able to calculate depths values of pixels in an image according to geometric distribution information of the image, i.e., a global feature of the image, and absolute features of preset regions and relative features among the preset regions in the image, to obtain the predicted depth map.

The method for the depth estimation of a monocular image provided by the embodiments obtains a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image on the basis of a depth estimation neural network, and obtains a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image. By means of the mutual supplementation between relative features and absolute features of preset regions in an image in the depth estimation of monocular images, the accuracy of the relative distance prediction in the depth estimation is improved, such that the accuracy of the depth estimation of monocular images can be improved.

According to one or more disclosure of the present disclosure, before the obtaining a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image, feature extraction is further performed on the monocular image by means of a first neural network to obtain features of preset regions in the monocular image, the features of the preset regions are taken as absolute features of the present regions in

the monocular image, and then relative features among the preset regions in the monocular image are obtained according to the absolute features of the preset regions in the monocular image. For example, the first neural network may be a coder network consisting of a convolution layer and a pooling layer, and high-dimension  
5 features of the image can be obtained by performing feature extraction on the monocular image by means of the coder network.

In an optional example, vector operation is performed on the absolute features of the preset regions in the monocular image by means of an association layer to obtain the relative features among the preset regions in the monocular image. The  
10 relative features among the preset regions in the image may be relative features between the preset regions in the image and preset regions within a preset range surrounding said preset regions. For example, the relative features among the preset regions in the monocular image may be obtained by performing dot product operation on feature vectors between the preset regions in the monocular image and the preset  
15 regions within a preset range surrounding said preset regions.

According to one or more disclosure of the present disclosure, before performing feature extraction on the monocular image by means of a first neural network, downsampling is further performed on the monocular image to obtain a monocular image having a preset dimension, and the monocular image having a  
20 preset dimension is taken as the monocular image on which depth estimation is performed by the depth estimation neural network to reduce the computation amount and improve the data processing speed. The dimension of the monocular image is a multiple of the preset dimension. For example, the dimension of the monocular image is 8 times of the preset dimension.

25 Generally, the depth variation of the image in the vertical direction is greater than the depth variation in the horizontal direction. For example, in a driving scene, a road in an image always extends along the vertical direction to a place which is farther from the video camera. Therefore, the longitudinal variation law of the depth information of the image facilitates the estimation of an absolute distance of the  
30 image. Therefore, the longitudinal variation law of the depth information of the monocular image can be used for the depth estimation of the monocular image. For example, optimization can be performed on the predicated depth map according to the longitudinal variation law of the depth information of the monocular image.

In some embodiments, as shown in FIG. 2, after operation 204, i.e., obtaining a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image, the method further includes the following operations.

5           At 206, optimization is performed on the predicted depth image according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image.

          According to one or more disclosure of the present disclosure, residual estimation is performed on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map, and then optimization is performed on the predicted depth map according to the residual plot to obtain a target depth map of the monocular image.

10           In an optional example, residual estimation is performed on the predicted depth map by a residual estimation network according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map, and then pixel-by-pixel superposition operation is performed on the residual plot and the predicted depth map to obtain a target depth map of the monocular image.

15           According to one or more disclosure of the present disclosure, before performing optimization on the predicted depth image according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image, the longitudinal variation law of depth information of the monocular image is further obtained according to the predicted depth map.

20           In an optional example, processing is performed on the predicted depth map by means of a longitudinal pooling layer to obtain a longitudinal variation law of depth information of the monocular image. The longitudinal pooling layer takes a column vector as a pooling kernel to perform pooling processing on the predicted depth map. For example, the longitudinal pooling layer uses a pooling kernel having a size of  $H \times 1$  to perform average pooling processing on the predicted depth map, where  $H$  is an integer greater than 1.

25           The method for the depth estimation of a monocular image provided by the embodiments obtains a global feature of a monocular image according to absolute

features of preset regions and relative features among the preset regions in the monocular image on the basis of a depth estimation neural network, obtains a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image, and performs optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image. By means of the mutual supplementation between relative features and absolute features of preset regions in an image in the depth estimation of monocular images, the accuracy of the relative distance prediction in the depth estimation is improved, and by means of optimization performed according to a longitudinal variation law of depth information of the image, the accuracy of the absolute distance prediction in depth estimation is improved, such that the accuracy of the depth estimation of monocular images can be comprehensively improved.

In some embodiments, when downsampling is performed, before performing feature extraction on the monocular image by means of a first neural network, on the monocular image to obtain a monocular image having a preset dimension, and the monocular image having a preset dimension is taken as the monocular image on which depth estimation is performed by the depth estimation neural network, a multi-scale learning method is used to perform optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to improve the accuracy of depth estimation of the monocular image.

According to one or more disclosure of the present disclosure, a preset number of upsamplings are performed on the predicted depth map, a longitudinal variation law of depth information is obtained according to the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple, and optimization is performed on the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple according to the longitudinal variation law of the depth information thereof to obtain an optimized target depth map; where the optimized target depth map obtained by each of the upsamplings other than the last upsampling is taken as the predicted depth map of the next upsampling, the optimized target depth map obtained by the last upsampling is taken as the target

depth map of the monocular image, and the target depth map has the same dimension as the monocular image.

A flowchart of each scale optimization during performing optimization using multi-scale learning is described in details below with reference to FIG. 3.

5       As shown in FIG. 3, the method includes the following operations.

At 302, upsampling is performed on the predicted depth map having a first preset dimension to obtain a predicted depth map having a second preset dimension.

In the embodiments, the predicted depth map having a first preset dimension may be a predicted depth map obtained from a depth estimation neural network, and  
10       may also be an optimized target depth map obtained from a previous scale optimization process. The second preset dimension is a multiple of the first preset dimension, where the sizes of the first preset dimension and the second preset dimension are determined according to the number and frequency of upsamplings, the size of the monocular image, or the like.

15       In an optional example, upsampling is performed on the predicted depth map having a first preset dimension by means of an upsampling layer to obtain a predicted depth map having a second preset dimension.

At 304, a longitudinal variation law of corresponding depth information according to the predicted depth map having a second preset dimension.

20       In an optional example, processing is performed on the predicted depth map having a second preset dimension by means of a longitudinal pooling layer to obtain a longitudinal variation law of the corresponding depth information.

At 306, residual estimation is performed on the predicted depth map having a second preset dimension according to the longitudinal variation law of the  
25       corresponding depth information to obtain a corresponding residual map.

In an optional example, residual estimation is performed on the predicted depth map having a second preset dimension by a residual estimation network according to the longitudinal variation law of the corresponding depth information to obtain a corresponding residual map.

30       At 308, optimization is performed on the predicted depth map having the second preset dimension according to the corresponding residual map to obtain an optimized target depth map having the second preset dimension.

In an optional example, pixel-by-pixel superposition operation is performed on the corresponding residual plot and the predicted depth map having the second preset dimension to obtain an optimized target depth map having the second preset dimension.

5           FIGS. 4A to 4C are structural schematic diagrams of networks for implementing the method for the depth estimation of a monocular image according to some embodiments of the present disclosure.

          In the embodiments, as shown in FIG. 4A, a network for implementing the method for the depth estimation of a monocular image according to embodiments of  
10       the present disclosure includes: a convolutional neural network, a depth estimation neural network, and a depth optimization neural network. The convolutional neural network includes a downsampling layer and a first neural network. The downsampling layer performs 8-time downsampling on the monocular image to obtain a monocular image having a dimension which is  $1/8$  of the dimension of said  
15       monocular image, and then the first neural network performs feature extraction on the  $1/8$ -dimension monocular image to obtain absolute features of preset regions in the  $1/8$ -dimension monocular image.

          As shown in FIG. 4B, the depth estimation neural network includes: an association layer, a full connection layer, and a depth estimator, where the association  
20       layer obtains relative features among the preset regions in the  $1/8$ -dimension monocular image according to the absolute features of the preset regions in the  $1/8$ -dimension monocular image, the full connection layer obtains a global feature of the  $1/8$ -dimension monocular image according to the absolute features of the preset regions and the relative features among the preset regions in the  $1/8$ -dimension  
25       monocular image, and the depth estimator obtains a  $1/8$ -dimension predicted depth map according the global feature of the  $1/8$ -dimension image and the absolute features of the preset regions and the relative feature among the preset regions in the  $1/8$ -dimension monocular image.

          In the embodiments, the depth optimization neural network includes a first  
30       scale optimization network, a second scale optimization network, and a third scale optimization network, where the structure of each scale optimization network, as shown in FIG. 4C, includes an upsampling layer, a longitudinal pooling layer, a residual estimation network, and an additive operation unit.

The upsampling layer of the first scale optimization network performs 2-time upsampling on the  $1/8$ -dimension predicted depth map to obtain a  $1/4$ -dimension predicted depth map. The longitudinal pooling layer of the first scale optimization network obtains a longitudinal variation law of corresponding depth information according to the  $1/4$ -dimension predicted depth map. The residual estimation network of the first scale optimization network performs residual estimation on the  $1/4$ -dimension predicted depth map according to the longitudinal variation law of the depth information corresponding to the  $1/4$ -dimension predicted depth map to obtain a corresponding residual plot. The additive operation unit of the first scale optimization network performs pixel-by-pixel superposition operation on the corresponding residual plot and the  $1/4$ -dimension predicted depth map to obtain an optimized  $1/4$ -dimension target depth map, and takes the optimized  $1/4$ -dimension target depth map as the predicted depth map of the second scale optimization network.

The upsampling layer of the second scale optimization network performs 2-time upsampling on the optimized  $1/4$ -dimension target depth map to obtain a  $1/2$ -dimension predicted depth map. The longitudinal pooling layer of the second scale optimization network obtains a longitudinal variation law of corresponding depth information according to the  $1/2$ -dimension predicted depth map. The residual estimation network of the second scale optimization network performs residual estimation on the  $1/2$ -dimension predicted depth map according to the longitudinal variation law of the depth information corresponding to the  $1/2$ -dimension predicted depth map to obtain a corresponding residual plot. The additive operation unit of the second scale optimization network performs pixel-by-pixel superposition operation on the corresponding residual plot and the  $1/2$ -dimension predicted depth map to obtain an optimized  $1/2$ -dimension target depth map, and takes the optimized  $1/2$ -dimension target depth map as the predicted depth map of the third scale optimization network.

The upsampling layer of the third scale optimization network performs 2-time upsampling on the optimized  $1/2$ -dimension target depth map to obtain a predicted depth map having the same dimension as the monocular image. The longitudinal pooling layer of the third scale optimization network obtains a longitudinal variation law of corresponding depth information according to the predicted depth map having the same dimension as the monocular image. The residual estimation network of the third scale optimization network performs residual estimation on the predicted depth

map having the same dimension as the monocular image according to the longitudinal variation law of the depth information corresponding to the predicted depth map to obtain a corresponding residual plot. The additive operation unit of the third scale optimization network performs pixel-by-pixel superposition operation on the  
5 corresponding residual plot and the predicted depth map having the same dimension as the monocular image to obtain an optimized target depth map having the same dimension as the monocular image, and takes the optimized target depth map as the target depth map of the monocular image.

In an optional example, the depth estimation neural network of the forgoing  
10 embodiments is obtained by performing semi-supervised training using a dense depth map obtained by means of stereoscopic binocular image matching and a sparse depth map as marking data.

In the embodiments, because the "marking data" of training data obtained using other methods is sparse, i.e., there are a small number of valid pixel values in a  
15 depth map, a depth map obtained using binocular matching is taken as the "marking data" of the training data.

The method for the depth estimation of a monocular image provided by the embodiments of the present disclosure is used in the fields of geometric scene structure analysis, automatic driving, assisted driving, target tracking, automatic  
20 obstacle avoidance of robots, or the like. For example, in a driving scene, the method for the depth estimation of a monocular image provided by the embodiments of the present disclosure is used for predicting a distance from a front vehicle or a pedestrian. When photographing is performed by a mobile phone, a monocular blurring operation is performed using depth information predicted by the method for the depth  
25 estimation of a monocular image provided by the embodiments of the present disclosure; and the prediction result of the method for the depth estimation of a monocular image provided by the embodiments of the present disclosure facilitates improving an object tracking algorithm.

FIG. 5 is a schematic structural diagram of an apparatus for the depth  
30 estimation of a monocular image according to some embodiments of the present disclosure.

As shown in FIG. 5, the apparatus includes a depth estimation neural network  
510. The depth estimation neural network 510 is configured to obtain a global feature

of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image, and obtain a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image.

In the embodiments, the monocular image may be an image obtained from an image collection device, and may also be an image obtained from a storage apparatus. For example, the image collection device may be a still camera, a video camera, a scanner, or the like, and the storage apparatus may be a USB flash drive, an optical disk, a hard disk drive, or the like. The embodiments do not limit the obtaining mode of the monocular image. The absolute features of the preset regions in the monocular image are used for representing local appearances of the preset regions in the monocular image. For example, the absolute features include texture features, geometric features, or the like. The relative features among the preset regions in the monocular image are used for representing the differences between the local appearances of the preset regions in the monocular image. For example, the relative features include texture differences, geometric differences, or the like. The preset regions in the monocular image are set according to the features of the image. The depth map of the embodiments is a map that represents the distances between pixels in an image and the image collection device using the pixels values of the pixels in the image.

In an optional example, as shown in FIG. 5, the depth estimation neural network 510 includes: a full connection layer 511, configured to obtain a global feature of a monocular image by combining absolute features of preset regions and relative features among the preset regions in the monocular image. The depth estimation neural network 510 further includes: a depth estimator 512, configured to perform depth estimation according to the global feature of the monocular image and the absolute features of preset regions and relative features among the preset regions in the monocular image to obtain a predicted depth map of the monocular image. For example, the depth estimator may be a fully convolutional network which mainly consists of a convolution layer and a deconvolution layer. The depth estimator is able to calculate depths values of pixels in an image according to geometric distribution information of the image, i.e., a global feature of the image, and absolute features of

preset regions and relative features among the preset regions in the image, to obtain the predicted depth map.

The apparatus for the depth estimation of a monocular image provided by the embodiments obtains a global feature of a monocular image according to absolute  
5 features of preset regions and relative features among the preset regions in the monocular image on the basis of a depth estimation neural network, and obtains a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image. By means of the mutual supplementation between relative  
10 features and absolute features of preset regions in an image in the depth estimation of monocular images, the accuracy of the relative distance prediction in the depth estimation is improved, such that the accuracy of the depth estimation of monocular images can be improved.

FIG. 6 is a schematic structural diagram of an apparatus for the depth  
15 estimation of a monocular image according to other embodiments of the present disclosure;

As shown in FIG. 6, as compared with the embodiment of FIG. 5, the difference therebetween is that the apparatus further includes a first neural network 620..

20 The first neural network 620 is configured to perform feature extraction on the monocular image to obtain features of preset regions in the monocular image, and take the features of the preset regions as absolute features of the preset regions in the monocular image. For example, the first neural network may be a coder network consisting of a convolution layer and a pooling layer, and high-dimension features of  
25 the image can be obtained by performing feature extraction on the monocular image by means of the coder network.

The depth estimation neural network 610 is further configured to obtain relative features among the preset regions in the monocular image according to the absolute features of the preset regions in the monocular image.

30 In an optional example, as shown in FIG. 6, the depth estimation neural network 610 further includes: an association layer 613, configured to perform vector operation on the absolute features of the preset regions in the monocular image to obtain the relative features among the preset regions in the monocular image. The

relative features among the preset regions in the image may be relative features between the preset regions in the image and preset regions within a preset range surrounding said preset regions. For example, the relative features among the preset regions in the monocular image may be obtained by performing dot product operation  
5 on feature vectors between the preset regions in the monocular image and the preset regions within a preset range surrounding said preset regions.

According to one or more disclosure of the present disclosure, the apparatus further includes: a downsampling layer, configured to perform downsampling on the monocular image before performing feature extraction on the monocular image to  
10 obtain a monocular image having a preset dimension. In this case, the depth estimation neural network 610 performs depth estimation on the monocular image having a preset dimension to reduce the computation amount and improve the data processing speed. The dimension of the monocular image is a multiple of the preset dimension. For example, the dimension of the monocular image is 8 times of the  
15 preset dimension.

FIG. 7 is a schematic structural diagram of an apparatus for the depth estimation of a monocular image according to other embodiments of the present disclosure.

As shown in FIG. 7, as compared with the embodiment of FIG. 5, the  
20 difference therebetween is that the apparatus further includes a second neural network 730.

The second neural network 730 is configured to perform optimization on the predicted depth image according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image.

According to one or more disclosure of the present disclosure, the second  
25 neural network 730 is configured to perform residual estimation on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map, and then perform optimization on the predicted depth map according to the residual plot to  
30 obtain a target depth map of the monocular image.

In an optional example, as shown in FIG. 7, the second neural network 730 includes: a residual estimation network 731, configured to perform residual estimation on the predicted depth map according to a longitudinal variation law of depth

information of the monocular image to obtain a residual plot of the predicted depth map, and an additive operation unit 732, configured to perform pixel-by-pixel superposition operation on the residual plot and the predicted depth map to obtain a target depth map of the monocular image.

5           According to one or more disclosure of the present disclosure, the second neural network 730 is further configured to obtain a longitudinal variation law of depth information of the monocular image according to the predicted depth map.

          In an optional example, as shown in FIG. 7, the second neural network 730 further includes: a longitudinal pooling layer 733, configured to perform processing  
10   on the predicted depth map to obtain a longitudinal variation law of depth information of the monocular image. The longitudinal pooling layer takes a column vector as a pooling kernel to perform pooling processing on the predicted depth map. For example, the longitudinal pooling layer uses a pooling kernel having a size of  $H \times 1$  to perform average pooling processing on the predicted depth map, where  $H$  is an integer  
15   greater than 1.

          The apparatus for the depth estimation of a monocular image provided by the embodiment obtains a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image on the basis of a depth estimation neural network, obtains a  
20   predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image, and performs optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image. By means of

25           the mutual supplementation between relative features and absolute features of preset regions in an image in the depth estimation of monocular images, the accuracy of the relative distance prediction in the depth estimation is improved, and by means of optimization performed according to a longitudinal variation law of depth information of the image, the accuracy of the absolute distance prediction in depth  
30   estimation is improved, such that the accuracy of the depth estimation of monocular images can be comprehensively improved.

          In some embodiments, when downsampling is performed, before performing feature extraction on the monocular image by means of the first neural network, on

the monocular image by means of the downsampling layer to obtain a monocular image having a preset dimension, and the monocular image having a preset dimension is taken as the monocular image on which depth estimation is performed by the depth estimation neural network, a multi-scale learning method is used to perform optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to improve the accuracy of depth estimation of the monocular image.

According to one or more disclosure of the present disclosure, the apparatus further includes: an upsampling layer, configured to perform a preset number of upsamplings on the predicted depth map; a longitudinal pooling layer, configured to obtain a longitudinal variation law of depth information according to the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple; and the second neural network, configured to perform optimization on the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple according to the longitudinal variation law of the depth information thereof to obtain an optimized target depth map. The optimized target depth map obtained by each of the upsamplings other than the last upsampling is taken as the predicted depth map of the next upsampling, the optimized target depth map obtained by the last upsampling is taken as the target depth map of the monocular image, and the target depth map has the same dimension as the monocular image.

In an optional example, the depth estimation neural network of the forgoing embodiments is obtained by performing semi-supervised training using a dense depth map obtained by means of stereoscopic binocular image matching and a sparse depth map as marking data.

In an optional example, because the "marking data" of training data obtained using other methods is sparse, i.e., there are a small number of valid pixel values in a depth map, a depth map obtained using stereoscopic binocular image matching is taken as the "marking data" of the training data. The embodiments of the present disclosure further provide an electronic device which, for example, is a mobile terminal, a Personal Computer (PC), a tablet computer, a server, or the like. Referring to FIG. 8 below, a schematic structural diagram of an electronic device 800, which may be a terminal device or a server, suitable for implementing the embodiments of

the present disclosure is shown. As shown in FIG. 8, the electronic device 800 includes one or more processors, a communication part, or the like. The one or more processors are, for example, one or more Central Processing Units (CPUs) 801 and/or one or more Graphic Processing Units (GPUs) 813, and may execute appropriate actions and processing according to executable instructions stored in a Read-Only Memory (ROM) 802 or executable instructions loaded from a storage section 808 to a Random Access Memory (RAM) 803. The communication part 812 may include, but not limited to, a network card. The network card may include, but not limited to, an Infiniband (IB) network card. The processor may communicate with the ROM 802 and/or the RAM 830, to execute executable instructions. The processor is connected to the communication part 812 via a bus 804, and communicates with other target devices via the communication part 812, thereby implementing corresponding operations of any method provided in the embodiments of the present disclosure, for example, on the basis of a depth estimation neural network, obtaining a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image, and obtaining a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image.

In addition, the RAM 803 further stores various programs and data required for operations of an apparatus. The CPU 801, the ROM 802, and the RAM 803 are connected to each other by means of the bus 804. In the case that the RAM 803 exists, the ROM 802 is an optional module. The RAM 803 stores executable instructions, or writes the executable instructions into the ROM 802 during running, where the executable instructions cause the CPU 801 to execute corresponding operations of the foregoing communication method. An Input/Output (I/O) interface 805 is also connected to the bus 804. The communication part 812 may be integrated, or may be configured to have multiple sub-modules (for example, multiple IB network cards) connected to the bus.

The following components are connected to the I/O interface 805: an input section 806 including a keyboard, a mouse and the like; an output section 807 including a Cathode-Ray Tube (CRT), a Liquid Crystal Display (LCD), a speaker and the like; a storage section 808 including a hard disk and the like; and a

communication section 809 of a network interface card including an LAN card, a modem and the like. The communication section 809 performs communication processing via a network such as the Internet. A drive 810 is also connected to the I/O interface 805 according to requirements. A removable medium 811 such as a magnetic disk, an optical disk, a magneto-optical disk, and a semiconductor memory is installed on the drive 810 according to requirements, so that a computer program read from the removable medium is installed on the storage section 808 according to requirements.

It should be noted that the architecture illustrated in FIG. 8 is merely an optional implementation mode. During specific practice, the number and types of the components in FIG. 8 may be selected, decreased, increased, or replaced according to actual requirements. Different functional components may be separated or integrated or the like. For example, the GPU 813 and the CPU 801 may be separated, or the GPU 813 may be integrated on the CPU 801, and the communication part may be separated from or integrated on the CPU 801 or the GPU 813 or the like. These alternative implementations all fall within the scope of protection of the present disclosure.

Particularly, the process described above with reference to the flowchart according to the embodiments of the present disclosure may be implemented as a computer software program. For example, the embodiments of the present disclosure include a computer program product, which includes a computer program tangibly included in a machine-readable medium. The computer program includes a program code for executing a method shown in the flowchart. The program code may include corresponding instructions for correspondingly executing operations of the methods provided by the embodiments of the present disclosure, for example, on the basis of a depth estimation neural network, obtaining a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image, and obtaining a predicted depth map of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image. In such embodiments, the computer program is downloaded and installed from the network through the communication section 809, and/or is installed from the

removable medium 811. The computer program, when being executed by the CPU 801, executes the foregoing functions defined in the methods of the present disclosure.

In one or more optional implementations, the embodiments of the present disclosure also provide a computer program product configured to store computer-readable instructions, where when the instructions are executed, a computer executes the image recovery method in any of the possible embodiments.

The computer program product is specifically implemented by means of hardware, software, or a combination thereof. In an optional example, the computer program product is specifically represented by a computer storage medium. In another optional example, the computer program product is represented by a software product, such as a Software Development Kit (SDK).

In one or more optional implementations, the embodiments of the present disclosure also provide a method for the depth estimation of a monocular image and an apparatus corresponding thereto, an electronic device, a computer storage medium, a computer program, and a computer program product. The method includes: sending, by a first apparatus, an instruction for the depth estimation of a monocular image to a second apparatus, where the instruction causes the second apparatus to execute the method for the depth estimation of a monocular image according to any one of the possible embodiments; and receiving, by the first apparatus, a result of the depth estimation of the monocular image sent by the second apparatus.

In some embodiments, the instruction for the depth estimation of a monocular image is specifically an invoking instruction. The first apparatus instructs, by means of invoking, the second apparatus to execute the depth estimation of a monocular image. Accordingly, in response to receiving the invoking instruction, the second apparatus executes the operations and/process in any of the embodiments of the method for the depth estimation of a monocular image.

It should be understood that the terms such as "first" and "second" in the embodiments of the present invention are only used for distinguishing, and shall not be understood as limitations on the embodiments of the present invention.

It should also be understood that, in the present invention, "multiple" may refer to two or more, and "at least one" may refer to one, two or more.

It should also be understood that, for any component, data or structure mentioned in the present disclosure, if there is no explicit limitation or no opposite

motivation is provided in context, it is generally understood that the number of the component, data or structure is one or more.

It should also be understood that, the descriptions of the embodiments in the present disclosure focus on differences between the embodiments, and for same or  
5 similar parts in the embodiments, refer to these embodiments. For the purpose of brevity, details are not described again.

The methods and apparatuses in the present disclosure may be implemented in many manners. For example, the methods and apparatuses in the present disclosure may be implemented with software, hardware, firmware, or any combination of  
10 software, hardware, and firmware. The foregoing specific sequence of operations of the method is merely for description, and unless otherwise stated particularly, is not intended to limit the operations of the method in the present disclosure. In addition, in some embodiments, the present disclosure is also implemented as programs recorded in a recording medium. The programs include machine-readable instructions for  
15 implementing the methods according to the present disclosure. Therefore, the present disclosure further covers the recording medium storing the programs for performing the methods according to the present disclosure.

The descriptions of the present disclosure are provided for the purpose of examples and description, and are not intended to be exhaustive or limit the present  
20 disclosure to the disclosed form. Many modifications and changes are obvious to a person of ordinary skill in the art. The embodiments are selected and described to better describe a principle and an actual application of the present disclosure, and to make a person of ordinary skill in the art understand the present disclosure, so as to design various embodiments with various modifications applicable to particular use.

25

## CLAIMS

1. A method for depth estimation of a monocular image, comprising:
  - 5 obtaining, through a depth estimation neural network, a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image; and  
obtaining a predicted depth map of the monocular image according to the global feature, the absolute features of preset regions and relative features among the  
10 preset regions in the monocular image.
2. The method according to claim 1, before the obtaining a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image, further comprising:
  - 15 performing, through a first neural network, feature extraction on the monocular image to obtain features of preset regions in the monocular image, and taking the features of the preset regions as the absolute features of the preset regions in the monocular image; and  
obtaining the relative features among the preset regions in the monocular  
20 image according to the absolute features of the preset regions in the monocular image.
3. The method according to claim 2, wherein the obtaining relative features among the preset regions in the monocular image according to the absolute features of the preset regions in the monocular image comprises:
  - 25 performing, through an association layer, a vector operation on the absolute features of the preset regions in the monocular image to obtain the relative features among the preset regions in the monocular image.
4. The method according to claim 2 or 3, before performing, through a first neural  
30 network, feature extraction on the monocular image, further comprising:
  - performing downsampling on the monocular image to obtain a monocular image having a preset dimension, wherein a dimension of the monocular image is a multiple of the preset dimension.

5. The method according to any one of claims 1 to 4, wherein the obtaining a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image comprises:

obtaining, through a full connection layer, the global feature of the monocular  
5 image by combining the absolute features of the preset regions and the relative features among the preset regions in the monocular image .

6. The method according to any one of claims 1 to 5, wherein the obtaining a predicted depth map of the monocular image according to the global feature and the  
10 absolute features of preset regions and relative features among the preset regions in the monocular image comprises:

performing, through a depth estimator, depth estimation according to the global feature, the absolute features of the preset regions and the relative features among the preset regions in the monocular image to obtain the predicted depth map of  
15 the monocular image.

7. The method according to any one of claims 1 to 6, after the obtaining a predicted depth map of the monocular image according to the global feature, the absolute features of preset regions and relative features among the preset regions in the  
20 monocular image, further comprising:

performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image.

25 8. The method according to claim 7, wherein the performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image comprises:

performing residual estimation on the predicted depth map according to the longitudinal variation law of depth information of the monocular image to obtain a  
30 residual plot of the predicted depth map; and

performing optimization on the predicted depth map according to the residual plot to obtain the target depth map of the monocular image.

9. The method according to claim 8, wherein the performing residual estimation on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map comprises:

- 5 performing, through a residual estimation network, residual estimation on the predicted depth map according to the longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map; and
- the performing optimization on the predicted depth map according to the residual plot to obtain a target depth map of the monocular image comprises
- 10 performing a pixel-by-pixel superposition operation on the residual plot and the predicted depth map to obtain the target depth map of the monocular image.

10. The method according to any one of claims 7 to 9, before the performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the

15 monocular image, further comprising:

obtaining the longitudinal variation law of depth information of the monocular image according to the predicted depth map.

20 11. The method according to claim 10, wherein the obtaining a longitudinal variation law of depth information of the monocular image according to the predicted depth map comprises:

- performing, through a longitudinal pooling layer, processing on the predicted depth map to obtain the longitudinal variation law of depth information of the
- 25 monocular image.

12. The method according to claim 7, wherein the performing optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image comprises:

- 30 performing a preset number of upsamplings on the predicted depth map, obtaining the longitudinal variation law of depth information according to a predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple, and performing optimization on the predicted depth map, obtained by

each upsampling, having a dimension sequentially increased by a multiple according to the obtained longitudinal variation law of the depth information to obtain an optimized target depth map;

5 wherein the optimized target depth map obtained by each of the upsamplings other than a last upsampling is taken as a predicted depth map of a next upsampling, the optimized target depth map obtained by the last upsampling is taken as the target depth map of the monocular image, and the target depth map has the same dimension as the monocular image.

10 13. The method according to any one of claims 1 to 12, wherein the depth estimation neural network comprises an association layer, a full connection layer, and a depth estimator, and is obtained by training the depth estimation neural network by using a sparse depth map and a dense depth map obtained by stereo matching for binocular images as marking data.

15

14. An apparatus for the depth estimation of a monocular image, comprising:

a depth estimation neural network, configured to obtain a global feature of a monocular image according to absolute features of preset regions and relative features among the preset regions in the monocular image, and obtain a predicted depth map  
20 of the monocular image according to the global feature and the absolute features of preset regions and relative features among the preset regions in the monocular image.

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

a first neural network, configured to perform feature extraction on the monocular image to obtain features of preset regions in the monocular image, and  
25 take the features of the preset regions as the absolute features of the preset regions in the monocular image; and

the depth estimation neural network, further configured to obtain relative features among the preset regions in the monocular image according to the absolute  
30 features of the preset regions in the monocular image.

16. The apparatus according to claim 15, wherein the depth estimation neural network comprises:

an association layer, configured to perform a vector operation on the absolute features of the preset regions in the monocular image to obtain the relative features among the preset regions in the monocular image.

5 17. The apparatus according to claim 15 or 16, further comprising:

a downsampling layer, configured to perform downsampling on the monocular image before performing feature extraction on the monocular image to obtain a monocular image having a preset dimension, wherein a dimension of the monocular image is a multiple of the preset dimension.

10

18. The apparatus according to any one of claims 14 to 17, wherein the depth estimation neural network comprises:

a full connection layer, configured to obtain the global feature of the monocular image by combining the absolute features of the preset regions and the relative features among the preset regions in the monocular image.

15

19. The apparatus according to any one of claims 14 to 18, wherein the depth estimation neural network comprises:

a depth estimator, configured to perform depth estimation according to the global feature and the absolute features of the preset regions and the relative features among the preset regions in the monocular image to obtain the predicted depth map of the monocular image.

20

20. The apparatus according to any one of claims 14 to 19, further comprising:

a second neural network, configured to perform optimization on the predicted depth map according to a longitudinal variation law of depth information of the monocular image to obtain a target depth map of the monocular image.

25

21. The apparatus according to claim 20, wherein the second neural network is configured to perform residual estimation on the predicted depth map according to the longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map, and perform optimization on the predicted

30

depth map according to the residual plot to obtain the target depth map of the monocular image.

22. The apparatus according to claim 21, wherein the second neural network  
5 comprises:

a residual estimation network, configured to perform residual estimation on the predicted depth map according to the longitudinal variation law of depth information of the monocular image to obtain a residual plot of the predicted depth map; and

10 an additive operation unit, configured to perform a pixel-by-pixel superposition operation on the residual plot and the predicted depth map to obtain the target depth map of the monocular image.

23. The apparatus according to any one of claims 20 to 22, wherein the second neural  
15 network is further configured to obtain the longitudinal variation law of depth information of the monocular image according to the predicted depth map.

24. The apparatus according to claim 23, wherein the second neural network comprises:

20 a longitudinal pooling layer, configured to perform processing on the predicted depth map to obtain the longitudinal variation law of depth information of the monocular image.

25. The apparatus according to claim 20, further comprising:

25 an upsampling layer, configured to perform a preset number of upsamplings on the predicted depth map;

a longitudinal pooling layer, configured to obtain the longitudinal variation law of depth information according to the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple; and

30 the second neural network, configured to perform optimization on the predicted depth map, obtained by each upsampling, having a dimension sequentially increased by a multiple according to the obtained longitudinal variation law of the depth information to obtain an optimized target depth map;

wherein the optimized target depth map obtained by each of the upsamplings other than a last upsampling is taken as a predicted depth map of a next upsampling, the optimized target depth map obtained by the last upsampling is taken as the target depth map of the monocular image, and the target depth map has the same dimension as the monocular image.

26. The apparatus according to any one of claims 14 to 25, wherein the depth estimation neural network comprises an association layer, a full connection layer, and a depth estimator, and is obtained by training the depth estimation neural network by using a sparse depth map and a dense depth map obtained by stereo matching for binocular images as marking data.

27. An electronic device, comprising the apparatus according to any one of claims 14 to 26.

28. An electronic device, comprising:  
a memory, configured to store executable instructions; and  
a processor, configured to execute the executable instructions to implement the method according to any one of claims 1 to 13.

29. A computer program, comprising a computer-readable code that, when being executed in a device, cause a processor in the device to execute instructions for implementing the method according to any one of claims 1 to 13.

30. A computer storage medium, configured to store computer readable instructions that, when being executed, implement the method according to any one of claims 1 to 13.