

MODEL TRAINING AND IMAGE PROCESSING METHODS AND APPARATUSES, ELECTRONIC DEVICE, AND STORAGE MEDIUM

TECHNICAL FIELD

[0001] The disclosure relates, but not limited, to the technical field of computers, and particularly to model training and image processing methods and apparatuses, an electronic device, and a storage medium.

BACKGROUND

[0002] With the popularization of electronic devices, a large number of videos are generated everyday with increasing resolutions (such as 1080P or higher). Different videos may be applied to various fields, such as video editing, object attribute editing technologies, and Video Quality Assessment (VQA). When quality of a video is assessed, an assessment result may be affected by a resolution of the video, a size of the video, device performance, and other factors.

SUMMARY

[0003] Embodiments of the disclosure provide model training and image processing methods and apparatuses, an electronic device, and a storage medium.

[0004] The technical solutions of the embodiments of the disclosure are implemented as follows.

[0005] The embodiments of the disclosure provide an image processing method, which may include the following operations.

[0006] At least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image.

- [0007] A second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.
- [0008] First quality information of each second image frame sequence is determined.
- [0009] Second quality information of the video to be processed is determined based on each first quality information.
- [0010] The embodiments of the disclosure provide a model training method, which may include the following operations.
- [0011] At least one first image frame sequence is determined based on at least one first video sample, each first image frame sequence including at least one frame of first image.
- [0012] A second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.
- [0013] First quality information of each second image frame sequence is determined by use of a first neural network model to be trained.
- [0014] Model parameters of the first neural network model are updated at least once based on each first quality information to obtain a trained first neural network model.
- [0015] The embodiments of the disclosure provide an image processing apparatus, which may include a first determination module, a second determination module, a third determination module, and a fourth determination module.
- [0016] The first determination module may be configured to determine at least one first image frame sequence based on a video to be processed, each first image frame

sequence including at least one frame of first image.

[0017] The second determination module may be configured to determine, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[0018] The third determination module may be configured to determine first quality information of each second image frame sequence.

[0019] The fourth determination module may be configured to determine second quality information of the video to be processed based on each first quality information.

[0020] The embodiments of the disclosure provide a model training apparatus, which may include a fifth determination module, a sixth determination module, a seventh determination module, and a first updating module.

[0021] The fifth determination module may be configured to determine at least one first image frame sequence based on at least one first video sample, each first image frame sequence including at least one frame of first image.

[0022] The sixth determination module may be configured to determine, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[0023] The seventh determination module may be configured to determine first quality information of each second image frame sequence by use of a first neural network model to be trained.

[0024] The first updating module may be configured to update model parameters of

the first neural network model at least once based on each first quality information to obtain a trained first neural network model.

[0025] The embodiments of the disclosure provide an electronic device, which may include a processor and a memory. The memory may store a computer program capable of running in the processor. The processor may execute the computer program to implement the method.

[0026] The embodiments of the disclosure provide a computer-readable storage medium having stored thereon a computer program that, when executed by a processor, causes the method to be implemented.

[0027] The embodiments of the disclosure provide a computer program product, which may include a non-transitory computer-readable storage medium storing a computer program. The computer program may be read and executed to implement the method.

[0028] In the embodiments of the disclosure, at least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image. A second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image. First quality information of each second image frame sequence is determined. Second quality information of the video to be processed is determined based on each first quality information. As such, each first image block in the first image in the first image frame sequence is sampled to obtain each first pixel block in the second image in the corresponding second image frame sequence. On one hand, the second image is the same as the first image in resolution, so that quality information of the original video may be retained, and the accuracy of VQA may further be improved. On the other hand, a size of the second image is smaller than that of the first image, so that the calculation cost for quality assessment may be reduced, and the efficiency of VQA may further be improved.

[0029] It is to be understood that the above general description and the following detailed description are only exemplary and explanatory and not intended to limit the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0030] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments consistent with the disclosure and, together with the specification, serve to describe the technical solutions of the disclosure.

[0031] FIG. 1 is an implementation flowchart of an image processing method according to an embodiment of the disclosure.

[0032] FIG. 2 is an implementation flowchart of an image processing method according to an embodiment of the disclosure.

[0033] FIG. 3 is an implementation flowchart of an image processing method according to an embodiment of the disclosure.

[0034] FIG. 4 is an implementation flowchart of a model training method according to an embodiment of the disclosure.

[0035] FIG. 5 is an implementation flowchart of a model training method according to an embodiment of the disclosure.

[0036] FIG. 6A is a composition structure diagram of an image processing system according to an embodiment of the disclosure.

[0037] FIG. 6B is a schematic diagram of processing two frames of first images by a preprocessing portion according to an embodiment of the disclosure.

[0038] FIG. 6C is a schematic composition diagram of a network portion according to an embodiment of the disclosure.

[0039] FIG. 6D is a schematic diagram of a Gated Relative Position Bias (GRPB)

according to an embodiment of the disclosure.

[0040] FIG. 6E is a schematic diagram of processing a video frame image by FAST-VQA according to an embodiment of the disclosure.

[0041] FIG. 7 is a composition structure diagram of an image processing apparatus according to an embodiment of the disclosure.

[0042] FIG. 8 is a composition structure diagram of a model training apparatus according to an embodiment of the disclosure.

[0043] FIG. 9 is a schematic diagram of a hardware entity of an electronic device according to an embodiment of the disclosure.

DETAILED DESCRIPTION

[0044] In order to make the objectives, technical solutions, and advantages of the disclosure clearer, the disclosure will further be described below in combination with the drawings in detail. The described embodiments should not be considered as limitations on the disclosure. All other embodiments obtained by those of ordinary skill in the art without creative work shall fall within the scope of protection of the disclosure.

[0045] "Some embodiments" involved in the following descriptions describes a subset of all possible embodiments. However, it can be understood that "some embodiments" may be the same or different subsets of all the possible embodiments, and may be combined with one another without conflicts.

[0046] Term "first/second/third" involved in the following descriptions is only for distinguishing similar objects and does not represent a specific sequence of the objects. It can be understood that "first/second/third" may be interchanged to specific sequences or orders if allowed, so as to implement the embodiments of the disclosure described herein in sequences except the illustrated or described ones.

[0047] Unless otherwise defined, all technical and scientific terms used herein have

the same meanings as commonly understood by those skilled in the art that the disclosure belongs to. The terms used herein are only for a purpose of describing the embodiments of the disclosure and not intended to limit the disclosure.

[0048] With the popularization of electronic devices, a large number of videos are generated everyday with increasing resolutions (such as 1080P or higher). Different videos may be applied to various fields, such as video editing, object attribute editing technologies, and VQA. When quality assessment is performed on a video, an assessment result may be affected by a resolution of the video, a size of the video, device performance, and other factors. In a related art, a deep neural network-based VQA method is large in calculation amount and high in complexity during processing of a high-resolution video. As a result, the calculation cost (such as an internal memory and a Graphics Processing Unit (GPU)) is relatively high, the time consumption is relatively high, and the efficiency of VQA is further affected. Although the calculation cost or the assessment time is reduced in part of VQA methods by resolution reduction, cropping, etc., such processing may damage the quality of the original video. For example, resolution reduction may cause the loss of part of local texture information of a video frame image, thereby reducing the accuracy of VQA.

[0049] The embodiments of the disclosure provide an image processing method. Each first image block in a first image in a first image frame sequence is sampled to obtain each first pixel block in a second image in a corresponding second image frame sequence. On one hand, the second image is the same as the first image in resolution, so that quality information of an original video may be retained, and the accuracy of VQA may further be improved. On the other hand, a size of the second image is smaller than that of the first image, so that the calculation cost for quality assessment may be reduced, and the efficiency of VQA may further be improved. Both the image processing method and model training method provided in the embodiments of the disclosure may be executed by an electronic device. The electronic device may be various types of terminals, such as a notebook computer, a tablet computer, a desktop computer, a set-top box, and a mobile device (e.g., a mobile phone, a portable music player, a personal digital assistant, a dedicated messaging device, and a portable game device), or may be implemented as a server. The server may be an independent physical server, or a server cluster or

distributed system consisting of multiple physical servers, or a cloud server providing basic cloud computing service such as cloud service, a cloud database, cloud computing, a cloud function, cloud storage, network service, cloud communication, middleware service, domain name service, security service, a Content Delivery Network (CDN), and a big data and artificial intelligence platform.

[0050] The technical solutions in the embodiments of the disclosure will be described clearly and completely below in combination with the drawings in the embodiments of the disclosure.

[0051] FIG. 1 is an implementation flowchart of an image processing method according to an embodiment of the disclosure. As shown in FIG. 1, the method includes the following steps S11 to S14.

[0052] In S11, at least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image.

[0053] Here, the video to be processed may be a video of any suitable resolution, such as 1080P (1920×1080), 720P (1280×720), 2K (2048×1080), and 4K (3840×2160). In some implementation modes, a manner for acquiring the video to be processed may be determined according to a practical application scenario.

[0054] For example, the video to be processed may be uploaded or set on an operation interface by a user through an input component of an electronic device. The input component may include, but not limited to, a keyboard, a mouse, a touch screen, a touch pad, an audio input unit, etc. The operation interface includes an interaction interface for configuration operations on the video to be processed and information presentation. The operation interface may be displayed on any suitable electronic device with an interface interaction function. During implementation, the electronic device displaying the operation interface may be the same as or different from a device executing the image processing method, and no limits are made herein. For example, the electronic device executing the image processing method may be a notebook computer, the electronic device displaying the operation interface may also be the notebook computer, and the

operation interface may be an interaction interface of a client running in the notebook computer, or a web page displayed in a browser running in the notebook computer. For another example, the electronic device executing the image processing method may be a server, the electronic device displaying the operation interface may be a notebook computer, the operation interface may be an interaction interface of a client running in the notebook computer, or a web page displayed in a browser running in the notebook computer, and the notebook computer may access the server through the client or the browser.

[0055] For another example, the video to be processed sent by another electronic device may be received.

[0056] During implementation, those skilled in the art may determine the acquisition mode of the video to be processed as practically required, and no limits are made in the embodiments of the disclosure.

[0057] In some implementation modes, the video to be processed is sampled based on a preset sampling rule to obtain at least one first image frame sequence. The preset sampling rule may include, but not limited to, a sampling starting point, a sampling stride, sampling time, a sampling mode, etc. The sampling mode may include, but not limited to, uniform sampling, Gaussian distribution sampling, periodic sampling, fixed-frame-number sampling, etc. During implementation, those skilled in the art may determine the sampling rule as practically required, and no limits are made in the embodiments of the disclosure.

[0058] For example, if the video to be processed includes 100 frames of first images, the second, fourth,, and 64th frames of first images are determined as a first image frame sequence based on a preset sampling stride (such as two frames).

[0059] For another example, if the video to be processed includes 100 frames of first images, the tenth to 41st frames of first images are determined as a first image frame sequence based on a preset sampling starting point (such as the tenth frames).

[0060] For another example, if the video to be processed includes 100 frames of first images, the first to 30th, 31st to 60th, 61st to 90th, and 70th to 100th frames of first

images are determined as at least one first image frame sequence based on a preset sampling frame number (such as 30 frames).

[0061] For another example, 100 frames of first images in the video to be processed are determined as a first image frame sequence.

[0062] In some implementation modes, the sampling rule may be set according to a preset mode. The preset mode may include, but not limited to, a default sampling configuration of the electronic device, a customized definition, a user preference, a usage frequency, user operation information, etc. During implementation, those skilled in the art may set the preset mode independently as practically required, and no limits are made in the disclosure. For example, the electronic device provides a configuration option, and the user may customize the sampling rule through the configuration option.

[0063] In some implementation modes, the video to be processed is divided based on a preset partition rule to obtain at least one first image frame sequence. The preset partition rule may include, but not limited to, a partition starting point, a partition mode, etc. The partition mode may include, but not limited to, uniform partition, partition based on type, definition, quality, etc., random partition, etc. During implementation, those skilled in the art may determine the partition rule as practically required, and no limits are made in the embodiments of the disclosure.

[0064] For example, if the video to be processed includes 100 frames of first images, the video to be processed is divided uniformly to obtain four first image frame sequences. That is, the 0th to 24th frames of first images are taken as first image frame sequence 1, the 25th to 49th frames of first images are taken as first image frame sequence 2, the 50th to 74th frames of first images are taken as first image frame sequence 3, and the 75th to 99th frames of first images are taken as first image frame sequence 4.

[0065] For another example, if the video to be processed includes 100 frames of first images, where 30 frames of first images include a first object, 25 frames of first images include a second object, and the other first images include another object, the 30 frames of first images including the first object are determined as first image frame sequence 1, the 25 frames of first images including the second object are determined as first image

frame sequence 2, and the first images including the other object are determined as first image frame sequence 3. Here, the first object, the second object, and the other object may be objects of the same type or different types. Here, the type of the object may include, but not limited to, person, real object, animal, virtual object, background, etc. For example, the first object and the second object refer to different persons, and the other object may be a background. For another example, the first object is a person, the second object is a virtual building, and the other object is a background. During implementation, those skilled in the art may determine the partition of the types of the objects as practically required, and no limits are made in the embodiments of the disclosure.

[0066] In some implementation modes, each frame of first image of the video to be processed may include at least one object, such as at least one of a first object, a second object or another object. During implementation, object recognition is performed on each frame of first image of the video to be processed to obtain at least one object in the first image.

[0067] In S12, a second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[0068] Here, a sampling mode may include, but not limited to, a sampling mode based on a preset strategy, random sampling, uniform sampling, etc. The preset strategy may include, but not limited to, a position, a size, a distance, etc. A size of the second image is not larger than that of the first image.

[0069] For example, if the size of the first image is $H*W$, the first image is divided into $M*N$ first image blocks, a size of each first image block being $(H/M)*(W/N)$. Taking the top-left corner as a starting point, each first image block is sampled based on a preset size of $A*B$ to generate each first pixel block. In such case, if the top-left corner of the first image is an origin of coordinates, a coordinate position of the first pixel block in a first row and a first column is $[0:A, 0:B]$, a coordinate position of the first pixel

block in the first row and a second column is $[H/M:H/M+A, 0:B]$,, and a coordinate position of the first pixel block in an i -th row and a j -th column is $[(i \times H)/M:(i \times H)/M+A, (j \times W)/N:(j \times W)/N+B]$. The size of the second image is $(A \times N) \times (B \times M)$.

[0070] For another example, if the size of the first image is $H \times W$. The first image is divided into $M \times M$ first image blocks, a size of each first image block being $(H/M) \times (W/M)$. Each first image block is randomly sampled based on a preset size of $C \times C$ to generate each first pixel block ($C \times C$). In such case, the size of the second image is $(M \times C) \times (M \times C)$.

[0071] For another example, if the size of the first image is $H \times W$, the first image is divided into $N \times N$ first image blocks, a size of each first image block being $(H/N) \times (W/N)$. A pixel block whose size is $D \times D$ is randomly selected from each first image block in the first image, and the pixel block is determined as a first pixel block ($D \times D$) corresponding to the first image block. In such case, the size of the second image is $(N \times D) \times (N \times D)$.

[0072] In some implementation modes, each frame of second image is obtained based on at least one first pixel block selected from at least one first image block in a corresponding first image.

[0073] For example, a pixel block whose size is $A \times B$ is randomly selected from each first image block in the first image, the pixel block is determined as a first pixel block corresponding to the first image block, and each first pixel block is spliced to obtain a second image corresponding to the first image. Both A and B are positive integers. A and B may be the same or different in magnitude.

[0074] In some implementation modes, the second image frame sequence is determined by use of a trained neural network model based on the first image frame sequence.

[0075] The trained neural network model may be any suitable neural network model capable of obtaining an image frame sequence. For example, the trained neural network model is a trained first neural network model obtained by training a first neural network model by use of a video sample group. For another example, the trained neural network model is a second neural network model obtained by finely adjusting the trained first

neural network model.

[0076] In some implementation modes, the trained neural network model may include, but not limited to, a preprocessing portion, etc. The preprocessing portion may include, but not limited to, at least one of a partition portion, a sampling portion, an alignment portion, a splicing portion, etc.

[0077] The partition portion is configured to divide a first image based on a preset partition rule to generate at least one first image block. In some implementation modes, the partition rule may include, but not limited to, uniform partition (namely each first image block is the same in size), irregular partition, etc. During implementation, those skilled in the art may determine the partition rule independently as practically required, and no specific limits are made in the disclosure.

[0078] For example, if the size of the first image is $H*W$, the first image is divided into $M*N$ first image blocks, a size of each first image block being $(H/M)*(W/N)$. For another example, if the size of the first image is $H*W$, the first image is divided into $M*M$ first image blocks, a size of each first image block being $(H/M)*(W/M)$.

[0079] The sampling portion is configured to sample each first image block based on a preset rule to obtain a first pixel block corresponding to each first image block. Sampling based on the preset rule may include, but not limited to, a sampling mode, a sampling position, a sampling size, etc. In some implementation modes, the sampling mode may include, but not limited to, random sampling, uniform sampling, etc. During implementation, those skilled in the art may determine the sampling mode independently as practically required, and no specific limits are made in the disclosure.

[0080] For example, taking a preset position as a starting point, each first image block is uniformly sampled based on a preset size of $A*B$ to generate each first pixel block. A may be the same as or different from B . The preset position may include, but not limited to, a top-left corner, a central point, a specified position, etc. For example, for first image block 1, the preset position may be $(0, 0)$, $(H/2, W/2)$, $(0, W/2)$, and $(H/2, 0)$, where H represents a width of the first image block, and W represents a height of the first image block. During implementation, those skilled in the art may determine the preset position

independently as practically required, and no specific limits are made in the disclosure.

[0081] For another example, each first image block is randomly sampled based on a preset size of $C \times C$ to generate each first pixel block.

[0082] For another example, a pixel block ($D \times D$) is randomly selected from each first image block based on a preset size of $D \times D$, and the pixel block is determined as a first pixel block corresponding to the first image block.

[0083] The alignment portion is configured to align sampling positions of first image blocks at the same position in every two frames of first images. As such, the sampling positions of all frames of first images may be the same.

[0084] In some implementation modes, each first image block in the first image in the first image frame sequence may be sampled based on a reference position. As such, variation information between adjacent frames of first images in the first image frame sequence may be retained. During implementation, those skilled in the art may determine the reference position independently as practically required, and no limits are made in the disclosure.

[0085] In some implementation modes, a sampling position of each first image block in a certain frame of first image in the first image frame sequence may be determined as a reference position. For example, a sampling position of each first image block in a first frame of first image is determined as a reference position. For example, each first image is divided into seven rows and seven columns. In such case, when a third image block (in a zeroth row and a second column) in a fifth frame of first image is sampled, a sampling position of a third image block (in a zeroth row and a second column) in the first frame of first image may be acquired at first, the sampling position is determined as a reference position, and the third image block in the fifth frame of first image is sampled according to the reference position.

[0086] In some implementation modes, for a present frame of first image, a sampling position of each first image block in a previous frame of first image is determined as a reference position. For example, for a third frame of first image, a sampling position of each first image block in a second frame of first image is determined as a reference

position, and sampling is performed at each reference position. For example, each first image is divided into four rows and four columns. In such case, when a fifth image block (in a first row and a zeroth column) in the third frame of first image is sampled, a sampling position of a fifth image block (in a first row and a zeroth column) in the second frame of first image may be acquired at first, the sampling position is determined as a reference position, and the fifth image block in the third frame of first image is sampled according to the reference position.

[0087] The splicing portion is configured to splice each first pixel block to generate a second image. As such, each first pixel block is spliced to an original position to keep a correlation between adjacent first pixel blocks. Therefore, image information in the first image may be retained.

[0088] In S13, first quality information of each second image frame sequence is determined.

[0089] Here, the first quality information may include, but not limited to, a score, a grade, a class, etc.

[0090] In some implementation modes, the first quality information of each second image frame sequence is determined by use of a trained neural network model.

[0091] The trained neural network model may be any suitable neural network model capable of obtaining quality information of an image frame sequence. For example, the trained neural network model is a trained first neural network model obtained by training a first neural network model by use of a video sample group. For another example, the trained neural network model is a second neural network model obtained by finely adjusting the trained first neural network model.

[0092] In some implementation modes, the trained neural network model may include, but not limited to, a first network, a second network, etc.

[0093] The first network at least includes a feature extractor, configured to extract an image feature of the second image. The feature extractor may include, but not limited to, a Recurrent Neural Network (RNN), a Convolutional Neural Network (CNN), a

Transformer-based feature extraction network, etc. During implementation, those skilled in the art may use a suitable first network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[0094] In some implementation modes, the first network may include, but not limited to, a feature extractor, a bias portion, etc. The bias portion is configured to represent a relative position in an attention pair by use of a learnable relative bias table. For example, the first network may include Swin-Transformer (Swin-T) and a GRPB. The GRPB uses a learnable real position bias table (T_{real}) and pseudo position bias table (T_{pseudo}) instead of a learnable relative bias table (T), thereby expressing a self-attention matrix generated by the Swin-T more accurately. The Swin-T is configured to determine an image feature of each second image.

[0095] The second network is configured to obtain the first quality information by use of a preset neural network. The preset neural network may include, but not limited to, a linear regression network, a non-linear regression network, a logistic regression network, a classification network, etc. During implementation, those skilled in the art may use a suitable second network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[0096] In some implementation modes, the second network may include, but not limited to, a non-linear layer (R_{NL}) and a pooling layer (Pool). The non-linear layer is configured to regress the image feature of each second image to obtain quality information of each second image. The pooling layer is configured to pool the quality information of each second image to obtain the first quality information of the second image frame sequence.

[0097] In S14, second quality information of the video to be processed is determined based on each first quality information.

[0098] Here, the second quality information may include, but not limited to, a score, a grade, a class, etc. The second quality information may be a mean value of each first quality information, a variance of each first quality information, a mean square error of each first quality information, a mean value of each weighted first quality information, a

mean square error of each weighted first quality information, a variance of each weighted first quality information, etc. During implementation, those skilled in the art may determine the second quality information as practically required, and no limits are made in the embodiments of the disclosure.

[0099] In the embodiment of the disclosure, at least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image. A second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image. First quality information of each second image frame sequence is determined. Second quality information of the video to be processed is determined based on each first quality information. As such, each first image block in the first image in the first image frame sequence is sampled to obtain each first pixel block in the second image in the corresponding second image frame sequence. On one hand, the second image is the same as the first image in resolution, so that quality information of the original video may be retained, and the accuracy of VQA may further be improved. On the other hand, a size of the second image is smaller than that of the first image, so that the calculation cost for quality assessment may be reduced, and the efficiency of VQA may further be improved.

[00100] FIG. 2 is an implementation flowchart of an image processing method according to an embodiment of the disclosure. As shown in FIG. 2, the method includes the following steps S21 to S26.

[00101] In S21, at least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image.

[00102] Here, S21 corresponds to S11, and may be implemented with reference to the implementation modes of S11.

[00103] In S22, for each frame of first image in each first image frame sequence, the first image is divided to generate at least one first image block.

[00104] Here, a partition mode of the first image may include, but not limited to, uniform partition (namely each first image block is the same in size), irregular partition, etc. During implementation, those skilled in the art may determine the partition mode independently as practically required, and no specific limits are made in the disclosure.

[00105] For example, a t-th frame of first image $\mathcal{V}_t$ is divided into $G_f \times G_f$ first image blocks, recorded as $\mathcal{G}_t = \{g_t^{0,0}, \dots, g_t^{i,j}, \dots, g_t^{M-1,N-1}\}$, where $g_t^{i,j}$ represents a first image block in an i-th row and a j-th column, and M and N are the total numbers of rows and lines respectively. In such case, $g_t^{i,j}$ may be represented by the following formula (2-1):

[00106]
$$g_t^{i,j} = \mathcal{V}_t \left[\frac{i \times H}{G_f} : \frac{(i+1) \times H}{G_f}, \frac{j \times W}{G_f} : \frac{(j+1) \times W}{G_f} \right] \quad (2-1).$$

[00107] Here, H and W are used to represent a height and width of the first image respectively.

[00108] In S23, for each frame of first image in each first image frame sequence, a first pixel block is selected from each first image block in the first image based on a preset selection rule, and each first pixel block is spliced to obtain a second image corresponding to the first image.

[00109] Here, the selection rule may include, but not limited to, a selection mode, a selection position, a selection size, etc. In some implementation modes, the selection mode may include, but not limited to, random selection, uniform selection, etc. During implementation, those skilled in the art may determine the selection mode independently as practically required, and no limits are made in the disclosure.

[00110] For example, taking a preset position as a starting point, a first pixel block is selected from each first image block based on a preset size of $A \times B$. A may be the same as or different from B. The preset position may include, but not limited to, a top-left corner, a central point, a specified position, etc. During implementation, those skilled in the art may determine the preset position independently as practically required, and no specific

limits are made in the disclosure.

[00111] For another example, a first pixel block is selected from each first image block based on a preset size of $S_f \times S_f$. For example, random selection is performed on a first image block $g_t^{i,j}$ in an i-th row and a j-th column to obtain a first pixel block $\mathcal{MP}_t^{i,j}$. In some implementation modes, $\mathcal{MP}_t^{i,j}$ may be represented by the following formula (2-2):

$$[00112] \quad \mathcal{MP}_t^{i,j} = S_t^{i,j}(g_t^{i,j}), 0 \leq i, j < G_f \quad (2-2).$$

[00113] Here, $S_t^{i,j}$ represents a random selection operation on a first image block $g_t^{i,j}$ in an i-th row and a j-th column in a t-th frame of first image, i represents the i-th row, j represents the j-th column, and G_f represents a value of the number of rows (columns) to which the first image is divided.

[00114] In some implementation modes, for each first image, each first pixel block is spliced based on position information of each first image block of the first image to generate a second image corresponding to the first image. As such, a correlation between adjacent first pixel blocks may be kept. Therefore, image information in the first image may be retained. In some implementation modes, a first image block $\mathcal{F}_t^{i,j}$ in the second image may be represented by the following formula (2-3):

$$[00115] \quad \begin{aligned} \mathcal{F}_t^{i,j} &= \mathcal{F}_t[i \times S_f: (i + 1) \times S_f, j \times S_f: (j + 1) \times S_f] \\ &= \mathcal{MP}_t^{i,j}, 0 \leq i, j < G_f \end{aligned} \quad (2-3).$$

[00116] Here, t represents the t-th frame of first image, i represents the i-th row, j represents the j-th column. G_f represents a value of the number of rows (columns) to which the first image is divided, and S_f represents a width (height) of the first pixel block.

[00117] In S24, for each first image frame sequence, a second image frame sequence corresponding to the first image frame sequence is determined based on a frame of second image corresponding to each frame of first image in the first image frame sequence.

- [00118] Here, a frame of second image corresponding to each frame of first image in the first image frame sequence is combined as a second image frame sequence.
- [00119] In S25, first quality information of each second image frame sequence is determined.
- [00120] In S26, second quality information of the video to be processed is determined based on each first quality information.
- [00121] Here, S25 to S26 correspond to S13 to S14 respectively, and may be implemented with reference to the implementation modes of S13 to S14.
- [00122] In the embodiment of the disclosure, at least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image. For each frame of first image in each first image frame sequence, the first image is divided to generate at least one first image block. For each frame of first image in each first image frame sequence, a first pixel block is selected from each first image block in the first image based on a preset selection rule, and each first pixel block is spliced to obtain a second image corresponding to the first image. For each first image frame sequence, a second image frame sequence corresponding to the first image frame sequence is determined based on a frame of second image corresponding to each frame of first image in the first image frame sequence. First quality information of each second image frame sequence is determined. Second quality information of the video to be processed is determined based on each first quality information. First, the first image is divided to generate at least one first image block, and then quality assessment may be performed on each part of the first image, so that the accuracy of VQA may be improved. Second, a first pixel block is selected from each first image block. On one hand, the image size may be reduced, so that the calculation cost for quality assessment may be reduced, and the efficiency of VQA may further be improved. On the other hand, local texture information of the original image is retained, so that the accuracy of VQA may further be improved. Third, each first pixel block is spliced, so that overall texture information of the original image may be retained, and the accuracy of VQA may further be improved.

- [00123] In some implementation modes, the method further includes S27 to S28.
- [00124] In S27, position information of the first pixel block corresponding to each first image block in a first frame of first image in the first image frame sequence is determined.
- [00125] Here, the position information may include, but not limited to, starting point information, a size, an end position, etc.
- [00126] In some implementation modes, after each first image block in the first frame of first image is sampled, position information of each first pixel block is acquired and stored.
- [00127] In S28, for the first pixel block corresponding to each first image block in each frame of first image in the first image frame sequence except the first frame of first image, position information of the first pixel block is determined based on the position information of the first pixel block corresponding to the first image block in the first frame of first image.
- [00128] Here, a sampling position of each first image block in the first frame of first image in the first image frame sequence is determined as a reference position, and the other frames of first images (except the first frame) are sampled based on each reference position.
- [00129] For example, each first image is divided into four rows and four columns. In such case, when a four image block (in a zeroth row and a third column) in a third frame of first image is sampled, a sampling position of a four image block (in a zeroth row and a third column) in the first frame of first image may be acquired at first, the sampling position is determined as a reference position, and the four image block in the third frame of first image is sampled according to the reference position.
- [00130] In some implementation modes, a sampling position of the first image block at the same position may be restricted by the following formula (2-4):

$$[00131] \quad S_t^{i,j} = S_{\hat{t}}^{i,j} \quad \forall 0 \leq t, \hat{t} < T, 0 \leq i, j < G_f \quad (2-4).$$

[00132] Here, T represents a total frame number of the first image frame sequence, $\hat{t}$ represents another frame of first image except a t -th frame, (i, j) represents a first image block in an i -th row and a j -th column, and G_f represents a value of the number of rows (columns) to which the first image is divided.

[00133] In the implementation mode of the disclosure, sampling positions of the first image blocks at the same position in every two frames of first images are aligned to make the sampling positions of all frames of first images are the same. As such, variation information between adjacent frames of first images in the first image frame sequence may be retained.

[00134] FIG. 3 is an implementation flowchart of an image processing method according to an embodiment of the disclosure. As shown in FIG. 3, the method includes the following steps S31 to S34.

[00135] In S31, at least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image.

[00136] In S32, a second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[00137] Here, S31 to S32 correspond to S11 to S12 respectively, and may be implemented with reference to the implementation modes of S11 to S12.

[00138] In S33, for each second image frame sequence, first quality information of the second image frame sequence is determined by use of a trained neural network model based on the second image frame sequence.

[00139] Here, the first quality information may include, but not limited to, a score, a grade, a class, etc. The trained neural network model may be any suitable neural network

model capable of obtaining quality information of an image frame sequence. For example, the trained neural network model is a trained first neural network model obtained by training a first neural network model by use of a video sample group. For another example, the trained neural network model is a second neural network model obtained by finely adjusting the trained first neural network model.

[00140] In some implementation modes, the trained neural network model may include, but not limited to, a first network, a second network, etc.

[00141] The first network at least includes a feature extractor, configured to extract an image feature of the second image. The feature extractor may include, but not limited to, an RNN, a CNN, Transformer, Swin-T, etc. During implementation, those skilled in the art may use a suitable first network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00142] The second network is configured to obtain the first quality information by use of a preset neural network. The preset neural network may include, but not limited to, a linear regression network, a non-linear regression network, a logistic regression network, a classification network, etc., such as Intra-Patch Non-Linear Regression (IP-NLR). During implementation, those skilled in the art may use a suitable second network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00143] In S34, second quality information of the video to be processed is determined based on each first quality information.

[00144] Here, S34 corresponds to S14, and may be implemented with reference to the implementation modes of S14.

[00145] In the embodiment of the disclosure, at least one first image frame sequence is determined based on a video to be processed, each first image frame sequence including at least one frame of first image. A second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each second

image being obtained by sampling at least one first image block in a corresponding first image. For each second image frame sequence, first quality information of the second image frame sequence is determined by use of a trained neural network model based on the second image frame sequence. Second quality information of the video to be processed is determined based on each first quality information. First, each first image block in the first image in the first image frame sequence is sampled to obtain each first pixel block in the second image in the corresponding second image frame sequence. On one hand, the second image is the same as the first image in resolution, so that quality information of the original video may be retained, and the accuracy of VQA may further be improved. On the other hand, a size of the second image is smaller than that of the first image, so that the calculation cost for quality assessment may be reduced, and the efficiency of VQA may further be improved. Second, the quality of the video to be processed is determined by the trained neural network model, so that end-to-end quality assessment of the video to be processed is implemented without manually setting and maintaining a large number of assessment rules, the labor cost may be reduced greatly, and meanwhile, the accuracy of quality assessment is improved.

[00146] In some implementation modes, the trained neural network model at least includes a first network and a second network. S33 includes S331 and S332.

[00147] In S331, for each frame of second image in the second image frame sequence, an image feature corresponding to the second image is determined by use of the first network based on the second image.

[00148] Here, the first network may be any suitable neural network capable of obtaining an image feature. In some implementation modes, the first network may include, but not limited to, a feature extractor, etc. The feature extractor is configured to extract an image feature of the second image. The feature extractor may include, but not limited to, an RNN, a CNN, Transformer, Swin-T, etc. During implementation, those skilled in the art may use a suitable first network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00149] In some implementation modes, the first network may include a feature extractor and a bias portion. The bias portion is configured to represent a relative position

in an attention pair by use of a learnable relative bias table. For example, the first network may include Swin-T and a GRPB. The GRPB uses a learnable real position bias table and pseudo position bias table instead of T, thereby expressing a self-attention matrix generated by the Swin-T more accurately. The Swin-T is configured to determine an image feature of each second image.

[00150] In S332, the first quality information is determined by use of the second network based on each image feature.

[00151] Here, the second network may be any suitable neural network capable of obtaining quality information. In some implementation modes, the second network may include, but not limited to, a linear regression network, a non-linear regression network, a logistic regression network, a classification network, etc. During implementation, those skilled in the art may use a suitable second network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00152] In some implementation modes, the second network may include, but not limited to, a non-linear layer (R_{NL}) and a pooling layer (Pool). For example, the second network may be IP-NLR, including R_{NL} and Pool. R_{NL} regresses the image feature of each second image to obtain quality information of each second image. Pool is configured to perform a pooling operation on the quality information of each second image to obtain the first quality information of the second image frame sequence.

[00153] In some implementation modes, the first quality information s_{pred} may be obtained through the following formula (3-1):

[00154]
$$s_{pred} = \text{Pool}(R_{NL}(f)) \quad (3-1).$$

[00155] Here, f represents an image feature.

[00156] In the implementation mode of the disclosure, an image feature of an image frame sequence is extracted by a trained first network, and quality of the video to be processed is determined by a trained second network based on the image feature. End-to-end quality assessment of the video to be processed is implemented without manually setting and maintaining a large number of assessment rules, the labor cost may be

reduced greatly, and meanwhile, the accuracy of quality assessment is improved.

[00157] In some implementation modes, the first network includes a first subnetwork and a second subnetwork. The operation that an image feature corresponding to the second image is determined by use of the first network based on the second image includes S341 to S342.

[00158] In S341, a self-attention matrix is determined by use of the first subnetwork based on the second image.

[00159] Here, the first subnetwork may be any suitable neural network capable of obtaining a self-attention matrix, such as Swin-T. The self-attention matrix represents a correlation degree between each pair of image blocks in the second image.

[00160] In some implementation modes, the first subnetwork may include, but not limited to, a block partition layer, a self-attention layer, a bias portion, etc. The block partition layer is configured to divide the second image into at least one non-overlapping image patch, each image patch including at least one image block. The block partition layer may be any suitable neural network model capable of dividing an image, such as a partition layer (Patch Partition) of the Swin-T. The self-attention layer may be any suitable neural network model capable of dividing an image, such as a self-attention layer of the Swin-T. The bias portion is configured to represent a relative position in an attention pair by use of a learnable relative bias table. During implementation, those skilled in the art may use a suitable first subnetwork according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00161] For example, the first subnetwork may be a block partition layer and first three self-attention layers of the Swin-T, and a GRPB. The Swin-T includes the block partition layer and four self-attention layers. Each self-attention layer is configured to perform down-sampling once. The self-attention layer includes a linear embedding layer and a Swin Transformer block. The Swin Transformer block consists of a shifted window based Multihead Self-attention (MSA) module and two Multi-Layer Perception (MLP) layers. A LayerNorm (LN) layer is used for connection before each MSA module and each MLP layer. A residual is used for connection after each MSA module and each

MLP layer. In each self-attention layer, a correlation degree between each pair of image blocks is calculated in different image patches to obtain a self-attention matrix. That is, a feature vector of each image block is multiplied by a query mapping matrix W^Q , a key mapping matrix W^K and a value mapping matrix W^V to obtain a query matrix Q , a key matrix K , and a value matrix V respectively. The self-attention matrix A may be obtained through the following formula (3-2):

$$[00162] \quad A = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3-2).$$

[00163] Here, $\sqrt{d_k}$ is a scaling factor.

[00164] The GRPB uses a learnable real position bias table and pseudo position bias table instead of a learnable relative bias table, thereby expressing the self-attention matrix generated by the Swin-T more accurately.

[00165] In some implementation modes, the self-attention matrix may be updated based on a relative position bias between each pair of image blocks. The relative position bias may be obtained based on a bias table.

[00166] For example, the bias table may at least include a first bias table and a second bias table. A first distance ($B_{In,(i,j)}$) between each pair of image blocks in the first bias table and a second distance ($B_{Cr,(i,j)}$) in the second bias table are acquired respectively. The self-attention matrix is determined based on each first distance, each second distance, and a correlation degree between the query matrix Q and the key matrix K .

[00167] In some implementation modes, the self-attention matrix A may be obtained through the following formula (3-3):

$$[00168] \quad A = QK^T + G \otimes B_{In} + (1 - G) \otimes B_{Cr} \quad (3-3).$$

[00169] Here, QK^T represents the correlation degree between the query matrix Q and the key matrix K , G represents a correlation degree matrix between each pair of image blocks (namely each element represents position correlation information between two image blocks), B_{In} represents a first distance matrix (namely each element represents a

first distance corresponding to two image blocks), B_{Cr} represents a second distance matrix (namely each element represents a second distance corresponding to two image blocks), and $\otimes$ represents a Kronecker product.

[00170] In S342, the image feature is determined by use of the second subnetwork based on the self-attention matrix.

[00171] Here, the second subnetwork may be any suitable neural network capable of obtaining an image feature. For example, the second subnetwork may be the fourth self-attention layer of the Swin-T. During implementation, those skilled in the art may use a suitable second subnetwork according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00172] In the implementation mode of the disclosure, a self-attention matrix is determined by a trained first subnetwork based on the second image, and the image feature is determined by a trained second subnetwork based on the self-attention matrix. The image feature of the second image may be extracted rapidly and accurately, so that the accuracy of VQA may be improved.

[00173] In some implementation modes, the first subnetwork includes a third subnetwork and a fourth subnetwork, and the self-attention matrix represents a correlation degree between at least one pair of second image blocks. S341 includes S351 to S352.

[00174] In S351, the second image is divided by use of the third subnetwork to generate at least one first image patch, each first image patch including at least one second image block.

[00175] Here, the third subnetwork may be any suitable neural network capable of dividing an image, such as a block partition layer (Patch Partition) of the Swin-T. During implementation, those skilled in the art may use a suitable third subnetwork according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00176] In S352, the self-attention matrix is determined by use of the fourth subnetwork based on each second image block.

[00177] Here, the fourth subnetwork may be any suitable neural network capable of obtaining a self-attention matrix, such as Swin-T. In some implementation modes, the fourth subnetwork may include, but not limited to, a self-attention layer, a bias portion, etc. During implementation, those skilled in the art may use a suitable fourth subnetwork according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00178] In the implementation mode of the disclosure, the self-attention matrix is generated accurately by a fourth subnetwork, so that better image quality information may be obtained from the second image. Therefore, the accuracy of VQA may be improved.

[00179] In some implementation modes, S352 includes S361 to S362.

[00180] In S361, a correlation degree between each pair of second image blocks is determined by use of the fourth subnetwork based on each second image block.

[00181] Here, the fourth subnetwork may be any suitable neural network capable of obtaining a correlation degree between image blocks, such as a GRPB. During implementation, those skilled in the art may use a suitable fourth subnetwork according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00182] The correlation degree represents position correlation information between the two image blocks. In some implementation modes, the position correlation information may include, but not limited to, whether the two image blocks are located in the same image patch, etc. For example, when a pair of image blocks (i,j) are located in the same image patch, $G_{i,j}$ may be 1. When a pair of image blocks (i,j) are located in different image patches, $G_{i,j}$ may be 0.

[00183] In S362, the self-attention matrix is determined based on each correlation degree.

[00184] In some implementation modes, the self-attention matrix is determined based on information such as a correlation degree and relative position bias between each pair of image blocks.

- [00185] In some implementation modes, S362 includes S371 to S373.
- [00186] In S371, a correlation degree matrix is determined based on each correlation degree.
- [00187] Here, the correlation degree matrix includes at least one correlation degree. The correlation degree represents position correlation information between two image blocks. In some implementation modes, the position correlation information may include, but not limited to, whether the two image blocks are located in the same image patch, etc.
- [00188] In S372, a bias matrix is determined based on relative position information of each pair of second image blocks.
- [00189] Here, the bias matrix represents bias information corresponding to each relative position information in a preset bias table. In some implementation modes, the preset bias table includes a first bias table and a second bias table. The first bias table is used to represent real position bias information of each pair of second image blocks. The second bias table is used to represent pseudo position bias information of each pair of second image blocks.
- [00190] In S373, the self-attention matrix is determined based on the correlation degree matrix and the bias matrix.
- [00191] Here, the self-attention matrix may include, but not limited to, a product of the correlation degree matrix and the bias matrix, a Kronecker product of the correlation degree matrix and the bias matrix, a product of a weighted correlation degree matrix and bias matrix, a Kronecker product of the weighted correlation degree matrix and bias matrix, etc. During implementation, those skilled in the art may determine the self-attention matrix as practically required, and no limits are made in the embodiments of the disclosure.
- [00192] In the implementation mode of the disclosure, correlation degrees between different image blocks are expressed accurately by the fourth subnetwork, so that the accuracy of the self-attention matrix may be improved, the image feature may be obtained more accurately, and the accuracy of VQA may further be improved.

[00193] In some implementation modes, the bias matrix includes a first bias matrix and a second bias matrix. S372 includes S381 to S382.

[00194] In S381, the first bias matrix is determined based on the relative position information of each pair of second image blocks and the first bias table.

[00195] Here, the first bias matrix includes at least one first bias information, and the first bias information represents real position bias information corresponding to relative position information between two image blocks in the first bias table. In some implementation modes, the relative position information is taken as an index, and first bias information corresponding to the index in the first bias table is obtained.

[00196] For example, for an (i,j)th pair of second image blocks, relative position information of the (i,j)th pair of second image blocks is determined, and the first bias table is queried based on the relative position information to obtain real position bias information $B_{In,(i,j)}$.

[00197] In some implementation modes, the first bias information may be represented by the following formula (3-4):

$$[00198] \quad B_{In,(i,j)} = T_{real,FRP(i,j)} \quad (3-4).$$

[00199] Here, $FRP(i,j)$ represents the relative position information of the (i,j)th pair of second image blocks, and T_{real} represents the first bias table.

[00200] In S382, the second bias matrix is determined based on the relative position information of each pair of second image blocks and the second bias table.

[00201] Here, the second bias matrix includes at least one second bias information, and the second bias information represents pseudo position bias information corresponding to relative position information between two image blocks in the second bias table. In some implementation modes, the relative position information is taken as an index, and second bias information corresponding to the index in the second bias table is obtained.

[00202] For example, for an (i,j)th pair of second image blocks, relative position

information of the (i, j)th pair of second image blocks is determined, and the second bias table is queried based on the relative position information to obtain pseudo position bias information $B_{Cr,(i,j)}$.

[00203] In some implementation modes, the second bias information may be represented by the following formula (3-5):

[00204] $B_{Cr,(i,j)} = T_{pseudo,FRP(i,j)}$ (3-5).

[00205] Here, $FRP(i, j)$ represents the relative position information of the (i, j)th pair of second image blocks, and T_{pseudo} represents the second bias table.

[00206] In the implementation mode of the disclosure, bias information between different types of image blocks is distinguished by different bias tables, so that the probability of discontinuity between confused blocks and real video artifacts may be reduced, and the accuracy of VQA may further be improved.

[00207] In some implementation modes, S373 includes S391 to S393.

[00208] In S391, a first sub-matrix is determined based on the correlation degree matrix and the first bias matrix.

[00209] Here, the first sub-matrix may include, but not limited to, a product of the correlation degree matrix and the first bias matrix, a Kronecker product of the correlation degree matrix and the first bias matrix, a product of a weighted correlation degree matrix and first bias matrix, a Kronecker product of the weighted correlation degree matrix and first bias matrix, etc. During implementation, those skilled in the art may determine the first sub-matrix as practically required, and no limits are made in the embodiments of the disclosure.

[00210] In some implementation modes, the first sub-matrix A1 may be obtained through the following formula (3-6):

[00211] $A1 = G \otimes B_{In}$ (3-6).

[00212] Here, G represents the correlation degree matrix, B_{In} represents the first bias

matrix, and $\otimes$ represents the Kronecker product.

[00213] In S392, a second sub-matrix is determined based on the correlation degree matrix and the second bias matrix.

[00214] Here, the second sub-matrix may include, but not limited to, a product of the correlation degree matrix and the second bias matrix, a Kronecker product of the correlation degree matrix and the second bias matrix, a product of a weighted correlation degree matrix and second bias matrix, a Kronecker product of the weighted correlation degree matrix and second bias matrix, etc. During implementation, those skilled in the art may determine the second sub-matrix as practically required, and no limits are made in the embodiments of the disclosure.

[00215] In some implementation modes, the second sub-matrix A2 may be obtained through the following formula (3-7):

[00216]
$$A2 = (1 - G) \otimes B_{Cr} \quad (3-7).$$

[00217] Here, G represents the correlation degree matrix, B_{Cr} represents the second bias matrix, and $\otimes$ represents the Kronecker product.

[00218] In S393, the self-attention matrix is determined based on the first sub-matrix and the second sub-matrix.

[00219] Here, the self-attention matrix may include, but not limited to, a sum of the first sub-matrix and the second sub-matrix, a sum of a weighted first sub-matrix and second sub-matrix, etc. During implementation, those skilled in the art may determine the self-attention matrix as practically required, and no limits are made in the embodiments of the disclosure.

[00220] In some implementation modes, the self-attention matrix may be obtained through the following formula (3-8):

[00221]
$$A = A1 + A2 \quad (3-8).$$

[00222] Here, A1 represents the first sub-matrix, and A2 represents the second sub-

matrix.

[00223] In the implementation mode of the disclosure, bias information between different types of image blocks is expressed by different bias matrices, correlation degrees between different image blocks are expressed accurately by a correlation degree matrix, and the self-attention matrix is obtained based on a first bias matrix, a second bias matrix, and the correlation degree matrix. As such, the accuracy of the self-attention matrix may be improved. Therefore, the image feature may be obtained more accurately, and the accuracy of VQA may further be improved.

[00224] In some implementation modes, the second network includes a fifth subnetwork and a sixth subnetwork. S332 includes S3321 and S3322.

[00225] In S3321, for each image feature, non-linear regression is performed on the image feature by use of the fifth subnetwork to obtain quality sub information corresponding to the image feature.

[00226] Here, the fifth subnetwork may be any suitable neural network capable of obtaining quality information. For example, the fifth subnetwork may be R_{NL} . During implementation, those skilled in the art may use a suitable fifth subnetwork according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00227] In some implementation modes, i-th quality sub information s_i may be obtained through the following formula (3-9):

[00228]
$$s_i = R_{NL}(f_i) \quad (3-9).$$

[00229] Here, f_i represents an i-th image feature.

[00230] In S3322, the first quality information is determined by use of the sixth subnetwork based on each quality sub information.

[00231] Here, the first quality information may include, but not limited to, a mean value of each quality sub information, a variance of the mean value of each quality sub information, a mean square error of each quality sub information, a mean value of each weighted quality sub information, a variance of each weighted quality sub information, a

mean square error of each weighted quality sub information, etc. During implementation, those skilled in the art may determine the first quality information as practically required, and no limits are made in the embodiments of the disclosure.

[00232] The sixth subnetwork may be any suitable neural network capable of obtaining quality information. For example, the sixth subnetwork may be a pooling layer. During implementation, those skilled in the art may use a suitable sixth subnetwork according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00233] In the implementation mode of the disclosure, due to the discontinuity between image blocks, different image blocks include different quality information, and quality representations of the image blocks may be confused with each other. In such case, non-linear regression may be performed on the image feature by a trained fifth subnetwork to obtain the quality information of each image block accurately. Therefore, the accuracy of VQA may be improved.

[00234] FIG. 4 is an implementation flowchart of a model training method according to an embodiment of the disclosure. As shown in FIG. 4, the method includes the following steps S41 to S44.

[00235] In S41, at least one first image frame sequence is determined based on at least one first video sample, each first image frame sequence including at least one frame of first image.

[00236] Here, the first video sample may be a video sample in a training sample set, or a video obtained by processing the video sample. Here, the first video sample may be a video of any suitable resolution, such as 1080P (1920×1080), 720P (1280×720), 2K (2048×1080), and 4K (3840×2160).

[00237] In some implementation modes, each first video sample is sampled based on a preset sampling rule to obtain a first image frame sequence corresponding to each first video sample. The preset sampling rule may include, but not limited to, a sampling starting point, a sampling stride, sampling time, a sampling mode, etc. The sampling mode may include, but not limited to, uniform sampling, Gaussian distribution sampling, periodic sampling, fixed-frame-number sampling, etc. During implementation, those

skilled in the art may determine the sampling rule as practically required, and no limits are made in the embodiments of the disclosure.

[00238] For example, if there are four first video samples, each of which includes 100 frames of first images, the second, fourth,, and 64th frames of first images of each first video sample are determined as first image frame sequences respectively based on a preset sampling stride (such as two frames). For another example, the 100 frames of first images of each first video sample are determined as first image frame sequences respectively.

[00239] In S42, a second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[00240] Here, S42 corresponds to S12, and may be implemented with reference to the implementation modes of S12.

[00241] In S43, first quality information of each second image frame sequence is determined by use of a first neural network model to be trained.

[00242] Here, the first neural network model may be any suitable neural network model capable of obtaining quality information. In some implementation modes, the first neural network model may include, but not limited to, a first network, a second network, etc.

[00243] The first network at least includes a feature extractor, configured to extract an image feature of the second image. The feature extractor may include, but not limited to, an RNN, a CNN, Transformer, Swin-T, etc. In some implementation modes, the first network may include, but not limited to, a feature extractor and a bias portion. The bias portion is configured to represent a relative position in an attention pair by use of a learnable relative bias table. For example, the first network may include Swin-T and a GRPB. During implementation, those skilled in the art may use a suitable first network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00244] The second network is configured to obtain the first quality information by use of a preset neural network. The preset neural network may include, but not limited to, a linear regression network, a non-linear regression network, a logistic regression network, a classification network, etc., In some implementation modes, the second network may include, but not limited to, R_{NL} and a pooling layer. During implementation, those skilled in the art may use a suitable second network according to an actual situation, and no limits are made in the embodiments of the disclosure.

[00245] In S44, model parameters of the first neural network model are updated at least once based on each first quality information to obtain a trained first neural network model.

[00246] Here, an updating mode of the model parameters may include, but not limited to, at least one of a gradient descent method, a momentum updating method, a Newton momentum method, etc. During implementation, those skilled in the art may determine the updating mode independently as practically required, and no limits are made in the embodiments of the disclosure.

[00247] In some implementation modes, whether the model parameters of the first neural network model need to be updated may be determined based on a preset condition. When the model parameters of the first neural network model need to be updated, the model parameters of the first neural network model are updated by a suitable parameter updating algorithm. After the model parameters are updated, at least one new first image frame sequence is determined based on next at least one first video sample, at least one new second image frame sequence is determined based on each new first image frame sequence, and whether the model parameters of the first neural network model need to be continuously updated is determined based on first quality information of the at least one new second image frame sequence. When the model parameters of the first neural network model need not to be updated, a finally updated first neural network model is determined as the trained first neural network model. The preset condition may include, but not limited to, at least one of an iteration count, a loss value is not greater than a threshold, the loss value converges, etc.

[00248] For example, a target loss value may be determined based on each first quality

information. When the target loss value does not satisfy the preset condition, the model parameters of the first neural network model are updated. When the target loss value satisfies the preset condition, the model parameters of the first neural network model are stopped to be updated, and a finally updated first neural network model is determined as the trained first model.

[00249] In the embodiment of the disclosure, at least one first image frame sequence is determined based on at least one first video sample, each first image frame sequence including at least one frame of first image. A second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image. First quality information of each second image frame sequence is determined by use of a first neural network model to be trained. Model parameters of the first neural network model are updated at least once based on each first quality information to obtain a trained first neural network model. As such, an overall network structure of the first neural network model is trained with the second image that is the same as the first image in resolution and smaller than the first image in size. On one hand, the performance of the first neural network model is improved, and the trained first neural network model may accurately perform quality assessment on a video to be processed. On the other hand, end-to-end VQA is implemented without manually setting and maintaining a large number of quality assessment rules, the labor cost may be reduced greatly, and meanwhile, the accuracy of VQA is improved.

[00250] In some implementation modes, S41 includes S411 to S412.

[00251] In S411, the at least one first video sample in a first training set is divided into at least one video sample group based on a preset partition rule.

[00252] Here, the partition rule may include, but not limited to, a batch size, a type, a partition mode (uniform and random), etc. For example, each video sample group includes four first video samples. For another example, a video sample group includes eight first video samples of the same type. For another example, a first video sample

group includes five first video samples, and a second video sample group includes six first video samples.

[00253] In some implementation modes, the partition rule may be preset, or set according to a preset mode. The preset mode may include, but not limited to, a customized definition, a user preference, a usage frequency, user operation information, etc. During implementation, those skilled in the art may set the preset mode independently as practically required, and no limits are made in the disclosure. For example, an electronic device provides a configuration option, and a user may customize the partition rule through the configuration option.

[00254] In S412, for each video sample group, each first video sample in the video sample group is re-sampled based on a preset sampling rule to obtain the first image frame sequence corresponding to each first video sample.

[00255] Here, the sampling rule may include, but not limited to, a sampling starting point, a sampling stride, sampling time, a sampling mode, etc. The sampling mode may include, but not limited to, uniform sampling, Gaussian distribution sampling, periodic sampling, fixed-frame-number sampling, etc. During implementation, those skilled in the art may determine the sampling rule as practically required, and no limits are made in the embodiments of the disclosure.

[00256] For example, if a video sample group includes four first video samples, each of which includes 100 frames of first images, the second, fourth,, and 64th frames of first images of each first video sample are determined as corresponding first image frame sequences based on a preset sampling stride (such as two frames).

[00257] For another example, if a video sample group includes four first video samples, each of which includes 100 frames of first images, the 31st to 60th frames of first images of each first video sample are determined as corresponding first image frame sequences based on a preset sampling frame number (such as 30 frames) and a sampling starting point (the 31st frame).

[00258] In some implementation modes, the sampling rule may be set according to a preset mode. The preset mode may include, but not limited to, a default sampling

configuration of the electronic device, a customized definition, a user preference, a usage frequency, user operation information, etc. During implementation, those skilled in the art may set the preset mode independently as practically required, and no limits are made in the disclosure. For example, the electronic device provides a configuration option, and the user may customize the sampling rule through the configuration option.

[00259] In some implementation modes, the method further includes S45 to S46.

[00260] In S45, an initial second neural network model is determined based on the trained first neural network model.

[00261] Here, a network of the trained first neural network model may be adjusted according to a practical application scenario, and an adjusted first neural network model is determined as an initial second model.

[00262] In some implementation modes, the first neural network model at least includes a preprocessing portion and a network portion. The preprocessing portion in the trained first neural network model may be removed, the network portion of the first neural network model is adjusted according to a practical scenario, and an adjusted first network is determined as an initial second neural network model.

[00263] In S46, model parameters of the second neural network model are updated based on at least one second video sample to obtain a trained second neural network model.

[00264] Here, the second video sample may be the same as or different from the first video sample. In some implementation modes, the second video sample may include label information, or include no label information. During implementation, those skilled in the art may determine a suitable second video sample according to a practical application scenario, and no limits are made herein.

[00265] In some implementation modes, model parameters of the second neural network model may be finely adjusted and trained based on at least one second video sample to obtain a trained second neural network model.

[00266] In the implementation mode of the disclosure, an initial second neural network model is determined based on the trained first neural network model, and model parameters of the second neural network model are updated based on at least one second video sample to obtain a trained second neural network model. As such, the model parameters of the trained first neural network model are transferred to the second neural network model. First, this implementation mode is applied to multiple application scenarios, and the generalization ability of the model is further improved. Second, the calculation amount may be reduced during practical applications. Third, the training efficiency of the second neural network model and the VQA accuracy of the trained second neural network model may be improved.

[00267] FIG. 5 is an implementation flowchart of a model training method according to an embodiment of the disclosure. As shown in FIG. 5, the method includes the following steps S51 to S55.

[00268] In S51, at least one first image frame sequence is determined based on at least one first video sample, each first image frame sequence including at least one frame of first image.

[00269] In S52, a second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[00270] In S53, first quality information of each second image frame sequence is determined by use of a first neural network model to be trained.

[00271] Here, S51 to S53 correspond to S41 to S43 respectively, and may be implemented with reference to the implementation modes of S41 to S43.

[00272] In S54, a target loss value is determined based on each first quality information.

[00273] Here, the target loss value may include, but not limited to, at least one of a mean square error loss value, a cross-entropy loss value, a contrastive loss value, etc.

[00274] In some implementation modes, the target loss value may be calculated based on a Pearson Linear Correlation Coefficient (PLCC), a Spearman Rank-order Correlation Coefficient (SRCC), a Kendall Rank-order Correlation Coefficient (KRCC), etc.

[00275] In S55, model parameters of the first neural network model are updated at least once based on the target loss value to obtain a trained first neural network model.

[00276] Here, whether model parameters of the first neural network model need to be updated may be determined based on the target loss value. For example, the target loss value is compared with a threshold. When the target loss value is greater than the threshold, the model parameters of the first neural network model are updated. When the target loss value is not greater than the threshold, the first neural network model is determined as a trained first neural network model. For another example, the target loss value is compared with a previous target loss value. When the target loss value is greater than the previous target loss value, the model parameters of the first neural network model are updated. When the target loss value is substantially equal to the previous target loss value, the first neural network model is determined as a trained first neural network model.

[00277] In the embodiment of the disclosure, at least one first image frame sequence is determined based on at least one first video sample. A second image frame sequence is determined, for each first image frame sequence, based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image. First quality information of each second image frame sequence is determined by use of a first neural network model to be trained. A target loss value is determined based on each first quality information. Model parameters of the first neural network model are updated at least once based on each first quality information to obtain a trained first neural network model. As such, the model parameters of the first neural network model may be updated at least once when the target loss value does not satisfy a preset condition. Since the target loss value is determined based on the first quality information, the consistency of the trained first neural network model in

prediction of the same video sample may be improved, and the trained first neural network model may perform VQA more accurately.

[00278] In some implementation modes, each first video sample includes label information. S54 includes S541 to S542.

[00279] In S541, a first loss value is determined based on the first quality information corresponding to each second image frame sequence and the label information of the first video sample corresponding to the second image frame sequence.

[00280] Here, the first loss value may include, but not limited to, at least one of a mean square error loss value, a cross-entropy loss value, a contrastive loss value, etc. In some implementation modes, the first loss value may be calculated based on a PLCC, an SRCC, a KRCC, etc.

[00281] In some implementation modes, the first loss value may be calculated through the following formula (5-1):

$$[00282] \quad l1 = \frac{\sum_i (s_{pred_i} - \overline{s_{pred}})(s_{gt_i} - \overline{s_{gt}})}{\sqrt{\sum_i (s_{pred_i} - \overline{s_{pred}})^2 \sum_i (s_{gt_i} - \overline{s_{gt}})^2}} \quad (5-1).$$

[00283] Here, s_{pred_i} represents i-th first quality information, s_{gt_i} represents i-th label information, $\overline{s_{pred}}$ represents a mean value of each first quality information, and $\overline{s_{gt}}$ represents a mean value of the label information.

[00284] In S542, the target loss value is determined based on the first loss value.

[00285] Here, the target loss value may include, but not limited to, the first loss value, a weighted first loss value, a difference between a preset value and the first loss value, a difference between the preset value and the weighted first loss value, etc. During implementation, those skilled in the art may determine the target loss value as practically required, and no limits are made in the embodiments of the disclosure.

[00286] In some implementation modes, the target loss value may be calculated through the following formula (5-2):

[00287] $l = \frac{(1-l_1)}{2} \quad (5-2).$

[00288] Here, l_1 represents the first loss value.

[00289] In the implementation mode of the disclosure, the target loss value is determined based on the first quality information and preset label information. As such, the accuracy of the target loss value may be improved, and whether the first neural network model converges may be determined accurately.

[00290] In some implementation modes, S541 includes S5411 to S5413.

[00291] In S5411, a first bias value corresponding to each first quality information is determined based on the first quality information corresponding to each second image frame sequence.

[00292] Here, the first bias value may be a difference between the first quality information and a mean value of each first quality information, a difference between the first quality information and a variance of each first quality information, a difference between the first quality information and a mean square error of each first quality information, a difference between weighted first quality information and a weighted mean value of each first quality information, a difference between the weighted first quality information and a weighted variance of each first quality information, a difference between the weighted first quality information and a weighted mean square error of each first quality information, etc. During implementation, those skilled in the art may determine the first bias value as practically required, and no limits are made in the embodiments of the disclosure.

[00293] In some implementation modes, the i -th first bias value $D1_i$ may be calculated through the following formula (5-3):

[00294] $D1_i = s_{\text{pred}_i} - \overline{s_{\text{pred}}} \quad (5-3).$

[00295] Here, s_{pred_i} represents i -th first quality information, and $\overline{s_{\text{pred}}}$ represents the mean value of each first quality information.

[00296] In S5412, a second bias value corresponding to each label information is determined based on the label information of the first video sample corresponding to each second image frame sequence.

[00297] Here, the second bias value may be a difference between the label information and a mean value of each label information, a difference between the label information and a variance of each label information, a difference between the label information and a mean square error of each label information, a difference between weighted label information and a weighted mean value of each label information, a difference between the weighted label information and a weighted variance of each label information, a difference between the weighted label information and a weighted mean square error of each label information, etc. During implementation, those skilled in the art may determine the second bias value as practically required, and no limits are made in the embodiments of the disclosure.

[00298] In some implementation modes, the i-th second bias value $D2_i$ may be calculated through the following formula (5-4):

[00299]
$$D2_i = s_{gt_i} - \overline{s_{gt}} \quad (5-4).$$

[00300] Here, s_{gt_i} represents i-th label information, and $\overline{s_{gt}}$ represents the mean value of each label information.

[00301] In S5413, the first loss value is determined based on each first bias value and each second bias value.

[00302] Here, the first loss value may include, but not limited to, a sum/difference/product of the first bias value and the second bias value, a sum/difference/product of a weighted first bias value and second bias value, etc. During implementation, those skilled in the art may determine the first loss value as practically required, and no limits are made in the embodiments of the disclosure.

[00303] In some implementation modes, the first loss value may be calculated through the following formula (5-5):

[00304]
$$l1 = \frac{\sum_i D1_i D2_i}{\sqrt{\sum_i D1_i^2 \sum_i D2_i^2}} \quad (5-5).$$

[00305] In the implementation mode of the disclosure, the first loss value is determined based on a first bias value and a second bias value. As such, the accuracy of the first loss value may be improved, and whether the first neural network model converges may be determined accurately.

[00306] FIG. 6A is a composition structure diagram of an image processing system 60 according to an embodiment of the disclosure. As shown in FIG. 6A, the image processing system 60 includes an acquisition portion 61 and an assessment portion 62.

[00307] The acquisition portion 61 is configured to determine at least one first image frame sequence based on a video to be processed, each first image frame sequence including at least one frame of first image.

[00308] The assessment portion 62 is configured to determine second quality information of the video to be processed based on each first image frame sequence.

[00309] In some implementation modes, the assessment portion 62 may be a trained first neural network model and a trained second neural network model. In some implementation modes, the first neural network model/second neural network model at least includes a preprocessing portion and a network portion. In some implementation modes, the first neural network model/second neural network model at least includes a network portion, and in such case, the assessment portion 62 further includes a preprocessing portion.

[00310] The preprocessing portion is configured to determine, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image. The network portion is configured to determine first quality information of each second image frame sequence, and determine the second quality information of the video to be processed based on each first quality information.

[00311] In some implementation modes, the preprocessing portion includes a partition portion, a sampling portion, an alignment portion, and a splicing portion.

[00312] The partition portion is configured to divide the first image in the first image frame sequence based on a preset partition rule to generate at least one first image block corresponding to each first image. In some implementation modes, the partition rule may include, but not limited to, uniform partition (namely each first image block is the same in size), irregular partition, etc. During implementation, those skilled in the art may determine the partition rule independently as practically required, and no specific limits are made in the disclosure. As such, an original resolution of the first image is retained to retain texture related local quality information.

[00313] The sampling portion is configured to sample each first image block based on a preset rule to obtain the first pixel block corresponding to each first image block. Sampling based on the preset rule may include, but not limited to, a sampling mode, a sampling position, a sampling size, etc. In some implementation modes, the sampling mode may include, but not limited to, random sampling, uniform sampling, etc. During implementation, those skilled in the art may determine the sampling mode independently as practically required, and no specific limits are made in the disclosure. Each first image block is globally sampled, so that different first pixel blocks have different quality information, and all the first pixel blocks cover global quality information.

[00314] For example, each first image block is randomly sampled based on a preset size of $C \times C$ to generate each first pixel block. For another example, a pixel block ($D \times D$) is randomly selected from each first image block based on a preset size of $D \times D$, and the pixel block is determined as a first pixel block corresponding to the first image block.

[00315] The alignment portion is configured to align sampling positions of first image blocks at the same position in every two frames of first images. As such, the sampling positions of all frames of first images may be the same.

[00316] In some implementation modes, each first image block in the first image in the first image frame sequence may be sampled based on a reference position. As such, variation information between adjacent frames of first images in the first image frame

sequence may be retained. During implementation, those skilled in the art may determine the reference position independently as practically required, and no limits are made in the disclosure. As such, variation information between adjacent frames is retained to keep the sensitivity of quality to time. Therefore, a distorted part in the video may be found to distinguish a video with jitters from a relatively stable video.

[00317] In some implementation modes, a sampling position of each first image block in a certain frame of first image in the first image frame sequence may be determined as a reference position. For example, a sampling position of each first image block in a first frame of first image is determined as a reference position.

[00318] In some implementation modes, for a present frame of first image, a sampling position of each first image block in a previous frame of first image is determined as a reference position.

[00319] The splicing portion is configured to splice each first pixel block to generate the second image. As such, each first pixel block is spliced to an original position to keep a correlation between adjacent first pixel blocks. Therefore, image information in the first image may be retained for the neural network model to learn global scenario information of the original video frame.

[00320] FIG. 6B is a schematic diagram of processing two frames of first images by a preprocessing portion according to an embodiment of the disclosure. As shown in FIG. 6B, the preprocessing portion includes a partition portion, a sampling portion, an alignment portion, and a splicing portion. The partition portion divides a t-th frame of first image 601 into $G_f \times G_f$ first image blocks 611, and divides a (t+1)th frame of first image 602 into $G_f \times G_f$ first image blocks 612. The sampling portion is configured to randomly select two pixel blocks from each first image block 611 and each second first image block 612 respectively based on a preset size of $S_f \times S_f$ as a first pixel block 621 corresponding to each first image block 611 and a first pixel block 622 corresponding to each first image block 612. The alignment portion is configured to align sampling positions of first image blocks at the same position in two frames of first images, i.e., the first image 601 and the first image 602. That is, a sampling position of first image block 5 611 (in a zeroth row and a fourth column) in the first image 601 is the same as that of

first image block 5 612 (in a zeroth row and a fourth column) in the first image 602. The splicing portion is configured to splice each first pixel block 621 corresponding to the first image 601 to generate a corresponding second image 631, and splice each first pixel block 622 corresponding to the first image 602 to generate a corresponding second image 632. A size of each second image is $(G_f \times S_f) * (G_f \times S_f)$.

[00321] In some implementation modes, the network portion includes a first network and a second network. The first network at least includes a feature extractor and a bias portion. The feature extractor is configured to extract an image feature. The bias portion is configured to represent a relative position in an attention pair by use of a learnable relative bias table. For example, the first network may include Swin-T and a GRPB. The second network is configured to obtain the first quality information by use of a preset neural network. The preset neural network may include, but not limited to, a linear regression network, a non-linear regression network, a logistic regression network, a classification network, etc., During implementation, those skilled in the art may use a suitable second network according to an actual situation, and no limits are made in the embodiments of the disclosure. In some implementation modes, the second network may include R_{NL} and a pooling layer, such as IP-NLR.

[00322] FIG. 6C is a schematic composition diagram of a network portion according to an embodiment of the disclosure. As shown in FIG. 6C, the network portion includes a first network 641 and a second network 642. The first network 641 includes Swin-T and a GRPB. The GRPB is added to first three self-attention layers of the Swin-T. The second network 642 includes IP-NLR R_{NL} and a pooling layer. R_{NL} includes two convolutional layers and an activation function. A score s_{pred} of the second image frame sequence may be obtained by the network portion.

[00323] FIG. 6D a schematic diagram of a GRPB according to an embodiment of the disclosure. As shown in FIG. 6D, each second image is divided into four image patches 651. Each image patch 651 includes multiple second image blocks 652. The GRPB represents a correlation degree between each pair of second image blocks 652 through a gate. When a pair of second image blocks 652 are located in the same image patch, the correlation degree may be represented as B_{In} . When a pair of second image blocks 652

are located in different image patches, the correlation degree may be represented as B_{Cr} .

[00324] Compared with a method in the related art, the method provided in the embodiments of the disclosure at least has the following improvements.

[00325] (1) In the related art, when quality of a video is assessed, the calculation cost or assessment time is reduced by resolution reduction, cropping, etc. A new VQA sampling strategy disclosed in the embodiments of the disclosure includes partition, sampling, alignment, and splicing.

[00326] (2) In the related art, when quality of a video is assessed, a relative position of an attention pair is represented by a learnable relative bias table. In the embodiments of the disclosure, it is proposed to use a learnable real position bias table and pseudo position bias table instead of the relative bias table. The real position bias table and the pseudo position bias table use the same mechanisms as the relative bias table, but learn respectively to represent a relative position of a second image block pair in the same image patch and a relative position of a second image block pair in different image patches respectively.

[00327] (3) In the related art, when quality of a video is assessed, image quality is regressed by linear regression. In the embodiments of the disclosure, non-linear regression is performed on the quality of each second image block.

[00328] (4) In the related art, when quality of a video is assessed, a fixed image feature is extracted in advance, and a model is trained based on the image feature. In the embodiments of the disclosure, the whole neural network model is trained by an end-to-end method.

[00329] (5) In the related art, when quality of a video is assessed, an extracted image feature is fixed, and the image feature is not learned during subsequent training, so the image feature may not be transferred. In the embodiments of the disclosure, the image feature is a feature learned from a large dataset, and may be transferred to another small dataset.

[00330] The method provided in the embodiments of the disclosure at least has the

following beneficial effects. (1) through the new sampling strategy, on one hand, the complexity of an original video is reduced, and the calculation cost for VQA is reduced; and on the other hand, local quality, unbiased global quality, and a context relationship may be maintained, not only is the quality information of the original video retained, but also the accuracy of VQA is improved. (2) The bias information between different types of image blocks is distinguished by different bias tables, so that the probability of discontinuity between confused blocks and real video artifacts may be reduced, and the accuracy of VQA may further be improved. (3) The quality information of each second image block may be represented accurately by a non-linear regression network, so that the probability of confusions between quality representations of the second image blocks is reduced, and the accuracy of VQA may further be improved. (4) End-to-end training is performed with video samples retaining the quality of the original video, so that quality related features may be learned effectively on one hand, and on the other hand, it is unnecessary to manually set and maintain a large number of quality assessment rules, which greatly reduces the labor cost and improves the accuracy of VQA. (5) The learned feature is transferred to the neural network model, so that the generalization ability of the neural network model is improved on one hand, and on the other hand, the training efficiency of the second neural network model and the VQA accuracy of the trained second neural network model may be improved.

[00331] In order to describe the beneficial effects of the embodiments of the disclosure better, experimental data of the method provided in the embodiments of the disclosure and the method in the related art will be described comparatively below.

[00332] (1) Dataset

[00333] A Logarithmic Spherical Vector Quantization (LSVQ) dataset is used to train FAST-VQA. For assessment, four test sets LSVQtest, LSVQ1080p, KoNViD-1k, and LIVE-VQC are selected to test a model trained based on LSVQ.

[00334] (2) Assessment indicator

[00335] A PLCC and a SRCC are used as assessment indicators.

[00336] (3) Experimental result

[00337] (1) Comparison results of the method provided in the embodiments of the disclosure and the existing method on the test dataset

[00338] Table 1 shows test results of each method on the four datasets LSVQtest, LSVQ1080p, KoNViD-1k, and LIVE-VQC. Comparison with each existing and VQA method in a classical group and a deep group and Full-res Swin-T feature (baseline-based full-resolution Swin-T using feature regression but no end-to-end training) shows that, on dataset LSVQ1080p, the assessment indicators in FAST-VQA (normal sampling frequency) are improved by 10% compared with Predictive Vector Quantization (PVQ), the assessment indicators in FAST-VQA-M (relatively low sampling frequency) are also higher than PVQ, and even the assessment indicators in Full-res Swin-T feature is also higher than PVQ.

[00339] Table 1 Test Results of Each Method on Four Datasets LSVQtest, LSVQ1080p, KoNViD-1k, and LIVE-VQC

Type		Intra-dataset				Cross-dataset			
Test set		LSVQtest		LSVQ1080p		KoNViD-1k		LIVE-VQC	
Group	Method	SRC C	PLC C	SRC C	PLC C	SRCC	PLCC	SRC C	PLCC
Classical	BRISQUE	0.569	0.576	0.497	0.531	0.646	0.647	0.524	0.536
	TLVQM	0.772	0.774	0.589	0.616	0.732	0.724	0.670	0.691
	VIDEVAL	0.794	0.783	0.545	0.554	0.751	0.741	0.630	0.640
Deep	VSFA	0.801	0.796	0.675	0.704	0.784	0.794	0.734	0.772
	PVQwo/patch	0.814	0.816	0.686	0.708	0.781	0.781	0.747	0.776
	PVQw/patch	0.827	0.828	0.711	0.739	0.791	0.795	0.770	0.807

Full-res Swin-T feature	0.835	0.833	0.739	0.753	0.825	0.828	0.794	0.809
FAST-VQA-M	0.852	0.854	0.739	0.773	0.841	0.832	0.788	0.810
FAST-VQA	0.876	0.877	0.779	0.814	0.859	0.855	0.823	0.844
Compared with PVQw/patch, improved by	6%	6%	10%	10%	9%	8%	7%	5%

[00340] (2) Comparison results of the VQA efficiency of the method provided in the embodiments of the disclosure and the existing method

[00341] In order to prove the assessment efficiency of FAST-VQA, used Floating-point Operations Per Second (FLOPs) and running time (running averagely for ten times for each sample) of a Central Processing Unit (CPU)/Graphics Processing Unit (GPU) during assessment of videos of different resolutions by FAST-VQA and the existing method are compared, as shown in Table 2. Thus it can be seen that, compared with PVQ, FAST-VQA reduces the FLOPs by 210 times and the running time by 247 times, and meanwhile, achieves higher performance. FAST-VQA-M may achieve relatively high performance only by 1/1,273 of FLOPs of PVQ and 1/258 of FLOPs of Full-res Swin-T feature. FAST-VQA (particularly FAST-VQA-M) may run fast even on the CPU, which reduces a hardware requirement of application of the deep VQA method.

[00342] Table 2 Comparison Results of VQA Efficiency of Various Methods

Method	540P		720P		1080P	
	FLOPs (G)	Running time (second)	FLOPs (G)	Running time (second)	FLOPs (G)	Running time (second)
VSFA	10249 _{36.7x}	2.603/92.761	18184 _{65.2x}	3.571/134.9	40919 _{147x}	11.14/465.6

PVQ	14646 _{52.5x}	3.091/97.85	22029 _{79.0x}	4.143/144.6	58501 _{210x}	13.79/538.4
Full-res Swin-T feature	3032 _{10.9x}	3.226/102.0	5357 _{19.2x}	5.049/166.2	11852 _{42.5x}	8.753/234.9
FAST- VQA-M	279 _{1x}	0.044/9.019	279 _{1x}	0.044/9.530	279 _{1x}	0.045/9.142
FAST- VQA	46 _{0.165x}	0.019/0.729	46 _{0.165x}	0.019/0.613	46 _{0.165x}	0.019/0.714

[00343] (3) Transfer learning of video quality related features

[00344] FAST-VQA pre-trains a VQA dataset in an end-to-end manner to learn quality related features, and then transfers the features to a specific VQA scenario available for a small dataset only. LSVQ is used as a large dataset, and four small datasets representing different scenarios are selected, including LIVE-VQC (real-world mobile photography, 230P to 1080P), KoNViD-1k (various contents collected on line, 540P), CVD2014 (synthetic acquisition distortion, 480P to 720P), and YouTube-UGC (contents generated by the user, including computer graphics contents, 360P to 2,160P). Each dataset is randomly split for ten times, and an average splitting test result is reported. As shown in Table 3, due to video quality related functions, FAST-VQA is much better than the existing method.

[00345] Table 3 Training Results of Various Methods on Multiple Training sets

Test set		LIVE-VQC		KoNViD-1k		CVD2014		YouTube-UGC	
Group	Method	SRC C	PLCC	SRC C	PLC C	SRCC	PLCC	SRCC	PLCC
Classica	TLVQM	0.799	0.803	0.773	0.768	0.83	0.85	0.669	0.659

I	VIDEVAL	0.794	0.783	0.783	0.780	-	-	0.779	0.773
	RAPIQUE	0.755	0.786	0.803	0.817	-	-	0.759	0.768
Deep	VSFA	0.773	0.795	0.773	0.775	0.870	0.868	0.724	0.743
	PVQ	0.827	0.837	0.791	0.786	-	-	-	-
	GST-VQA	-	-	0.814	0.825	0.831	0.844	-	-
	CoINVQ	-	-	0.767	0.764	-	-	0.816	0.802
Full-res Swin-T feature		0.799	0.808	0.841	0.838	0.868	0.870	0.798	0.796
FAST-VQA-M		0.803	0.828	0.873	0.872	0.877	0.892	0.768	0.765
FAST-VQA		0.849	0.865	0.891	0.892	0.891	0.903	0.855	0.852

[00346] (4) Influence of a preprocessing portion of FAST-VQA on VQA

[00347] In the first group in Table 4, a new sampling method in the preprocessing portion is compared with two common sampling methods: bilinear size adjustment and random cropping. Apparently, the new sampling method is better than bilinear size adjustment in LSVQ1080p (+4%) and LIVEVQC (+4%). Compared with bilinear size adjustment in a low-resolution dataset where size adjustment problems are not so serious, the new sampling method still has an improvement of 2%, and this indicates that it is crucial for VQA to retain local textures. The new sampling method is also much better than single random cropping and a combination of multiple croppings, and this indicates that it is also crucial for VQA to retain unified global quality.

[00348] In the second group in Table 4, comparison between the preprocessing portion and variation 1 (including no partition part) shows the effectiveness of uniform grid partition. Comparison between the preprocessing portion and variation 2 (including no splicing part) shows the importance of retaining a context relationship. Compared with variation 3 (including no alignment part), the preprocessing portion presents higher performance particularly for higher performance particularly for a high-resolution video,

and this indicates that it is necessary to retain inter-frame time variations.

[00349] Table 4 Results of Comparison of the Preprocessing Portion With Size Adjustment, Cropping (First Group), and Different Variations (Second Group)

Test set	LSVQtest		LSVQ1080p		KoNViD-1k		LIVE-VQC	
Method	SRC C	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLC C
First group: sampling mode								
Bilinear size adjustment	0.857	0.859	0.752	0.786	0.841	0.84	0.772	0.814
Random cropping	0.807	0.812	0.643	0.677	0.734	0.776	0.740	0.773
Combination of three croppings	0.838	0.835	0.727	0.754	0.841	0.827	0.785	0.809
Combination of six croppings	0.843	0.844	0.734	0.761	0.845	0.834	0.796	0.817
Second group: variations of the preprocessing part								
Variation 1	0.857	0.861	0.754	0.790	0.844	0.845	0.792	0.818
Variation 2	0.858	0.863	0.761	0.799	0.849	0.847	0.796	0.821
Variation 3	0.850	0.853	0.736	0.779	0.823	0.816	0.764	0.802
Preprocessing part	0.876	0.877	0.779	0.814	0.859	0.855	0.823	0.844

[00350] (5) Influence of a network portion of FAST-VQA on VQA

[00351] In the network portion FANet, a GRPB and IP-NLR are used. In Table 5, the

FANet is compared with four variations respectively, i.e., variation 1 (the GRPB is not used in all of four self-attention layers of Swin-T), variation 2 (the GRPB is used in two self-attention layers of Swin-T), variation 3 (linear regression), and variation 4 (non-linear regression, pooling is performed before regression). It can be seen that the two modules achieve non-negligible improvements, particularly in LSVQ1080p or LIVE-VQC. Since the discontinuity between image blocks is more obvious in a high-resolution video, this result shows that a corrected position bias and non-linear regress are favorable for solving problems caused by the discontinuity.

[00352] Table 5 Influences of Using GRPB and IP-NLR in FANet

Test set	LSVQtest		LSVQ1080p		KoNViD-1k		LIVE-VQC	
Method	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Variation 1	0.873	0.872	0.769	0.805	0.854	0.853	0.808	0.832
Variation 2	0.873	0.875	0.772	0.809	0.856	0.851	0.812	0.838
Variation 3	0.872	0.873	0.768	0.803	0.847	0.849	0.810	0.835
Variation 4	0.873	0.874	0.771	0.805	0.851	0.850	0.813	0.834
FANet	0.876	0.877	0.779	0.814	0.859	0.855	0.823	0.844

[00353] (6) Quality assessment of a video by FAST-VQA

[00354] FIG. 6E is a schematic diagram of processing a video frame image by FAST-VQA according to an embodiment of the disclosure. As shown in FIG. 6E, a zeroth frame of first image has clear texture information, and a 12th frame of first image and a 24th frame of first image have fuzzy texture information. Images of the first column represent first images 661. Images of the second column represent quality images 662 obtained by re-projecting the first images 661. Images of the third column represent second images 663. Images of the fourth column represent quality images 664 of multiple second image blocks. It can be seen that FAST-VQA taking the second image 663 as an input is sensitive to local texture quality. In addition, there is a significant

difference between quality of an action related region and quality of a background region, and this shows that FAST-VQA learns global scenario information and context relationship in the video effectively.

[00355] Based on the above-mentioned embodiments, an embodiment of the disclosure provides an image processing apparatus. FIG. 7 is a composition structure diagram of an image processing apparatus according to an embodiment of the disclosure. As shown in FIG. 7, the image processing apparatus 70 includes a first determination module 71, a second determination module 72, a third determination module 73, and a fourth determination module 74.

[00356] The first determination module 71 is configured to determine at least one first image frame sequence based on a video to be processed, each first image frame sequence including at least one frame of first image.

[00357] The second determination module 72 is configured to determine, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[00358] The third determination module 73 is configured to determine first quality information of each second image frame sequence.

[00359] The fourth determination module 74 is configured to determine second quality information of the video to be processed based on each first quality information.

[00360] In some implementation modes, the second determination module 72 is further configured to divide, for each frame of first image in the first image frame sequence, the first image to generate at least one first image block; for each frame of first image, select a first pixel block from each first image block in the first image based on a preset selection rule, and splice each first pixel block to obtain a second image corresponding to the first image; and determine the second image frame sequence based on each second image.

[00361] In some implementation modes, each first pixel block includes position information. The apparatus further includes a first determination submodule and a second determination submodule. The first determination submodule is configured to determine the position information of the first pixel block corresponding to each first image block in a first frame of first image in the first image frame sequence. The second determination submodule is configured to determine, for the first pixel block corresponding to each first image block in each frame of first image in the first image frame sequence except the first frame of first image, the position information of the first pixel block based on the position information of the first pixel block corresponding to the first image block in the first frame of first image.

[00362] In some implementation modes, the third determination module 73 is further configured to determine, for each second image frame sequence, the first quality information of the second image frame sequence by use of a trained neural network model based on the second image frame sequence.

[00363] In some implementation modes, the trained neural network model at least includes a first network and a second network. The third determination module 73 is further configured to determine, for each frame of second image in the second image frame sequence, an image feature corresponding to the second image by use of the first network based on the second image; and determine the first quality information by use of the second network based on each image feature.

[00364] In some implementation modes, the first network includes a first subnetwork and a second subnetwork. The third determination module 73 is further configured to determine a self-attention matrix by use of the first subnetwork based on the second image; and determine the image feature by use of the second subnetwork based on the self-attention matrix.

[00365] In some implementation modes, the first subnetwork includes a third subnetwork and a fourth subnetwork, and the self-attention matrix represents a correlation degree between at least one pair of second image blocks. The third determination module 73 is further configured to divide the second image by use of the third subnetwork to generate at least one first image patch, each first image patch

including at least one second image block; and determine the self-attention matrix by use of the fourth subnetwork based on each second image block.

[00366] In some implementation modes, the third determination module 73 is further configured to determine a correlation degree between each pair of second image blocks by use of the fourth subnetwork based on each second image block; and determine the self-attention matrix based on each correlation degree.

[00367] In some implementation modes, the third determination module 73 is further configured to determine a correlation degree matrix based on each correlation degree; determine a bias matrix based on relative position information of each pair of second image blocks; and determine the self-attention matrix based on the correlation degree matrix and the bias matrix.

[00368] In some implementation modes, the bias matrix includes a first bias matrix and a second bias matrix. The third determination module 73 is further configured to determine the first bias matrix based on the relative position information of each pair of second image blocks and the first bias table; determine the second bias matrix based on the relative position information of each pair of second image blocks and the second bias table; determine a first sub-matrix based on the correlation degree matrix and the first bias matrix; determine a second sub-matrix based on the correlation degree matrix and the second bias matrix; and determine the self-attention matrix based on the first sub-matrix and the second sub-matrix.

[00369] In some implementation modes, the second network includes a fifth subnetwork and a sixth subnetwork. The third determination module 73 is further configured to perform, for each image feature, non-linear regression on the image feature by use of the fifth subnetwork to obtain quality sub information corresponding to the image feature; and determine the first quality information by use of the sixth subnetwork based on each quality sub information.

[00370] Based on the above-mentioned embodiments, an embodiment of the disclosure provides a model training apparatus. FIG. 8 is a composition structure diagram of a model training apparatus according to an embodiment of the disclosure. As shown in FIG.

8, the model training apparatus 80 includes a fifth determination module 81, a sixth determination module 82, a seventh determination module 83, and a first updating module 84.

[00371] The fifth determination module 81 is configured to determine at least one first image frame sequence based on at least one first video sample, each first image frame sequence including at least one frame of first image.

[00372] The sixth determination module 82 is configured to determine, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, each frame of second image in the second image frame sequence corresponding to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image being obtained by sampling at least one first image block in a corresponding first image.

[00373] The seventh determination module 83 is configured to determine first quality information of each second image frame sequence by use of a first neural network model to be trained.

[00374] The first updating module 84 is configured to update model parameters of the first neural network model at least once based on each first quality information to obtain a trained first neural network model.

[00375] In some implementation modes, the first updating module 84 is further configured to determine a target loss value based on each first quality information; and update the model parameters of the first neural network model at least once based on the target loss value to obtain the trained first neural network model.

[00376] In some implementation modes, each first video sample includes label information. The first updating module 84 is further configured to determine a first loss value based on the first quality information corresponding to each second image frame sequence and the label information of the first video sample corresponding to the second image frame sequence; and determine the target loss value based on the first loss value.

[00377] In some implementation modes, the first updating module 84 is further

configured to determine a first bias value corresponding to each first quality information based on the first quality information corresponding to each second image frame sequence; determine a second bias value corresponding to each label information based on the label information of the first video sample corresponding to each second image frame sequence; and determine the first loss value based on each first bias value and each second bias value.

[00378] In some implementation modes, the fifth determination module 81 is further configured to divide the at least one first video sample in a first training set into at least one video sample group based on a preset partition rule; and re-sample, for each video sample group, each first video sample in the video sample group based on a preset sampling rule to obtain the first image frame sequence corresponding to each first video sample.

[00379] In some implementation modes, the apparatus further includes an eighth determination module and a second updating module. The eighth determination module is configured to determine an initial second neural network model based on the trained first neural network model. The second updating module is configured to update model parameters of the second neural network model based on at least one second video sample to obtain a trained second neural network model.

[00380] The descriptions about the above apparatus embodiment are similar to those about the method embodiment, and beneficial effects similar to those of the method embodiment are achieved. Technical details undisclosed in the apparatus embodiment of the disclosure may be understood with reference to the descriptions about the method embodiment of the disclosure.

[00381] It is to be noted that, in the embodiments of the disclosure, the method may be stored in a computer-readable storage medium when being implemented in form of a software function module and sold or used as an independent product. Based on such an understanding, the technical solutions of the embodiments of the disclosure substantially or parts making contributions to the related art may be embodied in form of a software product. The software product is stored in a storage medium, including a plurality of instructions for enabling an electronic device (which may be a personal computer, a

server, a network device, or the like) to execute all or part of the method in each embodiment of the disclosure. The storage medium includes various media capable of storing program codes, such as a U disk, a mobile hard disk, a Read Only Memory (ROM), a magnetic disk, or an optical disk. Therefore, the embodiments of the disclosure are not limited to any specific combination of hardware and software.

[00382] An embodiment of the disclosure provides an electronic device, which includes a memory and a processor. The memory stores a computer program capable of running in the processor. The processor executes the computer program to implement the method.

[00383] An embodiment of the disclosure provides a computer-readable storage medium having stored thereon a computer program that, when executed by a processor, cause the method to be implemented. The computer-readable storage medium may be transitory or non-transitory.

[00384] An embodiment of the disclosure provides a computer program product, which includes a non-transitory computer-readable storage medium storing a computer program. The computer program is read and executed by a computer to implement part or all of the steps in the method. The computer program product may specifically be implemented by hardware, software, or a combination thereof. In an optional embodiment, the computer program product is specifically embodied as a computer storage medium. In another optional embodiment, the computer program product is specifically embodied as a software product, such as a Software Development Kit (SDK).

[00385] It is to be noted that FIG. 9 is a schematic diagram of a hardware entity of an electronic device according to an embodiment of the disclosure. As shown in FIG. 9, the hardware entity of the electronic device 900 includes a processor 901, a communication interface 902, and a memory 903.

[00386] The processor 901 usually controls overall operations of the electronic device 900.

[00387] The communication interface 902 may enable the electronic device to communicate with another terminal or server through a network.

[00388] The memory 903 is configured to store an instruction and application executable for the processor 901, may further cache data (such as image data, audio data, voice communication data, and video communication data) to be processed or having been processed by the processor 901 and each module in the electronic device 900, and may be implemented by a flash or a Random Access Memory (RAM). The processor 901, the communication interface 902, and the memory 903 may perform data transmission with one another through a bus 904.

[00389] It is to be pointed out here that the descriptions about the above storage medium and device embodiments are similar to those about the method embodiment, and beneficial effects similar to those of the method embodiment are achieved. Technical details undisclosed in the storage medium and device embodiments of the disclosure are understood with reference to those about the method embodiment of the disclosure.

[00390] It is to be understood that "one embodiment" or "an embodiment" mentioned throughout the specification means that specific features, structures, or characteristics related to the embodiment are included in at least one embodiment of the disclosure. Therefore, "in one embodiment" or "in an embodiment" mentioned throughout the specification does not necessarily refer to the same embodiment. In addition, these specific features, structures, or characteristics may be combined in one or more embodiments freely as appropriate. It is to be understood that, in each embodiment of the disclosure, the magnitude of the sequence number of each process does not mean an execution sequence, and the execution sequence of each process is determined by its function and an internal logic, and does not form any limitation on the implementation of the embodiments of the disclosure. The sequence numbers of the embodiments of the disclosure are only for description and do not represent superiority-inferiority of the embodiments.

[00391] It is to be noted that terms "include" and "contain" or any other variation thereof is intended to cover nonexclusive inclusions herein, so that a process, method, object, or device including a series of elements not only includes those elements but also includes other elements which are not clearly listed, or further includes elements intrinsic to the process, the method, the object, or the device. With no more restrictions, an

element defined by statement "including a/an....." does not exclude the existence of the same other elements in a process, method, object, or device including the element.

[00392] In some embodiments provided in the disclosure, it is to be understood that the disclosed device and method may be implemented in another manner. The device embodiment described above is only schematic. For example, the partition of the units is only logic function partition, and other partition manners may be used during practical implementation. For example, multiple units or components may be combined or integrated into another system, or some characteristics may be neglected or not executed. In addition, coupling or direct coupling or communication connection between each displayed or discussed component may be indirect coupling or communication connection, implemented through some interfaces, of the device or the units, and may be electrical or mechanical or use other forms.

[00393] The units described as separate parts may or may not be physically separated. Parts displayed as units may or may not be physical units. That is, they are located in the same place or distributed to multiple network units. Part or all of the units may be selected as practically required to achieve the objectives of the solutions of the embodiments.

[00394] In addition, each function unit in the embodiments of the disclosure may be integrated into a processing unit. Alternatively, each unit may serve as an independent unit. Alternatively, two or more than two units may be integrated into a unit. The integrated unit may be implemented in a hardware form, or in form of hardware plus a software function unit.

[00395] It can be understood by those of ordinary skill in the art that all or part of the steps of the method embodiment may be implemented by related hardware instructed through a program. The program may be stored in a computer-readable storage medium. When the program is executed, the steps of the method embodiment are executed. The storage medium includes various media capable of storing program codes, such as a mobile storage device, a ROM, a magnetic disk, or an optical disk.

[00396] Alternatively, the integrated unit of the disclosure may be stored in a

computer-readable storage medium when being implemented in form of a software function module and sold or used as an independent product. Based on such an understanding, the technical solutions of the disclosure substantially or parts making contributions to the related art may be embodied in form of a software product. The computer software product is stored in a storage medium, including a plurality of instructions for enabling an electronic device (which may be a personal computer, a server, a network device, or the like) to execute all or part of the method in each embodiment of the disclosure. The storage medium includes various media capable of storing program codes, such as a mobile hard disk, a ROM, a magnetic disk, or an optical disk.

[00397] The above is only the implementation mode of the disclosure and not intended to limit the scope of protection of the disclosure. Any variations or replacements apparent to those skilled in the art within the technical scope disclosed in the disclosure shall fall within the scope of protection of the disclosure.

CLAIMS

1. An image processing method, comprising:

determining at least one first image frame sequence based on a video to be processed, each first image frame sequence comprising at least one frame of first image;

determining, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, wherein each frame of second image in the second image frame sequence corresponds to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image is obtained by sampling at least one first image block in a corresponding first image;

determining first quality information of each second image frame sequence; and

determining second quality information of the video to be processed based on each first quality information.

2. The method of claim 1, wherein determining the second image frame sequence based on the first image frame sequence comprises:

dividing, for each frame of first image in the first image frame sequence, the first image to generate at least one first image block;

for each frame of first image, selecting a first pixel block from each first image block in the first image based on a preset selection rule, and splicing each first pixel block to obtain a second image corresponding to the first image; and

determining the second image frame sequence based on each second image.

3. The method of claim 2, wherein each first pixel block comprises position information, and

the method further comprises:

determining the position information of the first pixel block corresponding to each first image block in a first frame of first image in the first image frame sequence; and

determining, for the first pixel block corresponding to each first image block in each frame of first image in the first image frame sequence except the first frame of first image, the position information of the first pixel block based on the position information of the first pixel block corresponding to the first image block in the first frame of first image.

4. The method of any one of claims 1-3, wherein determining the first quality information of each second image frame sequence comprises:

determining, for each second image frame sequence, the first quality information of the second image frame sequence by use of a trained neural network model based on the second image frame sequence.

5. The method of claim 4, wherein the trained neural network model at least comprises a first network and a second network;

determining the first quality information of the second image frame sequence by use of the trained neural network model based on the second image frame sequence comprises:

determining, for each frame of second image in the second image frame sequence, an image feature corresponding to the second image by use of the first network based on the second image; and

determining the first quality information by use of the second network based on each image feature.

6. The method of claim 5, wherein the first network comprises a first subnetwork and a

second subnetwork;

determining the image feature corresponding to the second image by use of the first network based on the second image comprises:

determining a self-attention matrix by use of the first subnetwork based on the second image; and

determining the image feature by use of the second subnetwork based on the self-attention matrix.

7. The method of claim 6, wherein the first subnetwork comprises a third subnetwork and a fourth subnetwork, and the self-attention matrix represents a correlation degree between at least one pair of second image blocks;

determining the self-attention matrix by use of the first subnetwork based on the second image comprises:

dividing the second image by use of the third subnetwork to generate at least one first image patch, each first image patch comprising at least one second image block; and

determining the self-attention matrix by use of the fourth subnetwork based on each second image block.

8. The method of claim 7, wherein determining the self-attention matrix by use of the fourth subnetwork based on each second image block comprises:

determining a correlation degree between each pair of second image blocks by use of the fourth subnetwork based on each second image block; and

determining the self-attention matrix based on each correlation degree.

9. The method of claim 8, wherein determining the self-attention matrix based on each correlation degree comprises:

determining a correlation degree matrix based on each correlation degree;

determining a bias matrix based on relative position information of each pair of second image blocks; and

determining the self-attention matrix based on the correlation degree matrix and the bias matrix.

10. The method of claim 9, wherein the bias matrix comprises a first bias matrix and a second bias matrix;

determining the bias matrix based on relative position information of each pair of second image blocks comprises:

determining the first bias matrix based on the relative position information of each pair of second image blocks and a first bias table, and determining the second bias matrix based on the relative position information of each pair of second image blocks and a second bias table; and

determining the self-attention matrix based on the correlation degree matrix and the bias matrix comprises:

determining a first sub-matrix based on the correlation degree matrix and the first bias matrix, determining a second sub-matrix based on the correlation degree matrix and the second bias matrix, and determining the self-attention matrix based on the first sub-matrix and the second sub-matrix.

11. The method of any one of claims 5-10, wherein the second network comprises a fifth subnetwork and a sixth subnetwork;

determining the first quality information by use of the second network based on each image feature comprises:

performing, for each image feature, non-linear regression on the image feature by use of the fifth subnetwork to obtain quality sub information corresponding to the image feature; and

determining the first quality information by use of the sixth subnetwork based on each quality sub information.

12. A model training method, comprising:

determining at least one first image frame sequence based on at least one first video sample, each first image frame sequence comprising at least one frame of first image;

determining, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, wherein each frame of second image in the second image frame sequence corresponds to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image is obtained by sampling at least one first image block in a corresponding first image;

determining first quality information of each second image frame sequence by use of a first neural network model to be trained; and

updating model parameters of the first neural network model at least once based on each first quality information to obtain a trained first neural network model.

13. The method of claim 12, wherein updating the model parameters of the first neural network model at least once based on each first quality information to obtain a trained first neural network model comprises:

determining a target loss value based on each first quality information; and

updating the model parameters of the first neural network model at least once based on the target loss value to obtain the trained first neural network model.

14. The method of claim 13, wherein each first video sample comprises label information;

determining the target loss value based on each first quality information comprises:

determining a first loss value based on the first quality information corresponding to each second image frame sequence and the label information of the first video sample corresponding to the second image frame sequence; and

determining the target loss value based on the first loss value.

15. The method of claim 14, wherein determining the first loss value based on the first quality information corresponding to each second image frame sequence and the label information of the first video sample corresponding to the second image frame sequence comprises:

determining a first bias value corresponding to each first quality information based on the first quality information corresponding to each second image frame sequence;

determining a second bias value corresponding to each label information based on the label information of the first video sample corresponding to each second image frame sequence; and

determining the first loss value based on each first bias value and each second bias value.

16. The method of any one of claims 12-15, wherein determining at least one first image frame sequence based on at least one first video sample comprises:

dividing the at least one first video sample in a first training set into at least one video

sample group based on a preset partition rule; and

re-sampling, for each video sample group, each first video sample in the video sample group based on a preset sampling rule to obtain the first image frame sequence corresponding to each first video sample.

17. The method of any one of claims 12-16, further comprising:

determining an initial second neural network model based on the trained first neural network model; and

updating model parameters of the second neural network model based on at least one second video sample to obtain a trained second neural network model.

18. An image processing apparatus, comprising:

a first determination module, configured to determine at least one first image frame sequence based on a video to be processed, each first image frame sequence comprising at least one frame of first image;

a second determination module, configured to determine, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, wherein each frame of second image in the second image frame sequence corresponds to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image is obtained by sampling at least one first image block in a corresponding first image;

a third determination module, configured to determine first quality information of each second image frame sequence; and

a fourth determination module, configured to determine second quality information of the video to be processed based on each first quality information.

19. A model training apparatus, comprising:

a fifth determination module, configured to determine at least one first image frame sequence based on at least one first video sample, each first image frame sequence comprising at least one frame of first image;

a sixth determination module, configured to determine, for each first image frame sequence, a second image frame sequence based on the first image frame sequence, wherein each frame of second image in the second image frame sequence corresponds to a frame of first image in the first image frame sequence, and at least one first pixel block in each frame of second image is obtained by sampling at least one first image block in a corresponding first image;

a seventh determination module, configured to determine first quality information of each second image frame sequence by use of a first neural network model to be trained; and

a first updating module, configured to update model parameters of the first neural network model at least once based on each first quality information to obtain a trained first neural network model.

20. An electronic device, comprising a processor and a memory storing a computer program capable of running in the processor, wherein the method of any one of claims 1-17 is performed when the processor executes the computer program.

21. A computer-readable storage medium having stored thereon a computer program that, when executed by a processor, causes the method of any one of claims 1-17 to be performed.