CLAIMS

1. A method for image processing, comprising:

acquiring a reference face image and a reference face pose image;

encoding the reference face image to obtain face texture data of the reference face image;

performing face key point extraction on the reference face pose image to obtain a first face mask of the reference face pose image; and

obtaining a target image according to the face texture data and the first face mask;

wherein encoding the reference face image to obtain the face texture data of the reference face image comprises:

encoding the reference face image by a plurality of successive encoding layers, to obtain the face texture data of the reference face image, wherein the plurality of encoding layers comprise an sth encoding layer and an $(s+1)^{th}$ encoding layer, input data of a first encoding layer in the plurality of encoding layers is the reference face image, output data of the sth encoding layer serves as input data of the $(s+1)^{th}$ encoding layer, and s is a positive integer greater than or equal to 1.

2. The method of claim 1, wherein obtaining the target image according to the face texture data and the first face mask comprises:

decoding the face texture data to obtain first face texture data; and

performing n stages of target processing on the first face texture data and the first face mask to obtain the target image, wherein the n stages of target processing comprise an $(m-1)^{th}$ stage of target processing and an mth stage of target processing, input data of a first stage of target processing in the n stages of target processing is the face texture data, output data of the

$(m-1)^{th}$ stage of target processing serves as input data of the mth stage of target processing, an ith stage of target processing in the n stages of target processing comprises fusing input data of the ith stage of target processing with data obtained by resizing the first face mask to obtain fused data and decoding the fused data, n is a positive integer greater than or equal to 2, m is a positive integer greater than or equal to 2 and smaller than or equal to n, and i is a positive integer greater than or equal to 1 and smaller than or equal to n.

3. The method of claim 2, wherein fusing the input data of the ith stage of target processing with the data obtained by resizing the first face mask to obtain the fused data and decoding the fused data comprises:

obtaining, according to the input data of the ith stage of target processing, to-be-fused data of the ith stage of target processing;

fusing the to-be-fused data of the ith stage of target processing with an ith stage of face mask to obtain i^{th} stage of fused data, wherein the ith stage of face mask is obtained by downsampling the first face mask, and the ith stage of face mask has a same size as the input data of the ith stage of target processing; and

decoding the ith stage of fused data to obtain output data of the ith stage of target processing.

4. The method of claim 3, after encoding the reference face image to obtain the face texture data of the reference face image, the method further comprises:

performing j stages of decoding on the face texture data, wherein input data of a first stage of decoding in the j stages of decoding is the face texture data, the j stages of decoding comprises a $(k-1)^{th}$ stage of decoding and a kth stage of decoding, output data of the $(k-1)^{th}$ stage of decoding serves as input data of the kth stage of decoding, j is a positive integer greater than or equal to 2, and k is a positive integer greater than or equal to 2 and smaller than or equal to j; and

obtaining, according to the input data of the ith stage of target processing, the to-be-fused data of the ith stage of target processing comprises:

concatenating output data of an rth stage of decoding in the j stages of decoding and the input data of the ith stage of target processing to obtain an ith stage of concatenated data as the to-be-fused data of the ith stage of target processing, wherein the output data of the rth stage of decoding has a same size as the input data of the ith stage of target processing, and r is a positive integer greater than or equal to 1 and smaller than or equal to j.

5. The method of claim 4, wherein concatenating the output data of the rth stage of decoding in the j stages of decoding and the input data of the ith stage of target processing to obtain the ith stage of concatenated data comprises:

concatenating the output data of the rth stage of decoding and the input data of the ith stage of target processing in a channel dimension to obtain the ith stage of concatenated data.

6. The method of claim 4 or 5, wherein the rth stage of decoding comprises:

sequentially performing activation, deconvolution, and normalization on input data of the rth stage of decoding to obtain the output data of the rth stage of decoding.

7. The method of any one of claims 3-6, wherein fusing the to-be-fused data of the ith stage of target processing and the ith stage of face mask to obtain the ith stage of fused data comprises:

performing convolution on the ith stage of face mask by use of a convolution kernel with a first predetermined size, to obtain first feature data, and performing convolution on the ith stage of face mask by use of a convolution kernel with a second predetermined size, to obtain second feature data;

determining a normalization form according to the first feature data and the second feature data; and

normalizing, according to the normalization form, the to-be-fused data of the ith stage of target processing to obtain the ith stage of fused data.

8. The method of claim 7, wherein the normalization form comprises a target affine transformation form; and

normalizing, according to the normalization form, the to-be-fused data of the ith stage of target processing to obtain the ith stage of fused data comprises:

performing, according to the target affine transformation form, affine transformation on the to-be-fused data of the ith stage of target processing to obtain the ith stage of fused data.

9. The method of claim 1, wherein obtaining the target image according to the face texture data and the first face mask comprises:

fusing the face texture data with the first face mask to obtain target fused data; and

decoding the target fused data to obtain the target image.

10. The method of claim 1, wherein each of the plurality of encoding layers comprises a convolution layer, a normalization layer and an activation layer.

11. The method of any one of claims 1-10, further comprising:

performing face key point extraction on the reference face image and the target image respectively to obtain a second face mask of the reference face image and a third face mask of the target image;

determining a fourth face mask according to a pixel value difference between the second face mask and the third face mask, wherein a pixel value difference between a first

pixel in the reference face image and a second pixel in the target image is positively correlated with a pixel value of a third pixel in the fourth face mask, and a position where the first pixel is located in the reference face image is the same as a position where the second pixel is located in the target image and a position where the third pixel is located in the fourth face mask; and

fusing the fourth face mask, the reference face image and the target image to obtain a new target image.

12. The method of claim 11, wherein determining the fourth face mask according to the pixel value difference between the second face mask and the third face mask comprises:

determining an affine transformation form according to an average value of a pixel value of a pixel in the second face mask and a pixel value of a pixel in the third face mask and a variance of the pixel value of the pixel in the second face mask and the pixel value of the pixel in the third face mask, wherein a position where the pixel in the second face mask is located in the second face mask is the same as a position where the pixel in the third face mask is located in the third face mask; and

performing affine transformation on the second face mask and the third face mask according to the affine transformation form to obtain the fourth face mask.

13. The method of any one of claims 1-12, applied to a face generation network, wherein

a process of training the face generation network comprises:

inputting a training sample to the face generation network to obtain a first generated image of the training sample and a first reconstructed image of the training sample, wherein the training sample comprises a sample face image and a first sample face pose image, and the first reconstructed image is obtained by encoding and decoding the sample face image;

obtaining a first loss according to a face feature matching degree between the sample

face image and the first generated image; obtaining a second loss according to a difference between face texture information in the first sample face image and face texture information in the first generated image; obtaining a third loss according to a pixel value difference between a fourth pixel in the first sample face image and a fifth pixel in the first generated image; obtaining a fourth loss according to a pixel value difference between a sixth pixel in the first sample face image and a seventh pixel in the first reconstructed image; and obtaining a fifth loss according to truthness of the first generated image, wherein a position where the fourth pixel is located in the first sample face image is the same as a position where the fifth pixel is located in the first generated image, a position where the sixth pixel is located in the first sample face image is the same as a position where the seventh pixel is located in the first reconstructed image, and higher truthness of the first generated image represents a higher probability that the first generated image is a true picture;

obtaining a first network loss of the face generation network according to the first loss, the second loss, the third loss, the fourth loss and the fifth loss; and

adjusting a parameter of the face generation network based on the first network loss.

14. The method of claim 13, wherein the training sample further comprises a second sample face pose image, the second sample face pose image is obtained by imposing random disturbance to a second sample face image to change positions of five organs in the second sample face image, or to change a position of a face contour in the second sample face image, or to change both the positions of the five organs and the position of the face contour in the second sample face image, and

the process of training the face generation network further comprises:

inputting the second sample face image and the second sample face pose image to the face generation network to obtain a second generated image of the training sample and a second reconstructed image of the training sample, wherein the second reconstructed image is obtained by encoding and decoding the second sample face image;

obtaining a sixth loss according to a face feature matching degree between the second

sample face image and the second generated image; obtaining a seventh loss according to a difference between face texture information in the second sample face image and face texture information in the second generated image; obtaining an eighth loss according to a pixel value difference between an eighth pixel in the second sample face image and a ninth pixel in the second generated image; obtaining a ninth loss according to a pixel value difference between a tenth pixel in the second sample face image and an eleventh pixel in the second reconstructed image; and obtaining a tenth loss according to truthness of the second generated image, wherein a position where the eighth pixel is located in the second sample face image is the same as a position where the ninth pixel is located in the second generated image, a position where the tenth pixel is located in the second sample face image is the same as a position where the eleventh pixel is located in the second reconstructed image, and higher truthness of the second generated image represents a higher probability that the second generated image is a true picture;

obtaining a second network loss of the face generation network according to the sixth loss, the seventh loss, the eighth loss, the ninth loss and the tenth loss; and

adjusting a parameter of the face generation network based on the second network loss.

15. The method of any one of claims 1-14, wherein acquiring the reference face image and the reference face pose image comprises:

receiving a to-be-processed face image input by a user to a terminal;

acquiring a to-be-processed video containing a face; and

obtaining a target video by taking the to-be-processed face image as the reference face image and taking each image in the to-be-processed video as the face pose image.

16. An apparatus for image processing, comprising:

an acquisition unit, configured to acquire a reference face image and a reference face

pose image;

a first processing unit, configured to encode the reference face image to obtain face texture data of the reference face image and perform face key point extraction on the reference face pose image to obtain a first face mask of the reference face pose image; and

a second processing unit, configured to obtain a target image according to the face texture data and the first face mask,

wherein the first processing unit is configured to: encode the reference face image by a plurality of successive encoding layers, to obtain the face texture data of the reference face image, wherein the plurality of encoding layers comprise an sth encoding layer and an $(s+1)^{th}$ encoding layer, input data of a first encoding layer in the plurality of encoding layers is the reference face image, output data of the sth encoding layer serves as input data of the $(s+1)^{th}$ encoding layer, and s is a positive integer greater than or equal to 1.

17. The apparatus of claim 16, wherein the second processing unit is configured to:

decode the face texture data to obtain first face texture data; and

perform n stages of target processing on the first face texture data and the first face mask to obtain the target image, wherein the n stages of target processing comprise an $(m-1)^{th}$ stage of target processing and an mth stage of target processing, input data of a first stage of target processing in the n stages of target processing is the face texture data, output data of the $(m-1)^{th}$ stage of target processing serves as input data of the mth stage of target processing, an ith stage of target processing in the n stages of target processing comprises fusing input data of the ith stage of target processing with data obtained by resizing the first face mask to obtain fused data and decoding the fused data, n is a positive integer greater than or equal to 2, m is a positive integer greater than or equal to 2 and smaller than or equal to n, and i is a positive integer greater than or equal to 1 and smaller than or equal to n.

18. The apparatus of claim 17, wherein the second processing unit is configured to:

obtain, according to the input data of the ith stage of target processing, to-be-fused data of the ith stage of target processing;

fuse the to-be-fused data of the ith stage of target processing with an ith stage of face mask to obtain i^{th} stage of fused data, wherein the ith stage of face mask is obtained by downsampling the first face mask, and the ith stage of face mask has a same size as the input data of the ith stage of target processing; and

decode the ith stage of fused data to obtain output data of the ith stage of target processing.

19. The apparatus of claim 18, further comprising:

a decoding unit, configured to: after the reference face image is encoded to obtain the face texture data of the reference face image, perform j stages of decoding on the face texture data, wherein input data of a first stage of decoding in the j stages of decoding is the face texture data, the j stages of decoding comprises a $(k-1)^{th}$ stage of decoding and a kth stage of decoding, output data of the $(k-1)^{th}$ stage of decoding serves as input data of the kth stage of decoding, j is a positive integer greater than or equal to 2, and k is a positive integer greater than or equal to 2 and smaller than or equal to j; and

wherein the second processing unit is configured to: concatenate output data of an rth stage of decoding in the j stages of decoding and the input data of the ith stage of target processing to obtain an ith stage of concatenated data as the to-be-fused data of the ith stage of target processing, wherein the output data of the rth stage of decoding has a same size as the input data of the ith stage of target processing, and r is a positive integer greater than or equal to 1 and smaller than or equal to j.

20. The apparatus of claim 19, wherein the second processing unit is configured to:

concatenate the output data of the rth stage of decoding and the input data of the ith stage of target processing in a channel dimension to obtain the ith stage of concatenated data.

21. The apparatus of claim 19 or 20, wherein the rth stage of decoding comprises:

sequentially performing activation, deconvolution, and normalization on input data of the rth stage of decoding to obtain the output data of the rth stage of decoding.

22. The apparatus of any one of claims 18-21, wherein the second processing unit is configured to:

perform convolution on the ith stage of face mask by use of a convolution kernel with a first predetermined size, to obtain first feature data, and perform convolution on the ith stage of face mask by use of a convolution kernel with a second predetermined size, to obtain second feature data;

determine a normalization form according to the first feature data and the second feature data; and

normalize, according to the normalization form, the to-be-fused data of the ith stage of target processing to obtain the ith stage of fused data.

23. The apparatus of claim 22, wherein the normalization form comprises a target affine transformation form; and

the second processing unit is configured to: perform, according to the target affine transformation form, affine transformation on the to-be-fused data of the ith stage of target processing to obtain the ith stage of fused data.

24. The apparatus of claim 16, wherein the second processing unit is configured to:

fuse the face texture data with the first face mask to obtain target fused data; and

decode the target fused data to obtain the target image.

25. The apparatus of claim 16, wherein each of the plurality of encoding layers comprises a convolution layer, a normalization layer and an activation layer.

26. The apparatus of any one of claims 16-25, further comprising:

a face key point extraction unit, configured to perform face key point extraction on the reference face image and the target image respectively to obtain a second face mask of the reference face image and a third face mask of the target image;

a determination unit, configured to determine a fourth face mask according to a pixel value difference between the second face mask and the third face mask, wherein a pixel value difference between a first pixel in the reference face image and a second pixel in the target image is positively correlated with a pixel value of a third pixel in the fourth face mask, and a position where the first pixel is located in the reference face image is the same as a position where the second pixel is located in the target image and a position where the third pixel is located in the fourth face mask; and

a fusion unit, configured to fuse the fourth face mask, the reference face image and the target image to obtain a new target image.

27. The apparatus of claim 26, wherein the determination unit is configured to:

determine an affine transformation form according to an average value of a pixel value of a pixel in the second face mask and a pixel value of a pixel in the third face mask and a variance of the pixel value of the pixel in the second face mask and the pixel value of the pixel in the third face mask, wherein a position where the pixel in the second face mask is located in the second face mask is the same as a position where the pixel in the third face mask is located in the third face mask; and

perform affine transformation on the second face mask and the third face mask according to the affine transformation form to obtain the fourth face mask.

28. The apparatus of any one of claims 16-27, wherein a method for image processing executed by the apparatus is applied to a face generation network, the apparatus for image processing is configured to execute a process of training the face generation network, and

the process of training the face generation network comprises:

inputting a training sample to the face generation network to obtain a first generated image of the training sample and a first reconstructed image of the training sample, wherein the training sample comprises a sample face image and a first sample face pose image, and the first reconstructed image is obtained by encoding and decoding the sample face image;

obtaining a first loss according to a face feature matching degree between the sample face image and the first generated image; obtaining a second loss according to a difference between face texture information in the first sample face image and face texture information in the first generated image; obtaining a third loss according to a pixel value difference between a fourth pixel in the first sample face image and a fifth pixel in the first generated image; obtaining a fourth loss according to a pixel value difference between a sixth pixel in the first sample face image and a seventh pixel in the first reconstructed image; and obtaining a fifth loss according to truthness of the first generated image, wherein a position where the fourth pixel is located in the first sample face image is the same as a position where the fifth pixel is located in the first generated image, a position where the sixth pixel is located in the first sample face image is the same as a position where the seventh pixel is located in the first reconstructed image, and higher truthness of the first generated image represents a higher probability that the first generated image is a true picture;

obtaining a first network loss of the face generation network according to the first loss, the second loss, the third loss, the fourth loss and the fifth loss; and

adjusting a parameter of the face generation network based on the first network loss.

29. The apparatus of claim 28, wherein the training sample further comprises a second sample face pose image, the second sample face pose image is obtained by imposing random disturbance to a second sample face image to change positions of five organs in the second sample face image, or to change a position of a face contour in the second sample face image, or to change both the positions of the five organs and the position of the face contour in the second sample face image, and

the process of training the face generation network further comprises:

inputting the second sample face image and the second sample face pose image to the face generation network to obtain a second generated image of the training sample and a second reconstructed image of the training sample, wherein the second reconstructed image is obtained by encoding and decoding the second sample face image;

obtaining a sixth loss according to a face feature matching degree between the second sample face image and the second generated image; obtaining a seventh loss according to a difference between face texture information in the second sample face image and face texture information in the second generated image; obtaining an eighth loss according to a pixel value difference between an eighth pixel in the second sample face image and a ninth pixel in the second generated image; obtaining a ninth loss according to a pixel value difference between a tenth pixel in the second sample face image and an eleventh pixel in the second reconstructed image; and obtaining a tenth loss according to truthness of the second generated image, wherein a position where the eighth pixel is located in the second sample face image is the same as a position where the ninth pixel is located in the second generated image, a position where the tenth pixel is located in the second sample face image is the same as a position where the eleventh pixel is located in the second reconstructed image, and higher truthness of the second generated image represents a higher probability that the second generated image is a true picture;

obtaining a second network loss of the face generation network according to the sixth loss, the seventh loss, the eighth loss, the ninth loss and the tenth loss; and

adjusting a parameter of the face generation network based on the second network loss.

30. The apparatus of any one of claims 16-29, wherein the acquisition unit is configured to:

receive a to-be-processed face image input by a user to a terminal;

acquire a to-be-processed video containing a face; and

obtain a target video by taking the to-be-processed face image as the reference face image and taking each image in the to-be-processed video as the face pose image.