CLAIMS

1. A sequence recognition method, implemented by using a sequence recognition network comprising at least an encoder network and a decoder network, the method comprising:

acquiring a to-be-processed image, the to-be-processed image comprising a to-be-recognized object sequence;

encoding the to-be-processed image by using the encoder network to obtain a first feature sequence;

decoding the first feature sequence by using the decoder network to obtain a second feature sequence; and

obtaining a sequence recognition result of the object sequence based on the second feature sequence, wherein the sequence recognition network is obtained by respectively supervising the encoder network and the decoder network,

wherein the sequence recognition network further comprises a feature extraction network; and

encoding the to-be-processed image by using the encoder network to obtain the first feature sequence comprises:

performing feature extraction on the to-be-processed image by using the feature extraction network to obtain image features; and

encoding the image features by using the encoder network to obtain the first feature sequence,

wherein performing feature extraction on the to-be-processed image by using the feature extraction network to obtain the image features comprises:

segmenting the to-be-processed image to obtain at least two image patches, wherein no overlap exists between different image patches in the at least two image patches;

performing feature extraction on each image patch to obtain an image patch feature corresponding to each image patch; and

obtaining the image features based on the image patch features,

wherein obtaining the image features based on the image patch features comprises:

combining image patch features corresponding to the at least two image patches to obtain a combined feature; and

fusing the combined feature in a first dimension to obtain the image features, which comprises:

fusing the combined feature in the first dimension by using an average pooling operation to obtain the image features, wherein the first dimension is a width dimension of the to-be-processed image.

2. The method of claim 1, wherein performing feature extraction on each image patch to obtain the image patch feature corresponding to each image patch comprises:

encoding each image patch by using a linear projection operation to obtain the image patch feature corresponding to each image patch.

3. The method of any one of claims 1 to 2, wherein encoding the image features by using the encoder network to obtain the first feature sequence comprises:

determining a positional feature, wherein the positional feature is used for indicating position information of different features of the image features;

integrating the image features and the positional feature to obtain first feature information; and

inputting the first feature information into the encoder network so as to be encoded to obtain the first feature sequence.

4. The method of claim 3, wherein the positional feature is obtained through training, and the positional feature and the image feature have a same size.

5. The method of claim 3 or 4, wherein decoding the first feature sequence by using the decoder network to obtain the second feature sequence comprises:

determining a query feature;

integrating the first feature sequence, the positional feature and the query feature to obtain second feature information; and

inputting the second feature information into the decoder network so as to be decoded to obtain the second feature sequence.

6. The method of claim 7, wherein the query feature is obtained through training, and a size of the query feature is determined by a feature dimension of the image patch feature and a sequence length of the object sequence.

7. The method of any one of claims 1 to 6, wherein the encoder network and the decoder network are respectively an encoder network and a decoder network in a

Transformer model.

8. The method of any one of claims 1 to 7, wherein in a case that the object sequence is a game token sequence, the sequence recognition result of the object sequence at least comprises one of:

a category of each game token in the game token sequence, a face value of each game token in the game token sequence, and a quantity of game tokens in the game token sequence.

9. The method of any one of claims 1 to 8, wherein the sequence recognition network is trained by:

acquiring a sample image;

encoding the sample image by using the encoder network to obtain a first sample feature sequence;

inputting the first sample feature sequence into a classifier to obtain a first sample sequence recognition result;

decoding the first sample feature sequence by using the decoder network to obtain a second sample feature sequence;

inputting the second sample feature sequence into the classifier to obtain a second sample sequence recognition result; and

training the sequence recognition network based on a target loss function, the first sample sequence recognition result and the second sample sequence recognition result to obtain a trained sequence recognition network.

10. The method of claim 9, wherein the target loss function comprises a first target loss function and a second target loss function; and training the sequence recognition network based on the target loss function, the first sample sequence recognition result and the second sample sequence recognition result comprises:

determining a first classification loss based on the first target loss function and the first sample sequence recognition result;

determining a second classification loss based on the second target loss function and the second sample sequence recognition result;

determining a total classification loss based on the first classification loss and the second classification loss; and

performing parameter optimization on the sequence recognition network by using the total classification loss.

11. The method of claim 10, wherein determining the total classification loss based on the first classification loss and the second classification loss comprises:

determining weight coefficients respectively corresponding to the first classification loss and the second classification loss, wherein the weight coefficients are obtained through training; and

determining the total classification loss based on the first classification loss, the second classification loss and the weight coefficients.

12. A sequence recognition apparatus, implemented by using a sequence recognition network comprising at least an encoder network and a decoder network, the apparatus comprising:

an acquisition unit, configured to acquire a to-be-processed image, the to-be-processed image comprising a to-be-recognized object sequence;

an encoder unit, configured to encode the to-be-processed image by using the encoder network to obtain a first feature sequence;

a decoder unit, configured to decode the first feature sequence by using the decoder network to obtain a second feature sequence; and

a recognition unit, configured to obtain a sequence recognition result of the object sequence based on the second feature sequence, wherein the sequence recognition network is obtained by respectively supervising the encoder network and the decoder network,

wherein the encoder unit comprises:

a feature extraction module configured to perform feature extraction on the to-be-processed image by using a feature extraction network to obtain image features; and

an encoder module, configured to encode the image features by using the encoder network to obtain the first feature sequence,

wherein the feature extraction module comprises:

a splitting module, configured to segment the to-be-processed image to obtain at least two image patches, where no overlap exists between different image patches in the at least two image patches; and

a feature extraction sub-module configured to perform feature extraction on each image patch to obtain an image patch feature corresponding to each image patch, and obtain the image features based on the image patch features,

wherein the feature extraction sub-module comprises:

a combination portion, configured to combine image patch features corresponding to the at least two image patches to obtain a combined feature; and

a fusing portion, configured to fuse the combined feature in a first dimension to obtain the image features, and comprising:

a fusing sub-portion, configured to fuse the combined feature in the first dimension by using an average pooling operation to obtain the image features, where the first dimension is a width dimension of the to-be-processed image.

13. An electronic device, comprising a memory and a processor, the memory storing a computer program executable on the processor, the processor executing the program to implement the steps in the method of any one of claims 1 to 11.

14. A computer-readable storage medium, storing a computer program, wherein a processor is configured to execute the computer program to implement the steps in the method of any one of claims 1 to 11.