CLAIMS

1. An object sequence recognition method, comprising:

acquiring an image comprising an object sequence;

performing feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a preset consistency condition, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range; and

determining a class of each object in the object sequence based on the feature sequence.

2. The method of claim 1, wherein the performing feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence comprises:

performing feature extraction on the image comprising the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map; and

splitting the feature map to obtain the feature sequence.

3. The method of claim 2, wherein the performing feature extraction on the image comprising

the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map comprises:

down-sampling the image comprising the object sequence using the convolutional subnetwork in a length dimension of the image comprising the object sequence in a first direction to obtain a first-dimensional feature, the first direction being different from an arrangement direction of objects in the object sequence;

extracting a feature in a length dimension of the image comprising the object sequence in a second direction based on a length of the image comprising the object sequence in the second direction to obtain a second-dimensional feature; and

obtaining the feature map based on the first-dimensional feature and the second-dimensional feature.

4. The method of claim 3, wherein the splitting the feature map to obtain the feature sequence comprises:

pooling the feature map in the first direction to obtain a pooled feature map; and

splitting the pooled feature map in the second direction to obtain the feature sequence.

5. A method for training an object sequence recognition network, comprising:

acquiring a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range;

inputting the sample image group to an object sequence recognition network to be trained, and performing feature extraction to obtain sample feature sequences;

performing class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determining a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjusting a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

6. The method of claim 5, wherein the acquiring a sample image group comprises:

acquiring a sample video stream comprising the sample object sequence;

performing sample object sequence detection on multiple frames of sample images in the sample video stream to obtain a sample position of the sample object sequence in each frame of sample image in the multiple frames of sample images; and

determining at least two frames of sample images which satisfy the preset timing condition and in which the sample positions of the sample object sequence satisfy the preset consistency condition in the multiple frames of sample images to form the sample image group.

7. The method of claim 5 or 6, wherein the inputting the sample image group to an object sequence recognition network to be trained and performing feature extraction to obtain sample feature sequences comprises:

performing feature extraction on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group; and

splitting the sample feature map of each sample image in the sample image group to obtain the sample feature sequence of each sample image in the sample image group.

8. The method of claim 7, wherein the performing feature extraction on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group comprises:

down-sampling each sample image in the sample image group using the convolutional subnetwork in a length dimension of each sample image in a first direction to obtain a first-dimensional sample feature, the first direction being different from an arrangement direction of sample objects in the sample object sequence;

extracting a feature in a length dimension of each sample image in the sample image group in a second direction based on a length of each sample image in the sample image group in the second direction to obtain a second-dimensional sample feature; and

obtaining the sample feature map of each sample image in the sample image group based on the first-dimensional sample feature and the second-dimensional sample feature.

9. The method of claim 8, wherein the splitting the sample feature map of each sample image in the sample image group to obtain the sample feature sequence of each sample image in the

sample image group comprises:

pooling the sample feature map of each sample image in the sample image group in the first direction to obtain a pooled sample feature map of each sample image in the sample image group; and

splitting the pooled sample feature map of each sample image in the sample image group in the second direction to obtain the sample feature sequence of each sample image in the sample image group.

10. The method of any one of claims 5-9, wherein the adjusting a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition comprises:

performing weighted fusion on the first loss and the second loss set to obtain a total loss; and

adjusting the network parameter of the object sequence recognition network to be trained according to the total loss to make the loss of the classification result output by the adjusted object sequence recognition network satisfy the convergence condition.

11. The method of claim 10, wherein the performing weighted fusion on the first loss and the second loss set to obtain a total loss comprises:

adjusting the first loss using a first preset weight to obtain a third loss;

determining a class supervision weight based on the number of the sample images in the sample image group, multiple different sample images in the same sample image group corresponding to the same class supervision weight;

fusing second losses in the second loss set based on the class supervision weight and a second preset weight to obtain a fourth loss; and

determining the total loss based on the third loss and the fourth loss.

12. The method of claim 11, wherein the fusing second losses in the second loss set based on the class supervision weight and a second preset weight to obtain a fourth loss comprises:

assigning the class supervision weight to each second loss in the second loss set to obtain an updated loss set comprising at least two updated losses;

fusing the updated losses in the updated loss set to obtain a fused loss; and

adjusting the fused loss using the second preset weight to obtain the fourth loss.

13. An object sequence recognition apparatus, comprising:

a first acquisition module, configured to acquire an image comprising an object sequence;

a first extraction module, configured to perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range; and

a first determination module, configured to determine a class of each object in the object sequence based on the feature sequence.

14. An apparatus for training an object sequence recognition network, comprising:

a second acquisition module, configured to acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range;

a second extraction module, configured to input the sample image group to an object sequence recognition network to be trained and perform feature extraction to obtain sample feature sequences;

a second prediction module, configured to perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

a second determination module, configured to determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

a first adjustment module, configured to adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

15. A computer device, comprising a memory and a processor, wherein a computer-executable instruction is stored in the memory; and when executing the computer-executable instruction in the memory, the processor is configured to:

acquire an image comprising an object sequence;

perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a preset consistency condition, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range; and

determine a class of each object in the object sequence based on the feature sequence.

16. A computer device, comprising a memory and a processor, wherein a computer-executable instruction is stored in the memory; and when executing the computer-executable instruction in the memory, the processor is configured to:

acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range;

input the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain sample feature sequences;

perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

17. A computer storage medium, in which a computer-executable instruction is stored, wherein when executed by a processor, the computer-executable instruction is configured to:

acquire an image comprising an object sequence;

perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a

preset consistency condition, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range; and

determine a class of each object in the object sequence based on the feature sequence.

18. A computer storage medium, in which a computer-executable instruction is stored, wherein when executed by a processor, the computer-executable instruction is configured to:

acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range;

input the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain sample feature sequences;

perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjust a network parameter of the object sequence recognition network to be trained

according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

19. A computer program, comprising computer instructions executable by an electronic device, wherein when executed by a processor in the electronic device, the computer instructions are configured to:

acquire an image comprising an object sequence;

perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a preset consistency condition, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range; and

determine a class of each object in the object sequence based on the feature sequence.

20. A computer program, comprising computer instructions executable by an electronic device, wherein when executed by a processor in the electronic device, the computer instructions are configured to:

acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling

information of a sample object sequence, and wherein the preset consistency condition is that a difference between the positions of the sample object sequence in the frames of sample image in the sample image group does not exceed a preset difference range;

input the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain sample feature sequences;

perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.