

IMAGE GENERATION METHOD AND APPARATUS, STORAGE MEDIUM AND ELECTRONIC DEVICE

[0001] This application claims the priority to Chinese Patent Application No. 202311804219.2, titled "IMAGE GENERATION METHOD AND APPARATUS, STORAGE MEDIUM AND ELECTRONIC DEVICE", filed on December 26, 2023 with the China National Intellectual Property Administration, which is incorporated herein by reference in its entirety.

FIELD

[0002] The present disclosure relates to the field of image processing, and in particular to a method and an apparatus for image generation, a storage medium, and an electronic device.

BACKGROUND

[0003] With increasingly high requirements for product quality in industrial production, various product quality inspection technologies have been widely introduced into related industries.

[0004] In product quality inspection technology based on image detection, a neural network model is used to inspect an appearance, a structure, and the like of a product, so as to eliminate defective products. When the product quality is inspected based on the appearance by using the neural network model, image data is required for training. However, image data in the industrial field is usually relatively scarce, and thus an effective image generation technology applied to industrial production is indispensable.

SUMMARY

[0005] The present disclosure is provided in view of the above issues.

[0006] According to a first aspect of the present disclosure, a method for image generation is provided. The method includes:

determining a first replacement region and a second replacement region in a target

image;

performing feature extraction on the target image to obtain one or more target feature maps of the target image;

5 determining, in each of the one or more target feature maps, a first feature region corresponding to a position of the first replacement region and a second feature region corresponding to a position of the second replacement region;

performing, in the target feature map, transformation on feature information of the first feature region based on feature information of the second feature region, to obtain a target composite map after the feature information of the first feature region is transformed; and

10 inputting one or more target composite maps into a trained image generation model to obtain a target composite image, where the trained image generation model is configured to perform reverse inference based on the one or more target composite maps to obtain the target composite image.

[0007] In a possible implementation, the performing, in the target feature map, transformation on the feature information of the first feature region based on the feature information of the second feature region includes:

for each of first feature values in the first feature region, determining a second feature value in the second feature region; and

replacing the first feature values in the first feature region with the second feature values.

20 **[0008]** In a possible implementation, the determining the second feature value in the second feature region includes: determining the second feature value in the second feature region in a random manner.

[0009] In a possible implementation, the performing transformation on the feature information of the first feature region based on the feature information of the second feature region, to obtain the target composite map after the feature information of the first feature region is transformed, includes: directly replacing the feature information of the first feature region with the feature information of the second feature region.

[0010] In a possible implementation, the performing feature extraction on the target image to obtain the one or more target feature maps of the target image includes:

inputting the target image into a feature extraction model to obtain a plurality of target feature maps,

and

the determining, in each of the one or more target feature maps, the first feature region
5 corresponding to the position of the first replacement region and the second feature region
corresponding to the position of the second replacement region includes:

determining, in the target feature map, first feature information corresponding to the
position of the first replacement region and second feature information corresponding to the
position of the second replacement region, where feature information of first feature regions
10 of the one or more target feature maps includes first feature information in all the one or more
target feature maps, and feature information of second feature regions of the one or more
target feature maps includes second feature information in all the one or more target feature
maps.

[0011] In a possible implementation, the feature extraction model is an unlabeled
15 knowledge distillation model.

[0012] In a possible implementation, the method further includes:

performing feature extraction on a sample image to obtain a sample feature map of the
sample image;

inputting the sample feature map into an original image generation model to obtain a
20 restored image; and

training the original image generation model based on a difference between the restored
image and the sample image to obtain the trained image generation model.

[0013] In a possible implementation, the determining the first replacement region and the
second replacement region in the target image includes: performing object detection on the
25 target image to obtain the first replacement region and the second replacement region.

[0014] In a possible implementation, the trained image generation model is a stable
diffusion model, and the inputting the one or more target composite maps into the trained
image generation model to obtain the target composite image includes:

inputting the one or more target composite maps and the target image into the stable

diffusion model obtained by training, where the one or more target composite maps serve as condition information of the stable diffusion model, and the target image serves as an input image of the stable diffusion model.

[0015] According to a second aspect of the present disclosure, an apparatus for image generation is further provided. The apparatus includes:

a first determination module, configured to determine a first replacement region and a second replacement region in a target image;

a feature extraction module, configured to perform feature extraction on the target image to obtain one or more target feature maps of the target image;

a second determination module, configured to determine, in each of the one or more target feature maps, a first feature region corresponding to a position of the first replacement region and a second feature region corresponding to a position of the second replacement region;

a feature transformation module, configured to perform, in the target feature map, transformation on feature information of the first feature region based on feature information of the second feature region, to obtain a target composite map after the feature information of the first feature region is transformed; and

an image generation module, configured to input one or more target composite maps into a trained image generation model to obtain a target composite image, where the trained image generation model is configured to perform reverse inference based on the one or more target composite maps to obtain the target composite image.

[0016] According to a third aspect of the present disclosure, a storage medium is further provided. The storage medium stores program instructions, and the program instructions, when executed, are configured to perform the method for image generation described above.

[0017] According to a fourth aspect of the present disclosure, an electronic device is further provided. The electronic device includes a processor and a memory, where the memory stores computer program instructions, and the computer program instructions, when executed by the processor, are configured to perform the method for image generation described above.

[0018] In the above technical solutions, the feature maps of the target image are modified,

and the modified feature maps are used as an input of the image generation model to generate the target composite image. Thus, the obtained target composite image is realistic, and the image effect is satisfactory. The target composite image helps to ensure the training effect of a learning model for product quality inspection.

5 **[0019]** The above description is only an overview of the technical solutions of the present disclosure. In order to understand the technical means of the present disclosure more clearly, the present disclosure can be implemented according to the content of the specification, and in order to make the above and other objectives, features, and advantages of the present disclosure more obvious and understandable, specific implementations of the present
10 disclosure are provided below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0020] By describing the embodiments of the present disclosure in more detail with reference to the drawings, the above and other objectives, features, and advantages of the
15 present disclosure are apparent. The drawings provide a further understanding of the embodiments of the present disclosure and constitute a part of the specification, and are used together with the embodiments of the present disclosure to explain the present disclosure and do not constitute a limitation to the present disclosure. In the drawings, the same reference numerals generally represent the same components or steps.

20 **[0021]** FIG. 1 shows a schematic flowchart of a method for image generation according to an embodiment of the present disclosure;

[0022] FIG. 2 shows a schematic flowchart of performing transformation on feature information of a first feature region based on feature information of a second feature region according to an embodiment of the present disclosure;

25 **[0023]** FIG. 3 shows a schematic flowchart of training an image generation model according to an embodiment of the present disclosure;

[0024] FIG. 4 shows a schematic diagram of a method for image generation according to yet another embodiment of the present disclosure;

[0025] FIG. 5 shows a schematic block diagram of an apparatus for image generation

according to an embodiment of the present disclosure; and

[0026] FIG. 6 shows a schematic block diagram of an electronic device according to an embodiment of the present disclosure.

5 DETAILED DESCRIPTION

[0027] In order to make the objectives, technical solutions, and advantages of the present disclosure more apparent, example embodiments according to the present disclosure are described in detail below with reference to the drawings. Obviously, the described embodiments are only some of the embodiments of the present disclosure, rather than all the
10 embodiments of the present disclosure. It should be understood that the present disclosure is not limited to the example embodiments described herein. Based on the embodiments of the present disclosure described herein, all other embodiments obtained by those skilled in the art without creative effort shall fall within the protection scope of the present disclosure.

[0028] The method for image generation according to the embodiments of the present
15 disclosure is based on feature distribution of an image, and different images are generated by smearing a foreground region in an existing image to generate a background region, or by smearing a background region in an existing image to generate a foreground region. Thus, more realistic images may be provided. This method for image generation may be applied to application scenarios requiring a large number of images, such as a scenario of training an
20 artificial intelligence learning model. Image samples are important input data for training an image recognition model, and realistic image samples can effectively improve the accuracy of image recognition performed by the trained model. For example, in industrial inspection, it is required to inspect appearance defects of products to eliminate defective products. At this time, an artificial intelligence model may be used to perform image recognition on the
25 obtained image of the appearance of the product to determine whether the product is a defective product. The artificial intelligence model requires a large amount of image sample data for training when performing image recognition. If the image samples are few and have low realness, the training effect of the artificial intelligence model may be poor, thereby affecting the accuracy of image recognition by using the artificial intelligence model. In

industrial inspection, real defective image samples are relatively few, which is insufficient to support the training of the artificial intelligence model for image recognition.

[0029] To at least partially solve the above issues, a method for image generation is provided according to an embodiment of the present disclosure. In this method, different
5 images are generated by modifying feature maps of an image based on image reverse inference.

[0030] FIG. 1 shows a schematic flowchart of a method for image generation according to an embodiment of the present disclosure. As shown in FIG. 1, the method includes the following steps S110 to S150.

10 **[0031]** In step S110, a first replacement region and a second replacement region are determined in a target image.

[0032] The target image may be any suitable image. In an embodiment, the target image may be an RGB image or a grayscale image. The target image may be a static image, or any video frame in a dynamic video. The target image may be an image with any suitable size and
15 resolution. A target background image may be an original image directly acquired by an image acquisition apparatus, or an image obtained after performing a pre-processing operation on the original image. The pre-processing operation includes all operations for improving the visual effect of the target image, enhancing the clarity of the target image, or highlighting certain features in the target image. Exemplarily rather than restrictively, the pre-processing
20 operation includes operations such as digitization, geometric transformation, normalization, and filtering of the original image. The target image may be a composite image on the basis of not affecting subsequent image processing.

[0033] The first replacement region and the second replacement region are different regions in the target image. The first replacement region is replaced with the second replacement
25 region. The first replacement region and the second replacement region may be determined according to requirements. If it is desired that a target composite image has a reduced foreground region relative to the target image, at least a part of a foreground region of a target replacement region may be determined as the first replacement region, and at least a part of a background region of the target replacement region may be determined as the second

replacement region. If it is desired that the target composite image has a reduced background region relative to the target image, the opposite operation is performed.

[0034] In an embodiment, positions of the first replacement region and the second replacement region are manually determined in the target image according to one's expectation, thereby obtaining the first replacement region and the second replacement region.

[0035] In step S120, feature extraction is performed on the target image to obtain a target feature map of the target image.

[0036] The feature map may be referred to as a feature matrix, and there may be one or more feature maps. In this step, a convolutional neural network is used to perform feature extraction on the target image to obtain the feature map. The convolutional neural network includes a series of convolution kernels, so as to perform convolution operations on the target image. The feature map may be regarded as an abstract representation of an original image, where each pixel includes some specific features.

[0037] In step S130, in the target feature map, a first feature region corresponding to a position of the first replacement region and a second feature region corresponding to a position of the second replacement region are determined.

[0038] Based on the position correspondence, the first feature region is determined in the target feature map based on the position of the first replacement region in the target image, and the second feature region is determined in the target feature map based on the position of the second replacement region in the target image.

[0039] In step S140, in the target feature map, transformation is performed on feature information of the first feature region based on feature information of the second feature region, to obtain a target composite map after the feature information of the first feature region is transformed.

[0040] After determining positions of the first feature region and the second feature region in step S130, the feature information of the first feature region and the feature information of the second feature region may be obtained. In step S140, based on the feature information of the second feature region in the target feature map, the feature information of the first feature region in the target feature map is transformed, whereby the feature information of the first

feature region is changed. It can be understood that the feature information of the second feature region in the target feature map is only used as a basis for transforming the feature information of the first feature region, and itself does not change.

[0041] The above transformation operation may be a replacement operation. In other words, the feature information of the first feature region is directly replaced with the feature information of the second feature region. In a possible implementation, when multiple target feature maps are obtained, in each target feature map, feature information of the first replacement region is replaced with feature information of the second replacement region.

[0042] In step S150, the target composite map is input into a trained image generation model to obtain a target composite image, where the trained image generation model is configured to perform reverse inference based on the target composite map to obtain the target composite image.

[0043] The image generation model may be a trained neural network model, in which reverse inference is performed based on the input feature map to generate an image corresponding to the feature map. It can be understood that the "reverse inference" operation is relative to the feature extraction operation. The feature extraction operation refers to extracting features from an input image, that is, the operation performed in step S120. The reverse inference operation is to obtain an image corresponding to the image features based on the input image features. The image generation model may be an existing or future developed model capable of generating an image by reverse inference based on the input feature map, which is not limited in the present disclosure.

[0044] Visually, the target composite image output by the image generation model is an image in which image information of the first replacement region is transformed based on image information of the second replacement region of the target image. The target composite image obtained through the image generation model avoids the abruptness of a replaced image caused by directly replacing the image region, making the finally obtained image more in line with the expected requirements and closer to a real image.

[0045] In the above technical solution, the feature maps of the target image are modified, and the modified feature maps are used as an input of the image generation model to generate

the target composite image. Thus, the obtained target composite image is realistic, and the image effect is satisfactory. The target composite image helps to ensure the training effect of a learning model for product quality inspection.

[0046] In an embodiment, the above first replacement region or second replacement region may be obtained by performing object detection on the target image in step S110. The object detection is used to determine the position of an interested object in an image. According to whether the region includes the object, the first replacement region or the second replacement region may be obtained through an object detection model. For example, the target image is an image of a die. The die is provided with one or more through-holes, and a region where the through-holes are located serves as a foreground region, while a remaining region serves as a background region. The region where one through-hole is located may serve as the first replacement region, and a part of the background region may serve as the second replacement region. In this example, the through-holes may be taken as the object, and object detection may be performed on the image of the die to determine the first replacement region.

[0047] The first replacement region and the second replacement region are automatically obtained through object detection, improving the response speed of the overall system while ensuring the accuracy of the system, thereby improving the efficiency of image generation.

[0048] In an embodiment, the performing feature extraction on the target image to obtain the target feature map of the target image in step S120 includes: inputting the target image into a feature extraction model to obtain multiple target feature maps. The feature extraction model may be any existing or future developed learning model for feature extraction. In an embodiment, the feature extraction model is obtained by training based on sample images. Alternatively, the feature extraction model is obtained by directly modeling based on data without training. The target feature map output by the feature extraction model may have the same or different size as the target image. It can be understood that image scaling may be performed on the target feature map output by the feature extraction model, so that the target feature map has the same size as the target image, facilitating subsequent image processing.

[0049] When multiple target feature maps are obtained, the determining, in the target feature map, the first feature region corresponding to the position of the first replacement

region and the second feature region corresponding to the position of the second replacement region in step S130 includes: in each target feature map, determining first feature information corresponding to the position of the first replacement region and second feature information corresponding to the position of the second replacement region. Feature information of first feature regions of the multiple target feature maps includes first feature information in all the multiple target feature maps, and feature information of second feature regions of the multiple target feature maps includes second feature information in all the multiple target feature maps. In this case, for the multiple target feature maps, feature values at the same pixel position form a feature vector, which may be referred to as a feature vector of a current pixel. Thus, a dimension of the multiple target feature maps is $C \times H \times W$, where C represents the number of the multiple target feature maps, and H and W represent a height and a width of the target feature maps, respectively.

[0050] For one target image, multiple target feature maps thereof are obtained to describe the features from multiple perspectives. These target feature maps serve as the basis for subsequent image generation, improving the effect of the finally obtained target composite image.

[0051] In an embodiment, the feature extraction model is an unlabeled knowledge distillation model.

[0052] The unlabeled knowledge distillation model is a self-distillation learning model, where a feature map with the same size as an input image is obtained based on the input image without using labels for training. Through the unlabeled knowledge distillation model, multiple feature maps are obtained based on one input image. Moreover, redundant information in the image is removed by using the unlabeled knowledge distillation model, thereby focusing on learning the essential features in the image.

[0053] In the above technical solution, the target feature maps of the target image are obtained by using the unlabeled knowledge distillation model, simplifying the complexity of the model and improving the generalization ability of the model. In addition, it is not required to train the unlabeled knowledge distillation model by using image samples, so that the accuracy of feature extraction can be guaranteed even if the number of the target images is

small. Thus, when there are only a few target images, the realistic effect of the target composite image generated by the method for image generation is guaranteed.

[0054] In an embodiment, the performing, in the target feature map, transformation on the feature information of the first feature region based on the feature information of the second feature region to obtain the target composite map after the feature information of the first feature region is transformed in step S140 includes: directly replacing the feature information of the first feature region with the feature information of the second feature region to obtain the target composite map. The second feature region may have the same shape and size as the first feature region. The feature information of the first feature region may be directly replaced with the feature information of the second feature region. Specifically, based on the respective positions of the first feature region and the second feature region in the target feature map, distances in different directions between the second feature region and the first feature region, such as in horizontal and vertical directions of the target feature map, are calculated. Then, the feature information of the second feature region are moved according to the calculated distances, to replace the feature information of the first feature region with the feature information of the second feature region.

[0055] In the above technical solution, the feature information of the first feature region is directly replaced with the feature information of the second feature region. In other words, in this solution, the replacement is performed on a region basis as a whole, thereby reducing the amount of computational data in the process of generating the target feature map, and improving the speed of image generation.

[0056] FIG. 2 shows a schematic flowchart of performing transformation on the feature information of the first feature region based on the feature information of the second feature region according to an embodiment of the present disclosure. As shown in FIG. 2, the performing, in the target feature map, transformation on the feature information of the first feature region based on the feature information of the second feature region to obtain the target composite map after the feature information of the first feature region is transformed in step S140 includes step S141 and step S142.

[0057] In step S141, for each of first feature values in the first feature region, a second

feature value is determined in the second feature region. In other words, the first feature value is a pixel value in the first feature region of the target feature map, and the second feature value is a pixel value in the second feature region of the target feature map.

[0058] It can be understood that the target feature map is obtained by performing feature
5 extraction on the target image, and feature values in the target feature map characterize pixels in the target image. The position of a feature value in the target feature map corresponds to the position of a pixel that the feature value characterizes in the target image. In an embodiment where the target feature map has the same size as the target image, one feature value in the target feature map characterizes one pixel in the target image. The correspondence, between
10 the position of the feature value in the target feature map and the position of the pixel that the feature value characterizes in the target image, means that horizontal and vertical coordinates of the feature value are the same as horizontal and vertical coordinates of the pixel, respectively. It can be understood that in an embodiment where the target feature map has a size different from the target image, for example, the size of the target image is larger than
15 that of the target feature map, one feature value in the target feature map may characterize multiple pixels in the target image.

[0059] In step S142, the first feature values in the first feature region are replaced with the second feature values.

[0060] In this step, the first feature values in the first feature region are traversed, and each
20 first feature value is replaced with a respective second feature value in the second feature region, until all the first feature values in the first feature region are replaced.

[0061] In the above technical solution, the replacement is performed on a feature-value basis in the target feature map. Thus, the generated target composite map has rich details, thereby ensuring that the target composite image generated by the image generation method is
25 realistic.

[0062] In an embodiment, when performing replacement on the feature information in the first feature region on the feature-value basis, step S141 includes: determining the second feature value in the second feature region in a random manner.

[0063] In this embodiment, when traversing the first feature values in the first feature

region, for each first feature value, a second feature value is randomly determined in the second feature region, instead of a second feature value corresponding to the position of the first feature value in the first feature region. Then, the feature value to be replaced in the first feature region is replaced with the randomly determined second feature value.

5 **[0064]** In the above technical solution, the first feature values in the first feature region are replaced based on the randomly determined second feature values in the second feature region. Since the second feature region corresponds to the predetermined second replacement region in the target image, the second feature values for replacement are obtained by randomly selecting the feature values in the second feature region, thereby enriching the generated
10 target composite map, reducing the amount of computation for determining the positions of the feature values during data replacement, and improving the efficiency of image generation.

[0065] FIG. 3 shows a schematic flowchart of training an image generation model according to an embodiment of the present disclosure. In an embodiment, the above method for image generation may further include the steps of training the image generation model. As
15 shown in FIG. 3, the training the image generation model includes the following steps S101 to S103.

[0066] In step S101, feature extraction is performed on a sample image to obtain a sample feature map of the sample image.

[0067] The sample image may be a real image obtained through detection, or a composite
20 image. In an embodiment, feature extraction is performed on the sample image through a trained feature extraction model to obtain the sample feature map.

[0068] Step S101 is similar to the aforementioned step S120, and for brevity, it will not be repeated here.

[0069] In step S102, the sample feature map is inputted into an original image generation
25 model to obtain a restored image.

[0070] The original image generation model may be an untrained or incompletely trained image generation model. The performance in generating images may be improved through training. The restored image is an image obtained by reverse inference performed by the original image generation model according to the input sample feature map. The restored

image is used for comparison with the sample image to determine a difference between the restored image and the sample image.

[0071] In step S103, the original image generation model is trained based on the difference between the restored image and the sample image, to obtain the trained image generation
5 model.

[0072] The difference between the restored image and the sample image characterizes the performance of the current image generation model. A greater difference indicates a poorer performance of the current image generation model; otherwise, the opposite is true. Based on the difference between the restored image and the sample image, the image generation model
10 is trained. For example, by modifying parameters of the image generation model, the performance in image generation is gradually improved until the trained image generation model that meets the requirements is obtained.

[0073] The trained image generation model may be used to perform reverse inference based on the target composite map to obtain a satisfactory target composite image. Compared with
15 the original image generation model, the image generated by the trained image generation model better meets the expected requirements and the trained image generation model is more efficient.

[0074] In an embodiment, the trained image generation model is a stable diffusion model. The inputting the target composite map into the trained image generation model to obtain the
20 target composite image in step S150 includes: inputting the target composite map and the target image into the stable diffusion model obtained by training, where the target composite map serves as condition information of the stable diffusion model, and the target image serves as an input image of the stable diffusion model.

[0075] According to an embodiment of the present disclosure, based on the regular
25 distribution of the target image, the distribution information contained in the target feature map is used as guidance to gradually denoise a noisy image and generate a target composite image matching the feature map. In an embodiment, during image generation by using the stable diffusion model, noise is gradually added to the image. As the noise is added, the stable diffusion model gradually learns features of the target image. Under the guidance of the

condition information provided by the target composite map, the desired target composite image is generated based on the target feature map.

[0076] The stable diffusion model may be based on a spatial transformer architecture, and may include a U-Net network. The spatial transformer architecture may include a cross-attention module and a basic transformer module. When receiving the target image as the input image of the stable diffusion model, the target composite map of the target image is used as the condition information, and the target composite map and the target image are modeled by using the cross-attention module. It can be understood that before modeling by using the cross-attention module, an encoder is used to encode the target composite map. In the cross-attention module, the target image serves as a query keyword, and the target composite map serves as a key value, so that the cross-attention module learns the correlation between the target image and the target composite map. The basic transformer module actually invokes the cross-attention module and generates the target composite image based on the correlation learned by the cross-attention module. In summary, when the stable diffusion model performs reverse inference of an image, in addition to inputting Gaussian noise and the target image, the target composite map corresponding to the target image is input as condition information, based on which the specified target composite image is generated.

[0077] In the above technical solution, the stable diffusion model is used as the image generation model, and the target composite map is input as the condition information to generate the target composite image. Compared with inputting text information into the stable diffusion model as the condition information, the specified image generated based on the input target composite map better meets the expected requirements.

[0078] FIG. 4 shows a schematic diagram of an image generation method according to yet another embodiment of the present disclosure.

[0079] In an embodiment, before performing image generation, the stable diffusion model is trained to obtain a trained stable diffusion model for image generation. During training, features of the sample image may be extracted through a trained unlabeled knowledge distillation model to obtain the sample feature map. For the training of the stable diffusion

model, the extracted sample feature map is encoded through an encoder in the stable diffusion model and then input into a deep network of the U-Net structure to obtain the restored image corresponding to the sample image. Based on the difference between the restored image and the sample image, the stable diffusion model is trained to obtain the trained stable diffusion model for image generation. It can be understood that the training process of the above stable diffusion model may be an iterative process, which continues until the trained stable diffusion model meets the requirements or meets other training completion conditions, so as to obtain a qualified trained stable diffusion model.

[0080] For example, as shown in FIG. 4, after acquiring the target image, multiple target feature maps of the target image are obtained through the trained unlabeled knowledge distillation model. The feature values at the same pixel position of all the target feature maps form a feature vector. The dimension of a feature set corresponding to the target feature maps output by the unlabeled knowledge distillation model is $C \times H \times W$, where C represents the number of the target feature maps, and H and W represent a height and a width of the target feature maps, respectively.

[0081] As shown in FIG. 4, a first replacement region and a second replacement region are determined in the target image. The first replacement region is a region to be smeared. The second replacement region is a region based on which the first replacement region is smeared. The first replacement region and the second replacement region have different functions but may be expressed in the same way. For brevity, the first replacement region is taken as an example for description below. The first replacement region is expressed by a position

$\text{Mask}_{x,y}$:

$$\text{Mask}_{x,y} = \{(x,y)_{ROI}\}$$

where (x, y) represents coordinates of a pixel in the target image, and $\{(x,y)_{ROI}\}$ represents a set of coordinates of pixels in the first replacement region.

[0082] After obtaining the target feature maps through the trained unlabeled knowledge distillation model, the first feature region $F(I)_{ROI}$ in each target feature map may be determined according to the first replacement region in the target image.

[0083] According to $\text{Mask}_{x,y}$, the first feature region $F(I)_{ROI}$ is determined by the following equation:

$$F(I)_{ROI} = R_{h,w}((h, w) \in \text{Mask}_{x,y})$$

where (h, w) represents coordinates of a feature value in the target feature map, the feature value corresponding to the pixel in the first replacement region, and $R_{h,w}$ represents a set of coordinates of feature values in the target feature map, the feature values corresponding to the pixels in the first replacement region.

[0084] Similarly, as shown in FIG. 4, after obtaining the target feature maps through the trained unlabeled knowledge distillation model, the second feature region in each target feature map may be determined according to the second replacement region in the target image. The data in the first feature region is replaced with the data in the second feature region.

[0085] According to the position $\text{Trget}_{x,y}$ of the second replacement region determined in the target image, the position $F(I)_{ROI'}$ of the second feature region corresponding to the second replacement region is determined in each target feature map. The position of the second replacement region may be determined manually, or a specified region may be selected as the second replacement region by an algorithm. The second feature region $F(I)_{ROI'}$ is determined by the following equation:

$$F(I)_{ROI'} = R_{h,w}((h, w) \in \text{Trget}_{x,y})$$

where $\text{Trget}_{x,y}$ represents a set of coordinates of pixels in the second replacement region, that is, the position of the second replacement region, (h, w) represents coordinates of a feature value in the target feature map, the feature value corresponding to the pixel in the second replacement region, and $R_{h,w}$ represents a set of coordinates of feature values in the target feature map, the feature values corresponding to the pixels in the second replacement region.

[0086] As shown in FIG. 4, after determining the position of the first feature region and the

position of the second feature region in each target feature map, transformation is performed on the feature information of the first feature region based on the feature information of the second feature region, so as to obtain a target composite map after the feature information of the first feature region is transformed. In each target feature map, any one second feature value is randomly determined in the second feature region. The first feature value in the first feature region is replaced based on the second feature value, until all first feature values in the first feature region in the target feature map are replaced, so as to obtain multiple target feature maps with some feature values replaced. When the first feature region has the same size and shape as the second feature region, in each target feature map, the data of the first feature region is directly filled and replaced with the data of the second feature region, so as to obtain multiple replaced target feature maps. After obtaining the multiple replaced target feature maps, the multiple replaced target feature maps are composited into one target composite map. In an embodiment, the target composite map is obtained through algorithms, manual work, or neural network models. Alternatively, the multiple replaced target feature maps are directly concatenated into one target composite map.

[0087] In an embodiment, as shown in FIG. 4, after obtaining the target composite map, the target composite map, in combination with Gaussian noise, serves as the input of the trained stable diffusion model. Reverse inference is performed through the U-Net network in the stable diffusion model to obtain the target composite image.

[0088] FIG. 5 shows a schematic block diagram of an apparatus 500 for image generation according to an embodiment of the present disclosure. As shown in FIG. 5, the apparatus 500 includes a first determination module 510, a feature extraction module 520, a second determination module 530, a feature transformation module 540, and an image generation module 550.

[0089] The first determination module 510 is configured to determine a first replacement region and a second replacement region in a target image.

[0090] The feature extraction module 520 is configured to perform feature extraction on the target image to obtain one or more target feature maps of the target image.

[0091] The second determination module 530 is configured to determine, in each of the one

or more target feature maps, a first feature region corresponding to a position of the first replacement region and a second feature region corresponding to a position of the second replacement region.

[0092] The feature transformation module 540 is configured to perform, in the target feature map, transformation on feature information of the first feature region based on feature information of the second feature region, to obtain a target composite map after the feature information of the first feature region is transformed.

[0093] The image generation module 550 is configured to input one or more target composite maps into a trained image generation model to obtain a target composite image, where the trained image generation model is configured to perform reverse inference based on the one or more target composite maps to obtain the target composite image.

[0094] According to another aspect of the present disclosure, an electronic device is further provided. FIG. 6 shows a schematic block diagram of an electronic device 600 according to an embodiment of the present disclosure. As shown in FIG. 6, the electronic device 600 includes a processor 610 and a memory 620, where the memory 620 stores computer program instructions, and the computer program instructions, when executed by the processor 610, are configured to perform the method for image generation described above.

[0095] In addition, according to yet another aspect of the present disclosure, a storage medium is further provided. Program instructions are stored on the storage medium. When the program instructions are executed by a computer or a processor, the computer or the processor is caused to execute the corresponding steps of the method for image generation according to the embodiments of the present disclosure, and to implement the corresponding modules in the apparatus for image generation or the corresponding modules in the above electronic device according to the embodiments of the present disclosure. The storage medium may, for example, include a memory card of a smart phone, a storage component of a tablet computer, a hard disk of a personal computer, a read-only memory (ROM), an erasable programmable read-only memory (EPROM), a portable compact disc read-only memory (CD-ROM), a USB memory, or any combination of the above storage media. The computer-readable storage medium may be any combination of one or more computer-readable storage media.

[0096] Those skilled in the art can understand the specific implementations and beneficial effects of the apparatus for image generation, the electronic device, and the storage medium by reading the above specific description of the method for image generation, which will not be repeated herein for brevity.

5 **[0097]** Embodiments:

[0098] Embodiment 1: A method for image generation, comprising:

determining a first replacement region and a second replacement region in a target image;

performing feature extraction on the target image to obtain one or more target feature
10 maps of the target image;

determining, in each of the one or more target feature maps, a first feature region corresponding to a position of the first replacement region and a second feature region corresponding to a position of the second replacement region;

performing, in the target feature map, transformation on feature information of the first
15 feature region based on feature information of the second feature region, to obtain a target composite map after the feature information of the first feature region is transformed; and

inputting one or more target composite maps into a trained image generation model to obtain a target composite image, wherein the trained image generation model is configured to perform reverse inference based on the one or more target composite maps to obtain the target
20 composite image.

[0099] Embodiment 2: The method according to Embodiment 1, wherein the performing transformation on the feature information of the first feature region based on the feature information of the second feature region comprises:

for each of first feature values in the first feature region, determining a second feature
25 value in the second feature region; and

replacing the first feature values in the first feature region with the second feature values.

[0100] Embodiment 3: The method according to Embodiments 1 or 2, wherein the determining the second feature value in the second feature region comprises:

determining the second feature value in the second feature region in a random manner.

[0101] Embodiment 4: The method according to any one of Embodiments 1 to 3, wherein the performing transformation on the feature information of the first feature region based on the feature information of the second feature region comprises:

5 directly replacing the feature information of the first feature region with the feature information of the second feature region.

[0102] Embodiment 5: The method according to any one of Embodiments 1 to 4, wherein

the performing feature extraction on the target image to obtain the one or more target feature maps of the target image comprises:

10 inputting the target image into a feature extraction model to obtain a plurality of target feature maps,

and

the determining, in each of the one or more target feature maps, the first feature region corresponding to the position of the first replacement region and the second feature region corresponding to the position of the second replacement region comprises:

15 determining, in the target feature map, first feature information corresponding to the position of the first replacement region and second feature information corresponding to the position of the second replacement region, wherein feature information of first feature regions of the one or more target feature maps comprises first feature information in all the one or more target feature maps, and feature information of second feature regions of the one or more target feature maps comprises second feature information in all the one or more target feature maps.

[0103] Embodiment 6: The method according to any one of Embodiments 1 to 5, wherein the feature extraction model is an unlabeled knowledge distillation model.

25 **[0104]** Embodiment 7: The method according to any one of Embodiments 1 to 6, further comprising:

performing feature extraction on a sample image to obtain a sample feature map of the sample image;

inputting the sample feature map into an original image generation model to obtain a restored image; and

training the original image generation model based on a difference between the restored image and the sample image to obtain the trained image generation model.

[0105] Embodiment 8: The method according to any one of Embodiments 1 to 7, wherein the determining the first replacement region and the second replacement region in the target image comprises:

performing object detection on the target image to obtain the first replacement region and the second replacement region.

[0106] Embodiment 9: The method according to any one of Embodiments 1 to 8, wherein the trained image generation model is a stable diffusion model, and

the inputting the one or more target composite maps into the trained image generation model to obtain the target composite image comprises:

inputting the one or more target composite maps and the target image into the stable diffusion model obtained by training, wherein the one or more target composite maps serve as condition information of the stable diffusion model, and the target image serves as an input image of the stable diffusion model.

[0107] Embodiment 10: An apparatus for image generation, comprising:

a first determination module, configured to determine a first replacement region and a second replacement region in a target image;

a feature extraction module, configured to perform feature extraction on the target image to obtain one or more target feature maps of the target image;

a second determination module, configured to determine, in each of the one or more target feature maps, a first feature region corresponding to a position of the first replacement region and a second feature region corresponding to a position of the second replacement region;

a feature transformation module, configured to perform, in the target feature map, transformation on feature information of the first feature region based on feature information of the second feature region, to obtain a target composite map after the feature information of the first feature region is transformed; and

an image generation module, configured to input one or more target composite maps into

a trained image generation model to obtain a target composite image, wherein the trained image generation model is configured to perform reverse inference based on the one or more target composite maps to obtain the target composite image.

[0108] Embodiment 11: A storage medium, storing program instructions, wherein the program instructions, when executed, are configured to perform the method for image generation according to any one of Embodiments 1 to 9.

[0109] Embodiment 12: An electronic device, comprising a processor and a memory, wherein the memory stores computer program instructions, and the computer program instructions, when executed by the processor, are configured to perform the method for image generation according to any one of Embodiments 1 to 9.

[0110] Although example embodiments have been described herein with reference to the drawings, it should be understood that the above example embodiments are merely exemplary and are not intended to limit the scope of the present disclosure thereto. Those skilled in the art can make various changes and modifications therein without departing from the scope and spirit of the present disclosure. All these changes and modifications are intended to be included within the scope of the present disclosure as claimed in the appended claims.

[0111] Those skilled in the art can appreciate that the units and algorithm steps of each example described in conjunction with the embodiments disclosed herein may be implemented by electronic hardware, or a combination of computer software and electronic hardware. Whether these functions are executed by hardware or software depends on the specific application and design constraints of the technical solution. Professional technicians can use different methods to implement the described functions for each specific application, but such implementation should not be considered beyond the scope of the present disclosure.

[0112] In the several embodiments provided in the present disclosure, it should be understood that the disclosed devices and methods may be implemented in other ways. For example, the device embodiments described above are merely illustrative. For example, the division of the units is only a logical function division, and there may be other division manners in actual implementation. For example, multiple units or components may be combined or integrated into another device, or some features may be ignored or not executed.

[0113] In the specification provided herein, a large number of specific details are described. However, it can be understood that the embodiments of the present disclosure can be practiced without these specific details. In some embodiments, well-known methods, structures, and techniques are not shown in detail so as not to obscure the understanding of this specification.

[0114] Similarly, it should be understood that in order to simplify the present disclosure and help understand one or more of the various inventive aspects, in the description of the exemplary embodiments of the present disclosure, the various features of the present disclosure are sometimes grouped together into a single embodiment, figure, or description thereof. However, the method in the present disclosure should not be interpreted as reflecting the intention that the claimed protection of the present disclosure requires more features than the features explicitly stated in each claim. Rather, as reflected by the corresponding claims, the inventive point lies in that the corresponding technical issue can be solved with features that are fewer than all the features of a single disclosed embodiment. Thus, the claims following the detailed description are hereby explicitly incorporated into the detailed description, wherein each claim itself serves as a separate embodiment of the present disclosure.

[0115] Those skilled in the art can understand that all features disclosed in this specification (including the accompanying claims, abstract, and drawings) and all processes or units of any method or device so disclosed can be combined in any combination, except where features are mutually exclusive. Unless explicitly stated otherwise, each feature disclosed in this specification (including the accompanying claims, abstract, and drawings) can be replaced by alternative features that provide the same, equivalent, or similar purpose.

[0116] In addition, those skilled in the art can understand that although some embodiments described herein comprise certain features included in other embodiments but not other features, the combination of features of different embodiments is meant to be within the scope of the present disclosure and form different embodiments. For example, in the claims, any of the claimed embodiments can be used in any combination.

[0117] Various component embodiments of the present disclosure can be implemented by

hardware, or by software modules running on one or more processors, or by a combination thereof. Those skilled in the art should understand that a microprocessor or a digital signal processor (DSP) may be used in practice to implement some or all functions of some modules in the apparatus for image generation according to the embodiments of the present disclosure.

5 The present disclosure can also be implemented as an apparatus program (for example, a computer program and a computer program product) for executing part or all of the method described herein. Such a program implementing the present disclosure can be stored on a computer-readable medium, or can have the form of one or more signals. Such signals can be downloaded from an Internet website, or provided on a carrier signal, or provided in any other
10 form.

[0118] It should be noted that the above embodiments illustrate the present disclosure rather than limit the present disclosure, and those skilled in the art can design alternative embodiments without departing from the scope of the appended claims. In the claims, any reference signs placed between parentheses should not be construed as limiting the claims.

15 The word "comprising" does not exclude the presence of elements or steps not listed in the claims. The word "a" or "an" preceding an element does not exclude the presence of a plurality of such elements. The present disclosure can be implemented by means of hardware comprising several distinct elements and by means of a suitably programmed computer. In a unit claim enumerating several apparatuses, several of these apparatuses can be embodied by
20 the same item of hardware. The use of the words "first", "second", "third", and the like does not indicate any order. These words may be interpreted as names.

[0119] The above is only the specific implementations of the present disclosure or the description of the specific implementations, and the protection scope of the present disclosure is not limited thereto. Any person familiar with the technical field can easily think of changes
25 or substitutions within the technical scope disclosed in the present disclosure, and these should be covered by the protection scope of the present disclosure. The protection scope of the present disclosure should be based on the protection scope of the claims.

CLAIMS

1. A method for image generation, comprising:

determining a first replacement region and a second replacement region in a target
5 image;

performing feature extraction on the target image to obtain one or more target feature
maps of the target image;

determining, in each of the one or more target feature maps, a first feature region
corresponding to a position of the first replacement region and a second feature region
10 corresponding to a position of the second replacement region;

performing, in the target feature map, transformation on feature information of the first
feature region based on feature information of the second feature region, to obtain a target
composite map after the feature information of the first feature region is transformed; and

inputting one or more target composite maps into a trained image generation model to
15 obtain a target composite image, wherein the trained image generation model is configured to
perform reverse inference based on the one or more target composite maps to obtain the target
composite image.

2. The method according to claim 1, wherein the performing transformation on the
20 feature information of the first feature region based on the feature information of the second
feature region comprises:

for each of first feature values in the first feature region, determining a second feature
value in the second feature region; and

replacing the first feature values in the first feature region with the second feature values.

3. The method according to claim 2, wherein the determining the second feature value in
the second feature region comprises:

determining the second feature value in the second feature region in a random manner.

4. The method according to claim 1, wherein the performing transformation on the feature information of the first feature region based on the feature information of the second feature region comprises:

5 directly replacing the feature information of the first feature region with the feature information of the second feature region.

5. The method according to any one of claims 1 to 4, wherein

the performing feature extraction on the target image to obtain the one or more target
10 feature maps of the target image comprises:

inputting the target image into a feature extraction model to obtain a plurality of target feature maps,

and

the determining, in each of the one or more target feature maps, the first feature region
15 corresponding to the position of the first replacement region and the second feature region corresponding to the position of the second replacement region comprises:

determining, in the target feature map, first feature information corresponding to the position of the first replacement region and second feature information corresponding to the position of the second replacement region, wherein feature information of first feature regions
20 of the one or more target feature maps comprises first feature information in all the one or more target feature maps, and feature information of second feature regions of the one or more target feature maps comprises second feature information in all the one or more target feature maps.

25 6. The method according to claim 5, wherein the feature extraction model is an unlabeled knowledge distillation model.

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

performing feature extraction on a sample image to obtain a sample feature map of the sample image;

inputting the sample feature map into an original image generation model to obtain a restored image; and

5 training the original image generation model based on a difference between the restored image and the sample image to obtain the trained image generation model.

8. The method according to any one of claims 1 to 4, wherein the determining the first replacement region and the second replacement region in the target image comprises:

10 performing object detection on the target image to obtain the first replacement region and the second replacement region.

9. The method according to any one of claims 1 to 4, wherein the trained image generation model is a stable diffusion model, and

15 the inputting the one or more target composite maps into the trained image generation model to obtain the target composite image comprises:

inputting the one or more target composite maps and the target image into the stable diffusion model obtained by training, wherein the one or more target composite maps serve as condition information of the stable diffusion model, and the target image serves as an input
20 image of the stable diffusion model.

10. An apparatus for image generation, comprising:

a first determination module, configured to determine a first replacement region and a second replacement region in a target image;

25 a feature extraction module, configured to perform feature extraction on the target image to obtain one or more target feature maps of the target image;

a second determination module, configured to determine, in each of the one or more target feature maps, a first feature region corresponding to a position of the first replacement

region and a second feature region corresponding to a position of the second replacement region;

a feature transformation module, configured to perform, in the target feature map, transformation on feature information of the first feature region based on feature information of the second feature region, to obtain a target composite map after the feature information of the first feature region is transformed; and

an image generation module, configured to input one or more target composite maps into a trained image generation model to obtain a target composite image, wherein the trained image generation model is configured to perform reverse inference based on the one or more target composite maps to obtain the target composite image.

11. A storage medium, storing program instructions, wherein the program instructions, when executed, are configured to perform the method for image generation according to any one of claims 1 to 9.

12. An electronic device, comprising a processor and a memory, wherein the memory stores computer program instructions, and the computer program instructions, when executed by the processor, are configured to perform the method for image generation according to any one of claims 1 to 9.