WHAT IS CLAIMED IS:

1. A method for video analysis, comprising:

obtaining a video stream;

for each of a plurality of frames in the video stream, determining a reference score corresponding to the frame based on a spatiotemporal feature of the frame, wherein the spatiotemporal feature of the frame associates a feature of the frame with a feature of at least one adjacent frame of the frame;

determining, from the plurality of frames in the video stream, one or more target frames based on reference scores corresponding to the plurality of frames.

2. The method of claim 1, wherein for each of the plurality of frames in the video stream, the determining the reference score corresponding to the frame based on the spatiotemporal feature of the frame includes:

determining, using a video analysis model, the reference score corresponding to the frame based on the spatiotemporal feature of the frame, wherein the video analysis model is a trained machine learning model.

3. The method of claim 2, wherein the video analysis model includes a video analysis network including:

a feature extraction layer configured to extract features of the plurality of frames in the video stream;

a feature association layer configured to determine the spatiotemporal feature of the frame by associating the feature of the frame with the feature of the at least one adjacent frame of the frame; and

a prediction layer configured to determine the reference score corresponding to the frame based on the spatiotemporal feature of the frame.

4. The method of claim 3, wherein the video analysis model further includes:

a plurality of cache units each of which is connected to the video analysis network and configured to cache a same number of frames as a number of frames that are inputted into the video analysis network in a single processing of the video analysis network.

5. The method of claim 3, wherein the video analysis network is obtained by a training process including:

obtaining a plurality of sample video streams, wherein each of the plurality of sample video streams involves an abnormal event and includes at least one first frame corresponding to a moment of the abnormal event and a plurality of second frames corresponding to other moments;

for each of the plurality of sample video streams, labeling a plurality of sample frames in the sample video stream, wherein the at least one first frame is labeled with a first score, each of the plurality of second frames is labeled with a second score, and the second score is determined based on a similarity between the at least one first frame and the second frame; and

determining the video analysis network by training a preliminary video analysis network based on the plurality of labeled sample video streams.

6. The method of claim 5, wherein the determining the video analysis network by training the preliminary video analysis network based on the plurality of labeled sample video streams includes:

for each of the plurality of labeled sample video streams,

determining, using the preliminary video analysis network, a prediction score corresponding to each sample frame in the labeled sample video stream; and

determining a loss function based on prediction scores and the labeled scores corresponding to the plurality of sample frames in the labeled sample video stream; and

determining the video analysis network by training the preliminary video analysis network based on the plurality of labeled sample video streams and corresponding values of the loss function.

7. The method of claim 6, wherein the determining the loss function based on prediction scores and the labeled scores corresponding to the plurality of sample frames in the labeled sample video stream includes:

for each pair of adjacent sample frames in the plurality of sample frames, determining a first difference of prediction scores corresponding to the pair of adjacent sample frames and a second difference of labeled scores corresponding to the pair of adjacent sample frames;

determining a deviation term based on first differences and second differences corresponding to multiple pairs of adjacent sample frames in the plurality of sample frames; and

determining the loss function based on the deviation term.

8. The method of claim 3, wherein the video analysis network is obtained by a training process including:

constructing the feature extraction layer in the video analysis network by a transfer learning; and

determining the feature association layer and the prediction layer by training a preliminary feature association layer and a preliminary prediction layer based on a plurality of labeled sample video streams.

9. The method of claim 1, wherein for each of the plurality of frames in the video stream, the determining the reference score corresponding to the frame based on the spatiotemporal feature of the frame includes:

determining an analysis region in the video stream; and

for each of the plurality of frames in the video stream, determining the reference score corresponding to the frame based on the spatiotemporal feature of a portion of the frame in the analysis region.

10. The method of claim 1, wherein the determining, from the plurality of frames in the video stream, the one or more target frames based on the reference scores corresponding to the plurality of frames includes:

for each of the plurality of frames in the video stream,

determining whether the reference score corresponding to the frame is larger than a first threshold;

in response to determining that the reference score corresponding to the frame is larger than the first threshold, determining a similarity degree between the frame and a previous frame of the frame;

determining whether the similarity degree is less than a second threshold; and

in response to determining that the similarity degree is less than the second threshold, designating

the frame as a target frame.

11. The method of claim 1, wherein the method further includes:

for each of the one or more target frames, determining, using a detection model, a detection result based on the target frame, wherein the detection result includes at least one of an occurrence region, in the target frame, of an event involved in the target frame or a probability that a type of the event is the same as a type of at least one abnormal event; and

determining an analysis result of the video stream based on detection results corresponding to the one or more target frames.

12. The method of claim 11, wherein the determining the analysis result of the video stream based on the detection results corresponding to the one or more target frames includes:

for each of the one or more target frames,

determining whether a center of the occurrence region of the event associated with the target frame is located in an analysis region in the video stream;

in response to determining that the center of the occurrence region of the event associated with the target frame is located in the analysis region in the video stream, determining whether the probability is larger than a probability threshold associated with the at least one abnormal event;

in response to determining that the probability is larger than the probability threshold associated with the at least one abnormal event, determining whether the occurrence region of the event associated with the target frame satisfies a condition; and

in response to determining that the detection result corresponding to the target frame satisfies the condition, determining the analysis result of the video stream based on the detection result corresponding to the target frame.

13. The method of claim 12, wherein the condition includes:

an intersection-over-union (IOU) of the occurrence region of the event in the detection result corresponding to the target frame and an occurrence region of the event in a candidate detection result in the detection results corresponding to the one or more target frames is less than a third threshold, wherein the candidate detection result is determined before the detection result and a probability in the candidate detection result is same as the probability in the detection result; or

an interval between the target frame corresponding to the detection result and a target frame corresponding to a previous detection result of the detection result is larger than an interval threshold.

14. A system, comprising:

at least one storage device including a set of instructions; and

at least one processor in communication with the at least one storage device, wherein when executing the set of instructions, the at least one processor is directed to perform operations including:

obtaining a video stream;

for each of a plurality of frames in the video stream, determining a reference score corresponding to the frame based on a spatiotemporal feature of the frame, wherein the spatiotemporal feature of the frame associates a feature of the frame with a feature of at least one adjacent frame of the frame;

determining, from the plurality of frames in the video stream, one or more target frames based on reference scores corresponding to the plurality of frames.

15. The system of claim 14, wherein for each of the plurality of frames in the video stream, the determining the reference score corresponding to the frame based on the spatiotemporal feature of the frame includes:

determining, using a video analysis model, the reference score corresponding to the frame based on the spatiotemporal feature of the frame, wherein the video analysis model is a trained machine learning model.

16. The system of claim 15, wherein the video analysis model includes a video analysis network including:

a feature extraction layer configured to extract features of the plurality of frames in the video stream;

a feature association layer configured to determine the spatiotemporal feature of the frame by associating the feature of the frame with the feature of the at least one adjacent frame of the frame; and

a prediction layer configured to determine the reference score corresponding to the frame based on the spatiotemporal feature of the frame.

17. The system of claim 16, wherein the video analysis network is obtained by a training process including:

obtaining a plurality of sample video streams, wherein each of the plurality of sample video streams involves an abnormal event and includes at least one first frame corresponding to a moment of the abnormal event and a plurality of second frames corresponding to other moments;

for each of the plurality of sample video streams, labeling a plurality of sample frames in the sample video stream, wherein the at least one first frame is labeled with a first score, each of the plurality of second frames is labeled with a second score, and the second score is determined based on a similarity between the at least one first frame and the second frame; and

determining the video analysis network by training a preliminary video analysis network based on the plurality of labeled sample video streams.

18. The system of claim 17, wherein the determining the video analysis network by training the preliminary video analysis network based on the plurality of labeled sample video streams includes:

for each of the plurality of labeled sample video streams,

determining, using the preliminary video analysis network, a prediction score corresponding to each sample frame in the labeled sample video stream; and

determining a loss function based on prediction scores and the labeled scores corresponding to the plurality of sample frames in the labeled sample video stream; and

determining the video analysis network by training the preliminary video analysis network based on the

plurality of labeled sample video streams and corresponding values of the loss function.

19. The system of claim 18, wherein the determining the loss function based on prediction scores and the labeled scores corresponding to the plurality of sample frames in the labeled sample video stream includes:

for each pair of adjacent sample frames in the plurality of sample frames, determining a first difference of prediction scores corresponding to the pair of adjacent sample frames and a second difference of labeled scores corresponding to the pair of adjacent sample frames;

determining a deviation term based on first differences and second differences corresponding to multiple pairs of adjacent sample frames in the plurality of sample frames; and

determining the loss function based on the deviation term.

20. A non-transitory computer readable medium, comprising executable instructions that, when executed by at least one processor, direct the at least one processor to perform a method, the method comprising:

obtaining a video stream;

for each of a plurality of frames in the video stream, determining a reference score corresponding to the frame based on a spatiotemporal feature of the frame, wherein the spatiotemporal feature of the frame associates a feature of the frame with a feature of at least one adjacent frame of the frame;

determining, from the plurality of frames in the video stream, one or more target frames based on reference scores corresponding to the plurality of frames.