

METHOD FOR ESTIMATING MONOCULAR DEPTH, APPARATUS AND DEVICE THEREFOR, AND STORAGE MEDIUM

CROSS REFERENCE TO RELATED APPLICATION

The present application claims priority to Chinese Patent Application No. 201810496541.6 filed on May 22, 2018, the disclosure of which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

Embodiments of the present application relate to the field of artificial intelligence, and in particular, to a method and apparatus for estimating a monocular depth, a device and a storage medium.

BACKGROUND

Monocular depth estimation is an important issue in computer vision. The specific task of monocular depth estimation refers to predicting the depth of each pixel in a picture. A picture composed of depth values of all pixels is also called a depth map. Monocular depth estimation is of great significance for obstacle detection, 3D scene reconstruction, and scene stereoscopic analysis in autonomous driving. In addition, monocular depth estimation can indirectly improve the performance of other computer vision tasks, such as object detection, target tracking, and target recognition.

The current problem is that training neural networks for monocular depth estimation requires a large amount of annotated data, but the cost for obtaining the annotated data is high. Annotated data in an outdoor environment can be obtained through a laser radar. However, the obtained annotated data is very sparse. A monocular depth estimation network trained with such annotated data has no clear edge and cannot capture the correct depth information of small objects.

SUMMARY

Embodiments of the present application provide a method and an apparatus for estimating a monocular depth, a device and a storage medium.

The technical solutions of the embodiments of the present application are implemented as follows.

The embodiments of the present application provide a method for estimating a monocular depth, including: obtaining an image to be processed; inputting the image to be processed into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, where the monocular depth estimation network model is trained with a disparity map output by a first binocular matching neural network model as supervision; and outputting the analysis result of the image to be processed.

The embodiments of the present application provide an apparatus for estimating a monocular depth, including: an obtaining module, an execution module, and an output module, where the obtaining module is configured to obtain an image to be processed; the execution module is configured to input the image to be processed into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, where the monocular depth estimation network model is trained by supervision of a disparity map output by a first binocular matching neural network model; and the output module is configured to output the analysis result of the image to be processed.

The embodiments of the present application provide a device for estimating a monocular depth, including a memory and a processor, where the memory stores computer programs capable of running in the processor, and when the programs being executed, the processor implements steps of the method for estimating a monocular depth provided by the embodiments of the present application are implemented.

The embodiments of the present application provide a computer readable storage medium, having stored therein computer programs that, when being executed by a processor, implement steps of the method for estimating a monocular depth provided by the embodiments of the present application are implemented.

In the embodiments of the present application, by obtaining an image to be processed, inputting the image to be processed into a trained monocular depth estimation network model to obtain an analysis result of the image to be processed, where the monocular depth estimation

network model is trained with a disparity map output by a first binocular matching neural network model as supervision, and outputting the analysis result of the image to be processed, it is possible to train the monocular depth estimation network with less or no data having a depth map tag, and a more efficient method for unsupervised fine-tuning binocular disparity networks is proposed, thereby indirectly improving the effect of monocular depth estimation.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1A is a first schematic flowchart for implementing a method for monocular depth estimation according to embodiments of the present application;

FIG. 1B is a schematic diagram for depth estimation of a single picture according to embodiments of the present application;

FIG. 1C is a schematic diagram of training a second binocular matching neural network model according to embodiments of the present application.

FIG. 1D is a schematic diagram for training a monocular depth estimation network model according to embodiments of the present application.

FIG. 1E is a schematic diagram of a loss function related picture according to embodiments of the present application.

FIG. 2A is a second schematic flowchart for implementing a method for estimating a monocular depth according to embodiments of the present application.

FIG. 2B is a schematic diagram of the loss function effect according to embodiments of the present application.

FIG. 2C is a schematic diagram of a visual depth estimation result according to embodiments of the present application.

FIG. 3 is a schematic structural diagram of an apparatus for estimating a monocular depth according to embodiments of the present application.

FIG. 4 is a schematic diagram of a hardware entity of a monocular depth estimation device according to embodiments of the present application.

DETAILED DESCRIPTION

To make the objectives, technical solutions, and advantages of embodiments of the present invention clearer, the following further describes in detail the specific technical solutions of the present invention with reference to the accompanying drawings in the embodiments of the present invention. The following embodiments are intended to illustrate the present application, but are not intended to limit the scope of the present application.

In the following description, the suffixes such as "module", "component" or "unit" used to represent an element are merely illustrative for the present application, and have no particular meaning *per se*. Therefore, "module", "component" or "unit" can be used in combination.

In general, the depth neural network is used to predict the depth map of a single picture, and only one picture is needed to perform three-dimensional modeling on a scene corresponding to the picture to obtain the depth of each pixel. The method for monocular depth estimation proposed by the embodiments of the present application is obtained by training of a neural network, and the training data is derived from the disparity map data of the binocular matching output, without requiring an expensive depth acquisition device such as a laser radar. The binocular matching algorithm for providing the training data is also implemented by means of a neural network, which can achieve a good effect by pre-training a large number of virtual binocular image pairs rendered by a rendering engine, and can further perform fine-tune training on the real data to achieve a better result.

The technical solutions of the present application are further described below in detail with reference to the accompanying drawings and embodiments.

The embodiments of the present application provide a method for estimating a monocular depth, and the method is applied to a computing device. The function implemented by the method can be implemented by invoking a program code by a processor in a server. Certainly, the program code may be saved in a computer storage medium. In view of the above, the server includes at least a processor and a storage medium. FIG. 1A is a schematic flowchart 1 for implementing a method for monocular depth estimation according to embodiments of the present application. As shown in FIG. 1A, the method includes the following steps.

At step S101, an image to be processed is obtained.

Here, the image to be processed may be obtained by means of a mobile terminal, and the image to be processed may include a picture of any scene. In general, the mobile terminal is any type of device having information processing capability in the process of implementation, for example, the mobile terminal includes a mobile phone, a Personal Digital Assistant (PDA), a navigator, a digital phone, a video phone, a smart watch, a smart bracelet, a wearable device, and

a tablet computer, etc. In the process of implementation, the server is a computing device having information processing capability such as a mobile terminal, e. g. , a mobile phone, a tablet computer, or a notebook computer, and a fixed terminal e. g. , a personal computer or a server cluster, and the like.

At step S102, the image to be processed is input into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, where the monocular depth estimation network model is trained with by supervision of a disparity map output by a first binocular matching neural network model.

In the embodiments of the present application, the monocular depth estimation network model is obtained mainly by the following three steps: step 1: pre-training a binocular matching neural network by using the synthetic binocular data rendered by the rendering engine; step 2: using data of a real scene to perform fine-tune training on the binocular matching neural network obtained in step 1; and step 3: providing supervision for the monocular depth estimation network by using the binocular matching neural network obtained in step 2, thereby obtaining a trained monocular depth estimation network. In the prior art, monocular depth estimation generally uses a large amount of real tag data for training, or an unsupervised method is used to train the monocular depth estimation network. However, the cost for obtaining a large amount of real tag data is quite high, and directly using the unsupervised method to obtain the monocular depth estimation network cannot deal with depth estimation of an occlusion region, and the resulting effect is relatively poor. While the sample data of the monocular depth estimation network model described in the present application is derived from the disparity map output by the first binocular matching neural network model. That is to say, the present application utilizes binocular disparity to guide the prediction of the monocular depth. Therefore, the method in the present application does not require a large amount of tag data, and a better training effect can be obtained.

At step S103, the analysis result of the image to be processed is output. Here, the analysis result of the image to be processed refers to a depth map corresponding to the image to be processed. The image to be processed, after being obtained, is input to the trained monocular depth estimation network model, and the monocular depth estimation network model generally outputs a disparity map corresponding to the image to be processed, rather than a depth map. Therefore, it is also required to determine a depth map corresponding to the image to be processed according to the disparity map output by the monocular depth estimation network model, a lens baseline distance of a camera that captures the image to be processed, and a lens focus of the camera that captures the image to be processed.

FIG. 1B is a schematic diagram of depth estimation of a single picture according to embodiments of the present application. As shown in FIG. 1B, a picture 11 with the reference numeral of 11 is the image to be processed, and a picture 12 with the reference numeral of 12 is the depth map corresponding to the picture 11 with the reference numeral of 11.

In practical applications, a ratio of a product of the lens baseline distance and the lens focus to a disparity map corresponding to the output image to be processed is determined as a depth map corresponding to the image to be processed.

Based on the foregoing method embodiments, the embodiments of the present application further provide a method for monocular depth estimation, including the following steps.

At step S111, a synthesized binocular picture with a depth tag is obtained as the synthetic sample data, where the synthesized binocular picture includes a synthesized left picture and a synthesized right picture.

In some embodiments, the method further includes: step S11, constructing a virtual 3D scene by means of the rendering engine; step S12, mapping the 3D scene into a binocular picture by means of two virtual cameras; step S13, obtaining the depth data of the synthesized binocular picture according to the position where the virtual 3D scene is constructed, the direction in which the virtual 3D scene is constructed, and the lens focus of the virtual camera; and step S14, annotating the binocular picture according to the depth data to obtain the synthesized binocular picture.

At step S112, a second binocular matching neural network model is trained according to the obtained synthetic sample data.

Here, in actual application, step S112 is implemented by means of the following step. At step S1121, the second binocular matching neural network model is trained according to the synthesized binocular picture to obtain the trained second binocular matching neural network model, where outputs of the trained second binocular matching neural network model are a disparity map and an occlusion map, the disparity map describing a disparity distance between each pixel point in the left picture and a corresponding pixel point in the right picture, the disparity distance being in units of pixels, and the occlusion map describing whether a corresponding pixel point in the right picture of each pixel point in the left picture is occluded by an object.

FIG. 1C is a schematic diagram of training the second binocular matching neural network model according to embodiments of the present application. As shown in FIG. 1C, a picture 11 with the reference numeral of 11 is a left picture of the synthesized binocular picture, a picture

12 with the reference numeral of 12 is a right picture of the synthesized binocular picture, I^L is a pixel value of all the pixels included in the left picture 11 with the reference numeral of 11, I^R is a pixel value of all the pixels included in the right picture 12 with the reference numeral of 12, a picture 13 with the reference numeral of 13 is an occlusion map output after training the second binocular matching neural network model, a picture 14 with the reference numeral of 14 is a disparity map output after training the second binocular matching neural network model, and a picture 15 with the reference numeral of 15 is the second binocular matching neural network model.

At step S113, a parameter of the trained second binocular matching neural network model is adjusted according to the obtained real sample data, to obtain the first binocular matching neural network model.

Here, step S113 is implemented in two ways, where the first implementation is implemented according to the following step. At step S1131a, supervised training is performed on the trained second binocular matching neural network model according to the obtained real binocular data with a depth tag, to adjust a weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model. Here, the obtained real binocular data with a depth tag is obtained, so that supervised training is directly performed on the second binocular matching neural network trained in step S112 by using the real binocular data with a depth tag, to adjust the weight of the trained second binocular matching neural network model, and to further improve the effect of the trained second binocular matching neural network model, so as to obtain the first binocular matching neural network model. In this section, the binocular disparity network needs to adapt the real data. The real binocular data with a depth tag is used, and fine-tune training is directly performed on the binocular disparity network by means of supervised training to adjust the network weights. The second implementation is implemented according to the following step. At step S1131b, unsupervised training is performed on the trained second binocular matching neural network model according to the obtained real binocular data without a depth tag, to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model. In the embodiments of the present application, unsupervised training is performed on the trained second binocular matching neural network model according to the real binocular data without a depth tag, to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model. Unsupervised training here refers to the use of binocular data for training without a deep data tag, which can be implemented using an unsupervised fine-tuning method.

At step S114, the monocular depth estimation network model is supervised by using the disparity map output by the first binocular matching neural network model, so as to train the monocular depth estimation network model.

Here, step S114 is implemented in two ways, where the first implementation is implemented according to the following steps. At step S1141a, a left or right picture in the real binocular data with depth tag is obtained as a training sample, where the real binocular data with depth tag includes a left picture and a right picture. At step S1142a, the monocular depth estimation network model is trained according to the left or right picture in the real binocular data with depth tag. Here, the depth neural network is used to predict the depth map of a single picture, and only one picture is needed to perform three-dimensional modeling on a scene corresponding to the picture to obtain the depth of each pixel. Therefore, the monocular depth estimation network model is trained according to the left or right picture in the real binocular data with a depth tag, where the real binocular data with a depth tag is the real binocular data with a depth tag depth used in step S1131a. The second implementation is implemented according to the following steps. At step S1141b, the real binocular data without a depth tag is input into the first binocular matching neural network model to obtain a corresponding disparity map, where the real binocular data without a depth tag includes a left picture and a right picture. At step S1142b, a depth map corresponding to the disparity map is determined according to the corresponding disparity map, a lens baseline distance of a camera that captures the real binocular data without a depth tag, and a lens focus of the camera that captures the real binocular data without a depth tag. At step S1143b, the left or right picture in the real binocular data without a depth tag is used as sample data, and the monocular depth estimation network model is supervised according to the depth map corresponding to the disparity map, so as to train the monocular depth estimation network model. Here, the depth neural network is used to predict the depth map of a single picture, and only one picture is needed to perform three-dimensional modeling on a scene corresponding to the picture to obtain the depth of each pixel. Therefore, the left or right picture in the real binocular data without a depth tag used in step S1131b is used as the sample data, the left or right picture in the real binocular data without a depth tag used in step S1141b is also used as the sample data, and the monocular depth estimation network model is supervised according to the depth map corresponding to the disparity map output in step S1141b, to train the monocular depth estimation network model to obtain a trained monocular depth estimation network model.

FIG. 1D is a schematic diagram for training a monocular depth estimation network model according to embodiments of the present application. FIG. (a) in FIG. 1D shows that the real

binocular data without a depth tag is input to the first binocular matching neural network model, to obtain a corresponding disparity map picture 13 with the reference numeral of 13, where the real binocular data without a depth tag includes a left picture 11 with the reference numeral of 11 and a right picture 12 with the reference numeral of 12, and a picture 15 with the reference numeral of 15 is the first binocular matching neural network model. FIG. (b) in FIG. 1D shows the left or right picture in the real binocular data without a depth tag as the sample data, and the monocular depth estimation network model is supervised according to a depth map corresponding to a disparity map picture 13 with the reference numeral of 13, so as to train the monocular depth estimation network model, where the output of the sample data by means of the monocular depth estimation network model is a disparity map picture 14 with the reference numeral of 14, and a picture 16 with the reference numeral of 16 is the monocular depth estimation network model.

At step S115, an image to be processed is obtained.

Here, after obtaining the trained monocular depth estimation network model, the monocular depth estimation network model can be used. That is, the depth map corresponding to the image to be processed is obtained by using the monocular depth estimation network model.

At step S116, the image to be processed is input into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, where the monocular depth estimation network model is trained with a disparity map output by a first binocular matching neural network model as supervision.

At step S117, the analysis result of the image to be processed is output, where the analysis result of the image to be processed includes a disparity map output by the monocular depth estimation network model.

At step S118, a depth map corresponding to the disparity map is determined according to the disparity map output by the monocular depth estimation network model, the lens baseline distance of the camera that captures a picture input into the monocular depth estimation network model, and the lens focus of the camera that captures the picture input into the monocular depth estimation network model.

At step S119, the depth map corresponding to the disparity map is output.

Based on the foregoing method embodiments, the embodiments of the present application further provide a method for monocular depth estimation, including the following steps.

At step S121, a synthesized binocular picture with a depth tag is obtained as the synthetic

sample data, where the synthesized binocular picture includes a synthesized left picture and a synthesized right picture.

At step S122, a second binocular matching neural network model is trained according to the obtained synthetic sample data.

Here, the use of synthetic data for training the second binocular matching neural network model has better generalization ability.

At step S123, the loss function is determined using formula (1): $L_{stereo-un\sup ft} = L_{photo} + \gamma_1 L_{abs} + \gamma_2 L_{rel}$ (1), where $L_{stereo-un\sup ft}$ represents the loss function proposed by the embodiments of the present application, L_{photo} represents a reconstruction error, L_{abs} represents that the disparity map output by the first binocular matching network model is less deviated from the disparity map output by the trained second binocular matching network model, L_{rel} represents that an output gradient of the first binocular matching network model is constrained to be consistent with an output gradient of the trained second binocular matching network model, and γ_1 and γ_2 represent strength coefficients. Here, L_{abs} and L_{rel} are regular terms.

In some embodiments, the formula (1) in step S123 can also be refined by means of the formula in the following steps. That is, the method further includes: step S1231, determining the reconstruction error by using formula (2) or formula (3):

$$L_{photo} = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^L - I_{ij}^{wR}| = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^L - I_{i(j-\bar{D}_{ij}^L)}^R| \quad (2), \quad \text{and}$$

$$L_{photo} = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^R - I_{ij}^{wL}| = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^R - I_{i(j+\bar{D}_{ij}^R)}^L| \quad (3),$$

where N represents the number of pixels in the picture, $\bar{M}_{old(ij)}$ represents a pixel value of the occlusion map output by the trained second binocular matching network model, I_{ij}^L represents a pixel value of the left picture in the real binocular data without a depth tag, I_{ij}^R represents a pixel value of the right picture in the real binocular data without a depth tag, I_{ij}^{wR} represents a pixel value of a picture synthesized after sampling the right picture, i.e., the reconstructed left picture, I_{ij}^{wL} represents a pixel value of a picture synthesized after sampling the left picture, i. e. , the reconstructed right picture, $\bar{D}_{ij}^L$ represents a pixel value of the disparity map output by the first binocular matching network model of the left picture in the real binocular data without a depth tag, $\bar{D}_{ij}^R$ represents a pixel value of the disparity map output by the first binocular matching network model of the right picture in the real binocular data without a depth tag, ij represents pixel coordinates of the pixel point, old represents the output of the trained second binocular matching network model,

R represents the right picture or the related data thereof, L represents the left picture or the related data thereof, and I represents a RGB (Red Green Blue) value of the picture pixel point. At step S1212, the disparity map output by the first binocular matching network model is determined to be less deviated from the disparity map output by the trained second binocular matching network model by using formula (4) or formula (5): $L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^L - \bar{D}_{old(ij)}^L|$ (4), and $L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^R - \bar{D}_{old(ij)}^R|$ (5), where N represents the number of pixels in the picture, $\bar{M}_{old(ij)}$ represents a pixel value of the occlusion map output by the trained second binocular matching network model, $\bar{D}_{old(ij)}^L$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the left picture in the sample data, $\bar{D}_{old(ij)}^R$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the right picture in the sample data, $\bar{D}_{ij}^L$ represents a pixel value of the disparity map output by the first binocular matching network model of the left picture in the real binocular data without a depth tag, $\bar{D}_{ij}^R$ represents a pixel value of the disparity map output by the first binocular matching network model of the right picture in the real binocular data without a depth tag, ij represents pixel coordinates of the pixel point, old represents the output of the trained second binocular matching network model, R represents the right picture or the related data thereof, L represents the left picture or the related data thereof, and γ_3 represents a strength coefficient. At step S1233: an output gradient of the first binocular matching network model is determined to be consistent with an output gradient of the second binocular matching network model by using formula (6) or formula (7): $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^L - \nabla \bar{D}_{old(ij)}^L|$ (6), and $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^R - \nabla \bar{D}_{old(ij)}^R|$ (7), where N represents the number of pixels in the picture, $\nabla \bar{D}_{ij}^L$ represents the gradient of the disparity map output by the first binocular matching network model of the left picture in the real binocular data without a depth tag, $\nabla \bar{D}_{ij}^R$ represents the gradient of the disparity map output by the first binocular matching network model of the right picture in the real binocular data without a depth tag, $\nabla \bar{D}_{old(ij)}^L$ represents the gradient of the disparity map output by the trained second binocular matching network model of the left picture in the sample data, $\nabla \bar{D}_{old(ij)}^R$ represents the gradient of the disparity map output by the trained second binocular matching network model of the right picture in the sample data, old represents the output of the trained second binocular matching network model, R represents the right picture or the related data thereof, and L represents the left picture or the related data thereof

At step S124, unsupervised training is performed on the trained second binocular

matching neural network model by using a loss function (Loss) according to the real binocular data without a depth tag, to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

Here, the loss function (Loss) uses the output of the trained second binocular matching neural network in step S122 to perform regularization on the fine-tuning training, avoiding the prediction blurring problem in the prior art existing in unsupervised fine-tuning and improving the effect of the first binocular matching network obtained by fine-tuning, thereby indirectly improving the effect of the monocular depth network supervised by the first binocular matching network. FIG. 1E is a schematic diagram of a loss function related picture according to embodiments of the present application. FIG. (a) in FIG. 1E is a left picture of the real binocular data without a depth tag. FIG. (b) in FIG. 1E is a right picture of the real binocular data without a depth tag. FIG. (c) in FIG. 1E is a disparity map output by the trained second binocular matching neural network in which the real binocular data without a depth tag consisting of FIGs. (a) and (b) is input. FIG. (d) in FIG. 1E is a picture obtained by reconstructing the left picture by combining the sampled right picture shown in FIG. (b) and the disparity map shown in FIG. (c). FIG. (e) in FIG. 1E is a picture obtained by subtracting the pixel in the left picture shown in FIG. (a) from the corresponding pixel in the reconstructed left picture shown in FIG. (d), i.e., a reconstruction error picture of the left picture. FIG. (f) in FIG. 1E is an occlusion map output by the trained second binocular matching neural network model in which the real binocular data without a depth tag consisting of FIGs. (a) and (b) is input. All red boxes 11 in FIG. (d) represent the parts of the reconstructed left picture that are different from the real left picture annotated by FIG. (a), and all red boxes 12 in FIG. (e) represent the parts of the reconstruction error map that has errors, i.e., the occluded parts. Here, when the unsupervised fine-tuning training of the binocular disparity network described in step S124 is implemented, it is necessary to reconstruct the left picture using the right picture, but the occlusion region cannot be reconstructed correctly, and therefore, the occlusion map is used to clean the wrong training signal of this section to improve the effect of unsupervised fine-tuning training.

At step S125, the monocular depth estimation network model is supervised by using the disparity map output by the first binocular matching neural network model, so as to train the monocular depth estimation network model.

Here, the sample picture of the monocular depth estimation network model is the left picture in the real binocular data without a depth tag, or the right picture in the real binocular data without a depth tag. If the left picture is used as the sample picture, the loss function is determined by formula (1), formula (2), formula (4), and formula (6). If the right picture is used

as the sample picture, the loss function is determined by formula (1), formula (3), formula (5), and formula (7).

In the embodiments of the present application, the wherein the monocular depth estimation network model is supervised according to a disparity map output by the first binocular matching neural network model, so as to train the monocular depth estimation network model refers to wherein the monocular depth estimation network model is supervised according to a depth map corresponding to the disparity map output by the first binocular matching neural network model, i.e., providing supervision information, so as to train the monocular depth estimation network model.

At step S126, an image to be processed is obtained.

At step S127, the image to be processed is input into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, where the monocular depth estimation network model is trained with a disparity map output by a first binocular matching neural network model as supervision.

At step S128, the analysis result of the image to be processed is output, where the analysis result of the image to be processed includes a disparity map output by the monocular depth estimation network model.

At step S129, a depth map corresponding to the disparity map is determined according to the disparity map output by the monocular depth estimation network model, the lens baseline distance of the camera that captures a picture input into the monocular depth estimation network model, and the lens focus of the camera that captures the picture input into the monocular depth estimation network model.

At step S130, a depth map corresponding to the disparity map is output.

In the embodiments of the present application, when the image to be processed is a street view image, the depth of the street view image is predicted by using the trained monocular depth estimation network model.

Based on the foregoing method embodiments, the embodiments of the present application further provide a method for monocular depth estimation. FIG. 2A is a schematic flowchart 2 of a method for monocular depth estimation according to embodiments of the present application. As shown in FIG. 2A, the method includes the following steps.

At step S201, a binocular matching network is trained by using synthetic data rendered by a rendering engine to obtain a disparity map of a binocular picture.

Here, the input of the binocular matching network is: a pair of binocular pictures (including the left picture and the right picture), and the outputs of the binocular matching network are: a disparity map and an occlusion map, that is, the binocular matching network uses the binocular pictures as the inputs to output the disparity map and the occlusion map. The disparity map is used to describe the disparity distance between each pixel point in the left picture and a corresponding pixel point in the right picture, in units of pixels. The occlusion map is used to describe whether a corresponding pixel point in the right picture of each pixel in the left picture is occluded by other objects. Due to the change in the angle of view, some regions in the left picture are occluded by other objects in the right picture, and the occlusion map is used to annotating whether the pixels in the left picture are occluded in the right picture. In this section, the binocular matching network is trained using the synthetic data generated by the computer rendering engine. First, some virtual 3D scenes are constructed by means of the rendering engine, and then the 3D scenes are mapped to binocular pictures by means of two virtual cameras, thereby obtaining the synthetic data. Moreover, data such as correct depth data and camera focus can also be obtained from the rendering engine, and thus supervised training can be performed on the binocular matching network directly by means of the tag data.

At step S202: fine-tuning of the binocular matching network obtained in step S201 is performed on the real binocular picture data using the loss function by means of an unsupervised fine-tuning method.

In this section, the binocular disparity network needs to adapt the real data. That is, unsupervised training is performed on the binocular disparity network using the real binocular data without a depth tag. Unsupervised training here refers to training using only binocular data in the case of no depth data tag. The embodiments of the present application propose a new unsupervised fine-tuning method, that is, unsupervised fine-tuning is performed by using the loss function in the foregoing embodiments. The main purpose of the loss function proposed by the embodiments of the present application is to perform fine-tuning of the binocular disparity network on the real binocular data without reducing the pre-training effect, and the preliminary output of the pre-training binocular disparity network obtained by step S201 is used in the fine-tuning process for guide and regularization. FIG. 2B is a schematic diagram showing the effect of the loss function in the embodiments of the present application. As shown in FIG. 2B, a picture 21 with the reference numeral of 21 is a disparity map obtained when the loss function in the prior art is used, and a picture 22 with the reference numeral of 22 is a disparity map obtained when the loss function proposed in the embodiments of the present application is used. The loss function in the prior art does not individually consider the occlusion region, and the

image reconstruction error of the occlusion region is also optimized to zero, which would cause a prediction disparity error of the occlusion region, and the edge of the disparity map is also blurred. However, the loss function in the present application cleans the wrong training signal in this section by using the occlusion map to improve the unsupervised fine-tuning training effect.

At step S203, supervision of the monocular depth estimation is performed on the real data by using the binocular matching network obtained in step S202, to finally obtain a monocular depth estimation network. Here, the input of the monocular depth estimation network is a single monocular picture, and the output of the monocular depth estimation network is a depth map. A binocular disparity network fine-tuned on the real data is obtained in step S202. For each pair of binocular pictures, the binocular disparity network predicts a disparity map, and a depth map corresponding to the disparity map is calculated according to the disparity map D , the binocular lens baseline distance b , and the lens focus f . That is, the depth map d corresponding to the disparity map is calculated according to formula (8): $d = bf / D$ (8). In order to train the monocular depth network to predict the depth map, the left picture is in the binocular picture pair is used as the input of the monocular deep network, and then supervision is performed by using the depth map output and calculated by the binocular disparity network, thereby training the monocular depth network to obtain a final result. In practical applications, an unmanned depth estimation module is trained by the method for monocular depth estimation in the embodiments of the present application, thereby performing three-dimensional reconstruction or obstacle detection on the scene. Moreover, the unsupervised fine-tuning method proposed in the embodiments of the present application improves the performance of the binocular disparity network.

In the prior art, it is limited and difficult for a supervised method for monocular depth estimation to obtain accurate tag data. The performance of reconstruction error-based unsupervised methods is generally limited by pixel matching ambiguity. In order to solve these problems, the embodiments of the present application provide a new method for monocular depth estimation, which solves the limitations in the prior art of supervised and unsupervised depth estimation methods. The method in the embodiments of the present application trains cross-modal synthetic data by using a binocular matching network and is used to supervise the monocular depth estimation network. The binocular matching network obtains the disparity based on the pixel matching relationship of the left and right pictures, instead of extracting from the semantic features. Therefore, the binocular matching network can be well generalized from the synthetic data to the real data. The method of the embodiments of the present application mainly includes three steps. First, the binocular matching network is trained with the synthetic

data, and the occlusion map and the disparity map are predicted from the binocular picture. Second, the trained binocular matching network is selectively adjusted based on the available real data with or without supervision. Third, the monocular depth estimation network is trained under the supervision of the binocular matching network obtained by the real data fine-tuning training obtained in the second step. This allows indirect use of binocular matching networks to make monocular depth estimation better utilize synthetic data to improve the performance.

At the first step, the synthetic data is used to train a binocular matching network, including: currently generating multiple composite images containing depth information by means of a graphics rendering engine. However, directly combining these composite image data with real data to train a monocular depth estimation network is generally inferior in performance because the monocular depth estimation is sensitive to the semantic information of the input scene. The large modal gap between the synthetic data and the real data makes the use of synthetic data-assisted training useless. However, binocular matching networks have better generalization capabilities, and binocular matching networks trained by the synthetic data can also obtain better disparity map output on real data. Therefore, the embodiments of the present application use binocular matching network training as a bridge between the synthetic data and the real data to improve the performance of monocular depth training. First, the binocular matching network is pre-trained using a large amount of synthetic binocular data. Unlike the conventional structure, the binocular matching network in the embodiments also estimates a multi-scale occlusion map based on the disparity map. The occlusion map indicates whether a corresponding pixel point in the right picture of the pixel in the left picture is occluded by other objects in a correct image. In next step, the unsupervised fine-tuning method uses the occlusion map to avoid wrong estimation. A horizontal disparity consistency test method is used to obtain an occlusion map M_{ij}^* with a correct tag from the correctly annotated disparity map by using formula (9): $M_{ij}^* = 1(|D_{ij}^{*L} - D_{ij}^{*wR}| \leq 1) = 1(|D_{ij}^{*L} - D_{i(j-D_{ij}^{*R})}^{*R}| \leq 1)$ (9), where the subscript i represents the value of the i -th row in the image, and the subscript j represents the value of the j -th column in the image. $D^{*L/R}$ represents a disparity map of the left and right pictures, and D^{*wR} represents a disparity map of the left picture reconstructed by the right picture. For the non-occluded region, the left disparity map is consistent with the disparity map of the left picture reconstructed by the right picture. The threshold for the consistency check is set to 1. The occlusion map is 0 in the occlusion region and 1 in the non-occlusion region. Therefore, the embodiments use the following formula (10) to calculate the loss (Loss) of training the binocular matching network using the synthetic data. At this stage, the loss function L_{stereo} consists of two parts, that is, a disparity map estimation error L_{disp} and an occlusion map estimation error L_{occ} .

The multi-scale intermediate layer of the binocular disparity network also produces disparity and occlusion prediction, and is directly applied to the loss weight w_m of multi-scale prediction, L_{disp}^m represents a corresponding disparity map estimation error of each layer, L_{occ}^m represents a corresponding occlusion map estimation error of each layer, and m represents the m -th layer: $L_{stereo} = L_{disp} + L_{occ} = \sum w_m (L_{disp}^m + L_{occ}^m)$ (10). In order to train the disparity map, the L1 loss function is used to avoid the influence of abnormal values, making the training process more robust. In order to train the occlusion map, formula (11) is used to represent the occlusion map estimation error L_{occ} , and the binary cross entropy loss is used as a classification task to train the occlusion map: $L_{occ} = -\frac{1}{N} \sum_{ij} M_{ij}^* \log(\bar{M}_{ij}) + (1 - M_{ij}^*) \log(1 - \bar{M}_{ij})$ (11), where N is the total number of pixels in the image, M_{ij}^* represents the correctly annotated occlusion map, and $\bar{M}_{ij}$ represents the occlusion map output by the trained binocular matching network model.

At the second step, the training the trained binocular matching network obtained in the first step on the real data by using the supervised or unsupervised fine-tuning method includes: performing fine-tuning on the trained binocular matching network in the embodiments of the present application in two ways. The supervised fine-tuning method improves the error of the previous pixel matching prediction only using the multi-scale L1 regression loss function $L_{stereo-sup\ ft}$, i.e., the disparity map estimation error L_{disp} , see formula (12): $L_{stereo-sup\ ft} = L_{disp} = \sum_m L_{disp}^m$ (12). The results show that a small amount of supervised data is used, e.g., 100 pictures, the binocular matching network can also be adapted from synthetic modal data to real modal data. The unsupervised fine-tuning method. For unsupervised network tuning, the unsupervised fine-tuning method in the prior art has a blurred disparity map and poor performance, as shown in picture 21 in FIG. 2B. This is due to the limitations of unsupervised losses and the ambiguity of matching pixels with only RGB values. Therefore, the embodiments of the present application introduce additional regular term constraints to improve performance. The corresponding occlusion map and disparity map are obtained from the trained binocular matching network without fine tuning by using the real data, and are annotated as $\bar{M}_{old}$ and $\bar{D}_{old}$, respectively. The two pieces of data are used to help standardize the training process. Further, the acquisition of the unsupervised fine-tuning loss function proposed in the embodiments of the present application, i.e., the loss function $L_{stereo-un\ sup\ ft}$, can be referred to the description in the previous embodiments.

At the third step, the training the monocular depth estimation network includes: we have conducted cross-modal training on the binocular matching network by means of a large amount

of synthetic data so far, and fine-tuning is performed using the real data. In order to train the final monocular depth estimation network, the embodiments of the present application provide training data by using a disparity map predicted by the trained binocular matching network. The loss L_{mono} of the monocular depth estimation is given by the following part, see formula (13):

$$L_{mono} = \frac{1}{N} \sum_{i,j} |\tilde{D}_{ij}^L - \bar{D}_{ij}^L| \quad (13),$$
 here, N is the sum of pixel points, $\tilde{D}_{ij}^L$ refers to the disparity map output by the monocular depth estimation network, $\bar{D}_{ij}^L$ refers to the disparity map output by the trained binocular matching network, or the disparity map output by a network subjected to fine-tuning of the binocular matching network. It should be noted that formulas (9) to (13) are all described by taking the left picture in the real data of the monocular depth estimation network as a training sample as an example. Experiment: since the monocular depth estimation network is sensitive to the change in the angle of view, no cropping or scaling is performed on the training data. The input of the monocular depth estimation network and the disparity map for supervising the monocular depth estimation network are both from the trained binocular matching network. FIG. 2C is a schematic diagram of a visual depth estimation result according to embodiments of the present application. As shown in FIG. 2C, FIG. 2C shows depth maps corresponding to three different street view pictures obtained by using the prior art and the method for monocular depth estimation in the embodiments of the present application, where the first line is the input of the monocular depth estimation network, i. e. , three different street view pictures; the second line is the depth data obtained by interpolating the sparse laser radar depth map by using the nearest neighbor algorithm, and the third to fifth lines are depth maps corresponding to the three input pictures obtained by the three different method for monocular depth estimations in the prior art. The results of the present application are shown in the last three lines. The binocular matching network trained by using the synthetic data obtained in the first step of the embodiments of the present application is used to supervise the monocular depth estimation network, and the depth maps corresponding to the three input pictures of the obtained monocular depth network are a picture 21 with the reference numeral of 21, a picture 22 with the reference numeral of 22, and a picture 23 with the reference numeral of 23. Fine-tuning is performed on the trained binocular matching network by using the unsupervised loss function proposed in the embodiments of the present application, and the disparity map output by the fine-tuned network is used as the training data of the monocular depth estimation network, and the depth maps corresponding to the three input pictures of the obtained monocular depth network are a picture 24 with the reference numeral of 24, a picture 25 with the reference numeral of 25, and a picture 26 with the reference numeral of 26. Supervised fine-tuning is performed on the trained binocular matching network, and the disparity map output by the fine-tuned network is used as the training data of

the monocular depth estimation network, and the depth maps corresponding to the three input pictures of the obtained monocular depth network are a picture 27 with the reference numeral of 27, a picture 28 with the reference numeral of 28, and a picture 29 with the reference numeral of 29. It can be seen from the picture 21 with the reference numeral of 21 to the picture 29 with the reference numeral of 29 that the models obtained by the method for monocular depth estimation in the embodiments of the present application can capture more detailed scene structures.

The embodiments of the present application provide a monocular depth estimation apparatus. FIG. 3 is a schematic structural diagram of the monocular depth estimation apparatus according to embodiments of the present application. As shown in FIG. 3, the apparatus 300 includes an obtaining module 301, an execution module 302, and an output module 303.

The obtaining module 301 is configured to obtain an image to be processed.

The execution module 302 is configured to input the image to be processed into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, where the monocular depth estimation network model is trained with a disparity map output by a first binocular matching neural network model as supervision.

The output module 303 is configured to output the analysis result of the image to be processed.

In some embodiments, the apparatus further includes a third training module, configured to perform supervision on the monocular depth estimation network model by using the disparity map output by the first binocular matching neural network model, so as to train the monocular depth estimation network model.

In some embodiments, the apparatus further includes: a first training module, configured to train a second binocular matching neural network model according to the obtained synthetic sample data; and a second training module, configured to adjust a parameter of the trained second binocular matching neural network model according to the obtained real sample data, to obtain the first binocular matching neural network model.

In some embodiments, the apparatus further includes: a first obtaining module, configured to obtain a synthesized binocular picture with a depth tag as the synthetic sample data, where the synthesized binocular picture includes a synthesized left picture and a synthesized right picture.

In some embodiments, the first training module includes: a first training unit, configured to train the second binocular matching neural network model according to the synthesized

binocular picture to obtain the trained second binocular matching neural network model, where outputs of the trained second binocular matching neural network model are a disparity map and an occlusion map, the disparity map describing a disparity distance between each pixel point in the left picture and a corresponding pixel point in the right picture, the disparity distance being in units of pixels, and the occlusion map describing whether a corresponding pixel point in the right picture of each pixel point in the left picture is occluded by an object.

In some embodiments, the apparatus further includes: a constructing module, configured to construct a virtual 3D scene by means of a rendering engine; a mapping module, configured to map the 3D scene into a binocular picture by means of two virtual cameras; a second obtaining module, configured to obtain the depth data of the synthesized binocular picture according to the position where the virtual 3D scene is constructed, the direction in which the virtual 3D scene is constructed, and the lens focus of the virtual camera; and a third obtaining module, configured to annotate the binocular picture according to the depth data to obtain the synthesized binocular picture.

In some embodiments, the second training module includes a second training unit, configured to perform supervised training on the trained second binocular matching neural network model according to the obtained real binocular data with a depth tag, to adjust a weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

In some embodiments, the second training unit in the second training module is further configured to perform unsupervised training on the trained second binocular matching neural network model according to the obtained real binocular data without a depth tag, to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

In some embodiments, the second training unit in the second training module includes a second training component, configured to perform unsupervised training on the trained second binocular matching neural network model by using a loss function according to the real binocular data without a depth tag to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

In some embodiments, the apparatus further includes: a first determining module, configured to determine the loss function using formula (14): $L_{stereo-un\sup ft} = L_{photo} + \gamma_1 L_{abs} + \gamma_2 L_{rel}$ (14), where $L_{stereo-un\sup ft}$ represents the loss function, L_{photo} represents a reconstruction error, L_{abs} represents that the disparity map output by the first binocular matching network model is less

deviated from the disparity map output by the trained second binocular matching network model, L_{rel} represents that an output gradient of the first binocular matching network model is constrained to be consistent with an output gradient of the trained second binocular matching network model, and γ_1 and γ_2 represent strength coefficients.

In some embodiments, the apparatus further includes: a second determining module, configured to determine the reconstruction error using formula (15) or formula (16):

$$L_{photo} = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^L - I_{ij}^{wR}| = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^L - I_{i(j-\bar{D}_{ij}^L)}^R| \quad (15) \quad \text{or} \quad L_{photo} = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^R - I_{ij}^{wL}| = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^R - I_{i(j+\bar{D}_{ij}^R)}^L|$$

(16), where N represents the number of pixels in the picture, $\bar{M}_{old(ij)}$ represents a pixel value of the occlusion map output by the trained second binocular matching network model, I_{ij}^L represents a pixel value of the left picture in the real binocular data without a depth tag, I_{ij}^R represents a pixel value of the right picture in the real binocular data without a depth tag, I_{ij}^{wR} represents a pixel value of a picture synthesized after sampling the right picture, I_{ij}^{wL} represents a pixel value of a picture synthesized after sampling the left picture, $\bar{D}_{ij}^L$ represents a pixel value of the disparity map output by the first binocular matching network model of the left picture in the real binocular data without a depth tag, $\bar{D}_{ij}^R$ represents a pixel value of the disparity map output by the first binocular matching network model of the right picture in the real binocular data without a depth tag, and ij represents pixel coordinates of the pixel point.

In some embodiments, the apparatus further includes: a third determining module, configured to determine, by using formula (17) or formula (18), that the disparity map output by the first binocular matching network model is less deviated from the disparity map output by the trained second binocular matching network model, $L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^L - \bar{D}_{old(ij)}^L|$ (17) or

$$L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^R - \bar{D}_{old(ij)}^R| \quad (18),$$

where $\bar{D}_{old(ij)}^L$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the left picture in the sample data, $\bar{D}_{old(ij)}^R$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the right picture in the sample data, and γ_3 represents a strength coefficient.

In some embodiments, the apparatus further includes: a fourth determining module, configured to determine, by using formula (19) or formula (20), that the output gradient of the first binocular matching network model is consistent with the output gradient of the second binocular matching network model, $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^L - \nabla \bar{D}_{old(ij)}^L|$ (19) or $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^R - \nabla \bar{D}_{old(ij)}^R|$ (20),

where $\nabla \bar{D}_{ij}^L$ represents the gradient of the disparity map output by the first binocular matching

network model of the left picture in the real binocular data without a depth tag, $\nabla \bar{D}_{ij}^R$ represents the gradient of the disparity map output by the first binocular matching network model of the right picture in the real binocular data without a depth tag, $\nabla \bar{D}_{old(ij)}^L$ represents the gradient of the disparity map output by the trained second binocular matching network model of the left picture in the sample data, and $\nabla \bar{D}_{old(ij)}^R$ represents the gradient of the disparity map output by the trained second binocular matching network model of the right picture in the sample data.

In some embodiments, the real binocular data with a depth tag includes a left picture and a right picture, and correspondingly, the third training module includes: a first obtaining unit, configured to obtain the left or right picture in the real binocular data with a depth tag as a training sample; and a first training unit, configured to train the monocular depth estimation network model according to the left or right picture in the real binocular data with a depth tag.

In some embodiments, the real binocular data without a depth tag includes a left picture and a right picture, and correspondingly, the third training module further includes: a second obtaining unit, configured to input the real binocular data without a depth tag into the first binocular matching neural network model to obtain a corresponding disparity map; a first determining unit, configured to determine a depth map corresponding to the disparity map according to the corresponding disparity map, a lens baseline distance of a camera that captures the real binocular data without a depth tag, and a lens focus of the camera that captures the real binocular data without a depth tag; and a second training unit, configured to use the left or right picture in the real binocular data without a depth tag as sample data, and perform supervision on the monocular depth estimation network model according to the depth map corresponding to the disparity map, so as to train the monocular depth estimation network model.

In some embodiments, the analysis result of the image to be processed includes a disparity map output by the monocular depth estimation network model, and correspondingly, the apparatus further includes: a fifth determining module, configured to determine a depth map corresponding to the disparity map according to the disparity map output by the monocular depth estimation network model, the lens baseline distance of the camera that captures a picture input into the monocular depth estimation network model, and the lens focus of the camera that captures the picture input into the monocular depth estimation network model; and a first output module, configured to output the depth map corresponding to the disparity map.

It should be noted here that the description of the foregoing apparatus embodiments is similar to the description of the foregoing method embodiments, and has similar advantages as the method embodiments. For the technical details that are not disclosed in the apparatus

embodiments of the present application, please refer to the description of the method embodiments of the present application. In the embodiments of the present application, when implemented in the form of a software functional module and sold or used as an independent product, the method for monocular depth estimation may also be stored in a computer readable storage medium. Based on such understanding, the technical solutions of the embodiments of the present application essentially, or the part contributing to the prior art may be implemented in the form of a software product. The computer software product is stored in a storage medium and includes several instructions for instructing a computer device to perform all or some of the methods in the embodiments of the present application. The foregoing storage medium includes any medium that can store program codes, such as a USB flash drive, a removable hard disk, a Read-Only Memory (ROM), a magnetic disk, or an optical disk. In this case, the embodiments of the present application are not limited to any particular combination of hardware and software. Accordingly, the embodiments of the present application provide a monocular depth estimation device, including a memory and a processor, where the memory stores a computer program running in the processor, and when the processor executes the program, steps of the method for monocular depth estimation are implemented. Accordingly, the embodiments of the present application provide a computer-readable storage medium, having a computer program stored thereon, where when the computer program is executed by a processor, steps of the method for monocular depth estimation are implemented. It should be noted here that the description of the foregoing storage medium and device embodiments is similar to the description of the foregoing method embodiments, and has similar advantages as the method embodiments. For the technical details that are not disclosed in the storage medium and apparatus embodiments of the present application, please refer to the description of the method embodiments of the present application.

It should be noted that FIG. 4 is a schematic diagram of a hardware entity of a monocular depth estimation device according to embodiments of the present application. As shown in FIG. 4, the hardware entity of the monocular depth estimation device 400 includes a memory 401, a communication bus 402, and a processor 403, where the memory 401 is configured to store instructions and application that can be executed by the processor 403, and cache data processed or to be processed by each module in the processor 403 and the monocular depth estimation device 400, which can be implemented by means of a FLASH or Random Access Memory (RAM). The communication bus 402 enables the monocular depth estimation device 400 to be in communication with other terminals or servers by means of a network, and also implements connection communication between the processor 403 and the memory 401. The processor 403 generally controls the overall operation of the monocular depth estimation device 400.

It should be noted that the term "comprising", "including" or any other variant thereof herein is intended to encompass a non-exclusive inclusion, such that a process, method, article, or apparatus that includes a series of elements includes those elements. Moreover, other elements not explicitly listed are also included, or elements that are inherent to the process, method, article, or apparatus are also included. An element defined by the phrase "including one ..." does not exclude the presence of additional same elements in the process, method, article, or apparatus that includes the element, without more limitations.

According to the foregoing description of implementations, persons skilled in the art could clearly know that the methods of the foregoing embodiments are implemented by means of software and necessary universal hardware platform, and certainly can also be implemented by means of hardware, but in many cases, the former is a better implementation. Based on such understanding, the technical solutions of the present application essentially, or the part contributing to the prior art may be implemented in the form of a software product. The computer software product is stored in a storage medium (such as an ROM/RAM, a magnetic disk, and an optical disk) and includes several instructions for instructing a computer device (which may be a mobile phone, a computer, a server, an air conditioner, or a network device, etc.) to perform the methods in the embodiments of the present application.

The present application is described with reference to flowcharts and/or block diagrams of methods, devices (apparatuses), and computer program products according to the embodiments of the present application. It should be understood that each flow and/or block in the flowchart and/or block diagram as well as a combination of flows and/or blocks in the flowchart and/or block diagram are implemented by computer program instructions. These computer program instructions can be provided to a processor of a general-purpose computer, a special-purpose computer, an embedded processor, or other programmable data processing devices to produce a machine, so that instructions executed by the processor of a computer or other programmable data processing devices produce an apparatus for implementing the functions specified in one or more flows of the flowchart and/or in one or more blocks of the block diagram. These computer program instructions can also be stored in a computer readable memory that can direct a computer or other programmable data processing device to operate in a particular manner, so that the instructions stored in the computer readable memory produce a product including an instruction apparatus. The instruction apparatus implements the functions specified in one or more flows of the flowchart and/or one or more blocks of the block diagram.

These computer program instructions can also be loaded onto a computer or other programmable data processing device so that a series of operational steps are performed on a

computer or other programmable device to produce computer-implemented processing, and thus the instructions executed on a computer or other programmable device provide steps for implementing the functions specified in one or more flows of the flowchart and/or one or more blocks of the block diagram.

The above are only preferred embodiments of the present application, and are not intended to limit the patent scope of the present application, and any equivalent structure or equivalent process transformations made by the specification and the drawings of the present application, or directly or indirectly applied to other related technical fields, are included in the scope of protection of the present application.

CLAIMS

1. A method for estimating a monocular depth, comprising:

obtaining an image to be processed;

inputting the image to be processed into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, wherein the monocular depth estimation network model is trained by supervision of a disparity map output by a first binocular matching neural network model; and

outputting the analysis result of the image to be processed.

2. The method according to claim 1, wherein a training process of the first binocular matching neural network model comprises:

training a second binocular matching neural network model according to obtained synthetic sample data; and

adjusting a parameter of the trained second binocular matching neural network model according to obtained real sample data, to obtain the first binocular matching neural network model.

3. The method according to claim 2, further comprising:

obtaining a synthesized binocular picture with a depth tag as the obtained synthetic sample data, wherein the synthesized binocular picture comprises a synthesized left picture and a synthesized right picture.

4. The method according to claim 3, wherein the training a second binocular matching neural network model according to the obtained synthetic sample data comprises:

training the second binocular matching neural network model according to the synthesized binocular picture to obtain the trained second binocular matching neural network model, wherein outputs of the trained second binocular matching neural network model are a disparity map and an occlusion map, the disparity map describing a disparity distance between each pixel point in the synthesized left picture and a corresponding pixel point in the synthesized right picture, the disparity distance being in units of pixels, and the occlusion map describing whether a respective pixel point in the synthesized right picture corresponding to each pixel point in the synthesized left picture is occluded by an object.

5. The method according to claim 2, wherein the adjusting a parameter of the trained second binocular matching neural network model according to the obtained real sample data, to obtain the first binocular matching neural network model comprises:

performing supervised training on the trained second binocular matching neural network model according to the obtained real binocular data with a depth tag, to adjust a weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

6. The method according to claim 2, wherein the adjusting a parameter of the trained second binocular matching neural network model according to the obtained real sample data, to obtain the first binocular matching neural network model, further comprises:

performing unsupervised training on the trained second binocular matching neural network model according to the obtained real binocular data without a depth tag, to adjust a weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

7. The method according to claim 6, wherein the performing unsupervised training on the trained second binocular matching neural network model according to the obtained real binocular data without a depth tag, to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model comprises:

performing unsupervised training on the trained second binocular matching neural network model according to the real binocular data without a depth tag by using a loss function, so as to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

8. The method according to claim 7, further comprising:

determining the loss function by using a formula $L_{stereo-un\ sup\ ft} = L_{photo} + \gamma_1 L_{abs} + \gamma_2 L_{rel}$, wherein $L_{stereo-un\ sup\ ft}$ represents the loss function, L_{photo} represents a reconstruction error, L_{abs} represents that the disparity map output by the first binocular matching network model is less deviated from the disparity map output by the trained second binocular matching network model, L_{rel} represents that an output gradient of the first binocular matching network model is constrained to be consistent with an output gradient of the trained second binocular matching network model, and γ_1 and γ_2 represent strength coefficients.

9. The method according to claim 8, further comprising:

determining the reconstruction error using a formula

$$L_{photo} = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^L - I_{ij}^{wR}| = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^L - I_{i(j-D_{ij}^L)}^R| \quad \text{or} \quad L_{photo} = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^R - I_{ij}^{wL}| = \frac{1}{N} \sum_{i,j} \bar{M}_{old(ij)} |I_{ij}^R - I_{i(j+D_{ij}^R)}^L|,$$

wherein N represents a number of pixels in a picture, $\bar{M}_{old(ij)}$ represents a pixel value of the occlusion map output by the trained second binocular matching network model, I_{ij}^L represents a pixel value of a left picture in the real binocular data without a depth tag, I_{ij}^R represents a pixel value of a right picture in the real binocular data without a depth tag, I_{ij}^{wR} represents a pixel value of a picture synthesized after sampling the right picture, I_{ij}^{wL} represents a pixel value of a picture synthesized after sampling the left picture, $\bar{D}_{ij}^L$ represents a pixel value of the disparity map output by the first binocular matching network model of the left picture in the real binocular data without a depth tag, $\bar{D}_{ij}^R$ represents a pixel value of the disparity map output by the first binocular matching network model of the right picture in the real binocular data without a depth tag, and ij represents a pixel coordinate of a pixel point.

10. The method according to claim 8, further comprising:

determining, by using a formula $L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^L - \bar{D}_{old(ij)}^L|$ or $L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^R - \bar{D}_{old(ij)}^R|$, that the disparity map output by the first binocular matching network model is less deviated from the disparity map output by the trained second binocular matching network model,

wherein N represents a number of pixels in a picture, $\bar{M}_{old(ij)}$ represents a pixel value of the occlusion map output by the trained second binocular matching network model, $\bar{D}_{ij}^L$ represents a pixel value of the disparity map output by the first binocular matching network model of a left picture in the real binocular data without a depth tag, $\bar{D}_{ij}^R$ represents a pixel value of the disparity map output by the first binocular matching network model of a right picture in the real binocular data without a depth tag, $\bar{D}_{old(ij)}^L$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the left picture, $\bar{D}_{old(ij)}^R$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the right picture, ij represents a pixel coordinate of a pixel point, and γ_3 represents a strength coefficient.

11. The method according to claim 8, further comprising:

determining, by using a formula $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^L - \nabla \bar{D}_{old(ij)}^L|$ or $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^R - \nabla \bar{D}_{old(ij)}^R|$,

that the output gradient of the first binocular matching network model is consistent with the output gradient of the second binocular matching network model,

wherein N represents a number of pixels in a picture, $\nabla \bar{D}_{ij}^L$ represents a gradient of the disparity map output by the first binocular matching network model of a left picture in the real binocular data without a depth tag, $\nabla \bar{D}_{ij}^R$ represents a gradient of the disparity map output by the first binocular matching network model of a right picture in the real binocular data without a depth tag, $\nabla \bar{D}_{old(ij)}^L$ represents a gradient of the disparity map output by the trained second binocular matching network model of the left picture, $\nabla \bar{D}_{old(ij)}^R$ represents a gradient of the disparity map output by the trained second binocular matching network model of the right picture, and ij represents a pixel coordinate of a pixel point.

12. The method according to claim 5, wherein the real binocular data with a depth tag comprises a left picture and a right picture, and correspondingly, the training process of the monocular depth estimation network model comprises:

obtaining the left picture or right picture in the real binocular data with a depth tag as a training sample; and

training the monocular depth estimation network model according to the left picture or right picture in the real binocular data with a depth tag.

13. The method according to any one of claims 6 to 11, wherein the real binocular data without a depth tag comprises a left picture and a right picture, and correspondingly, the training process of the monocular depth estimation network model comprises:

inputting the real binocular data without a depth tag into the first binocular matching neural network model to obtain a corresponding disparity map;

determining a depth map corresponding to the disparity map according to the corresponding disparity map, a lens baseline distance of a camera that captures the real binocular data without a depth tag, and a lens focus of the camera that captures the real binocular data without a depth tag; and

performing supervision on the monocular depth estimation network model according to the depth map corresponding to the disparity map by using the left picture or right picture in the real binocular data without a depth tag as sample data, to train the monocular depth estimation network model.

14. The method according to claim 12 or 13, wherein the analysis result of the image to be processed comprises a disparity map output by the monocular depth estimation network model,

and correspondingly, the method further comprises:

determining a depth map corresponding to the disparity map according to the disparity map output by the monocular depth estimation network model, the lens baseline distance of the camera that captures a picture input into the monocular depth estimation network model, and the lens focus of the camera that captures the picture input into the monocular depth estimation network model; and

outputting the depth map corresponding to the disparity map.

15. A monocular depth estimation apparatus, comprising: an obtaining module, an execution module, and an output module, wherein

the obtaining module is configured to obtain an image to be processed;

the execution module is configured to input the image to be processed into a trained monocular depth estimation network model, to obtain an analysis result of the image to be processed, wherein the monocular depth estimation network model is trained by supervision of a disparity map output by a first binocular matching neural network model; and

the output module is configured to output the analysis result of the image to be processed.

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

a first training module, configured to train a second binocular matching neural network model according to the obtained synthetic sample data; and

a second training module, configured to adjust a parameter of the trained second binocular matching neural network model according to the obtained real sample data, to obtain the first binocular matching neural network model.

17. The apparatus according to claim 16, further comprising: a first obtaining module, configured to obtain a synthesized binocular picture with a depth tag as the obtained synthetic sample data, wherein the synthesized binocular picture comprises a synthesized left picture and a synthesized right picture.

18. The apparatus according to claim 17, wherein the first training module comprises:

a first training unit, configured to train the second binocular matching neural network model according to the synthesized binocular picture to obtain the trained second binocular matching neural network model, wherein outputs of the trained second binocular matching neural network model are a disparity map and an occlusion map, the disparity map describing a disparity distance between each pixel point in the synthesized left picture and a corresponding

pixel point in the synthesized right picture, the disparity distance being in units of pixels, and the occlusion map describing whether a respective pixel point in the right picture corresponding to each pixel point in the synthesized left picture is occluded by an object.

19. The apparatus according to claim 16, wherein the second training module comprises a second training unit, configured to perform supervised training on the trained second binocular matching neural network model according to the obtained real binocular data with a depth tag, to adjust a weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

20. The apparatus according to claim 16, wherein the second training unit is further configured to perform unsupervised training on the trained second binocular matching neural network model according to the obtained real binocular data without a depth tag, to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

21. The apparatus according to claim 20, wherein the second training unit comprises a second training component, configured to perform unsupervised training on the trained second binocular matching neural network model according to the real binocular data without a depth tag by using a loss function, to adjust the weight of the trained second binocular matching neural network model to obtain the first binocular matching neural network model.

22. The apparatus according to claim 21, further comprising: a first determining module, configured to determine the loss function by using a formula $L_{\text{stereo-unsup ft}} = L_{\text{photo}} + \gamma_1 L_{\text{abs}} + \gamma_2 L_{\text{rel}}$,

wherein $L_{\text{stereo-unsup ft}}$ represents the loss function, L_{photo} represents a reconstruction error, L_{abs} represents that the disparity map output by the first binocular matching network model is less deviated from the disparity map output by the trained second binocular matching network model, L_{rel} represents that an output gradient of the first binocular matching network model is constrained to be consistent with an output gradient of the trained second binocular matching network model, and γ_1 and γ_2 represent strength coefficients.

23. The apparatus according to claim 22, further comprising: a second determining module, configured to determine the reconstruction error using a formula

$$L_{\text{photo}} = \frac{1}{N} \sum_{i,j} \bar{M}_{\text{old}(ij)} |I_{ij}^L - I_{ij}^{wR}| = \frac{1}{N} \sum_{i,j} \bar{M}_{\text{old}(ij)} |I_{ij}^L - I_{i(j-\bar{D}_j^L)}^R| \quad \text{or} \quad L_{\text{photo}} = \frac{1}{N} \sum_{i,j} \bar{M}_{\text{old}(ij)} |I_{ij}^R - I_{ij}^{wL}| = \frac{1}{N} \sum_{i,j} \bar{M}_{\text{old}(ij)} |I_{ij}^R - I_{i(j+\bar{D}_j^R)}^L|,$$

wherein N represents a number of pixels in a picture, $\bar{M}_{\text{old}(ij)}$ represents a pixel value of the occlusion map output by the trained second binocular matching network model, I_{ij}^L

represents a pixel value of a left picture in the real binocular data without a depth tag, I_{ij}^R represents a pixel value of a right picture in the real binocular data without a depth tag, I_{ij}^{wR} represents a pixel value of a picture synthesized after sampling the right picture, I_{ij}^{wL} represents a pixel value of a picture synthesized after sampling the left picture, $\bar{D}_{ij}^L$ represents a pixel value of the disparity map output by the first binocular matching network model of the left picture in the real binocular data without a depth tag, $\bar{D}_{ij}^R$ represents a pixel value of the disparity map output by the first binocular matching network model of the right picture in the real binocular data without a depth tag, and ij represents a pixel coordinate of a pixel point.

24. The apparatus according to claim 22, further comprising: a third determining module, configured to determine, by using a formula $L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^L - \bar{D}_{old(ij)}^L|$ or $L_{abs} = \frac{1}{N} \sum_{i,j} (1 - \bar{M}_{old(ij)} + \gamma_3) |\bar{D}_{ij}^R - \bar{D}_{old(ij)}^R|$, that the disparity map output by the first binocular matching network model is less deviated from the disparity map output by the trained second binocular matching network model,

wherein N represents a number of pixels in a picture, $\bar{M}_{old(ij)}$ represents a pixel value of the occlusion map output by the trained second binocular matching network model, $\bar{D}_{ij}^L$ represents a pixel value of the disparity map output by the first binocular matching network model of a left picture in the real binocular data without a depth tag, $\bar{D}_{ij}^R$ represents a pixel value of the disparity map output by the first binocular matching network model of a right picture in the real binocular data without a depth tag, $\bar{D}_{old(ij)}^L$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the left picture, $\bar{D}_{old(ij)}^R$ represents a pixel value of the disparity map output by the trained second binocular matching network model of the right picture, ij represents a pixel coordinate of a pixel point, and γ_3 represents a strength coefficient.

25. The apparatus according to claim 22, further comprising: a fourth determining module, configured to determine, by using a formula $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^L - \nabla \bar{D}_{old(ij)}^L|$ or $L_{rel} = \frac{1}{N} \sum_{i,j} |\nabla \bar{D}_{ij}^R - \nabla \bar{D}_{old(ij)}^R|$, that the output gradient of the first binocular matching network model is consistent with the output gradient of the second binocular matching network model,

wherein N represents a number of pixels in a picture, $\nabla \bar{D}_{ij}^L$ represents a gradient of the disparity map output by the first binocular matching network model of a left picture in the real binocular data without a depth tag, $\nabla \bar{D}_{ij}^R$ represents a gradient of the disparity map output by the first binocular matching network model of a right picture in the real binocular data without a

depth tag, $\nabla \bar{D}_{old(ij)}^L$ represents the gradient of the disparity map output by the trained second binocular matching network model of the left picture, $\nabla \bar{D}_{old(ij)}^R$ represents the gradient of the disparity map output by the trained second binocular matching network model of the right picture, and ij represents a pixel coordinate of a pixel point.

26. The apparatus according to claim 19, wherein the real binocular data with a depth tag comprises a left picture or a right picture, and correspondingly, the apparatus further comprises a third training module, configured to obtain the left or right picture in the real binocular data with a depth tag as a training sample; and train the monocular depth estimation network model according to the left picture or right picture in the real binocular data with a depth tag.

27. The apparatus according to any one of claims 20 to 25, wherein the real binocular data without a depth tag comprises a left picture and a right picture, and correspondingly, the apparatus further comprises:

a third training module, configured to input the real binocular data without a depth tag into the first binocular matching neural network model to obtain a corresponding disparity map; determine a depth map corresponding to the disparity map according to the corresponding disparity map, a lens baseline distance of a camera that captures the real binocular data without a depth tag, and a lens focus of the camera that captures the real binocular data without a depth tag; and perform supervision on the monocular depth estimation network model according to the depth map corresponding to the disparity map by using the left picture or right picture in the real binocular data without a depth tag as sample data, to train the monocular depth estimation network model.

28. The apparatus according to claim 26 or 27, wherein the analysis result of the image to be processed comprises a disparity map output by the monocular depth estimation network model, and correspondingly, the apparatus further comprises:

a fifth determining module, configured to determine a depth map corresponding to the disparity map according to the disparity map output by the monocular depth estimation network model, the lens baseline distance of the camera that captures a picture input into the monocular depth estimation network model, and the lens focus of the camera that captures the picture input into the monocular depth estimation network model; and

a first output module, configured to output the depth map corresponding to the disparity map.

29. A device for estimating a monocular depth, comprising a memory and a processor, wherein the memory stores computer programs capable of running in the processor, and when the

programs being executed, the processor implements steps of the method for estimating a monocular depth according to any one of claims 1 to 14.

30. A computer readable storage medium, having stored thereon computer programs that, when being executed by a processor, implement steps of the method for estimating a monocular depth according to any one of claims 1 to 14.