

THREE-DIMENSIONAL AUTO-ENCODER AND TRAINING METHOD THEREFOR, ELECTRONIC DEVICE, AND STORAGE MEDIUM

TECHNICAL FIELD

[1] The present invention relates to the technical field of artificial intelligence, and particularly relates to a three-dimensional auto-encoder and a training method therefor, a training method for a three-dimensional vision module, an electronic device, and a storage medium.

BACKGROUND

[2] All existing action recognition technologies based on a deep learning neural network are generally used for training based on a supervised learning approach, and the supervised learning means that all data have labels or truth values and calculated loss functions of network output results and labels are directly trained. All training sets in this way are relatively small, and covered scenarios are relatively limited; a trained model is only applicable to scenarios associated with the training datasets and cannot be conveniently transplanted to other scenarios, a labeled dataset is required for training, correct labeling itself requires a lot of human and arithmetic costs, and extracted features also rely on labels, rather than features of the data themselves. Self-supervised learning refers to supervised training by generating labels by itself, the ways of generating labels by itself usually include rotating, cropping, coloring, contrast learning and the like, and then fine-tuning is performed according to a downstream task.

[3] Action recognition such as fall recognition and object handling recognition usually involves an interaction between a human body and objects in surrounding scenarios, a type of action can be more accurately judged through a perceived 3D spatial position structure, independently of a spatial position of a camera; for example, whether a person is lying down in a chair or sitting on the floor can be more accurately deduced through spatial location information. 3D GANs (Generative adversarial networks) capable of perceiving 3D structures are different from ordinary 2D GANs and are based on the following two basic components:

[4] 1. inductive bias capable of perceiving 3D structures, introduced into a generative network; and

[5] 2. a neural network rendering engine, used for providing image results with consistent perspectives;

[6] the inductive bias capable of perceiving 3D structures can generally be modeled with an explicit voxel grid or can be modeled with an implicit neural network.

[7] Vision Transformers (ViT) is a vision module based on transformer modules, which can perform machine vision related tasks. In a general-purpose convolutional neural network module (e.g., ResNet) that performs machine vision tasks, saturation will be reached after a number of training data in the dataset is increased; for example, saturation will be reached after a volume of the training set reaches 100 million pictures, and an accuracy of recognition will no longer be increased even though training data are further added; however, ViT is suitable for use as a basic model for pre-training of large data volumes as its accuracy of recognition will also continue to rise at a training data volume of more than 100 million. Video Vision Transformers (ViViT) is a vision module based on transformer modules, which can perform machine vision related tasks. Similar to ViT which uses image patches for recognition and token division, ViViT uses video patches to perform recognition tasks and token division. An MAE (Masked Auto Encoders) mode is used to train Vision Transformers, the Vision Transformers can be trained by means of Self-Supervised Learning, and the trained Vision Transformers can achieve a high accuracy. Similarly, 3DViViT is a 3D vision module based on transformer modules, i.e., 3DViViT can perceive the three-dimensional features of an image.

SUMMARY

[8] The technical problem to be solved by the present invention is to provide a three-dimensional auto-encoder and a training method therefor, a training method for a three-dimensional vision module, an electronic device, and a storage medium in order to overcome the above defects in the prior art.

[9] The present invention solves the above technical problems through the following technical solution:

[10] A three-dimensional auto-encoder is provided, including an encoder and a decoder; the encoder is used for extracting spatial feature parameters of an input picture to be processed and outputting the spatial feature parameters to the decoder; the decoder is used for outputting, according to the spatial feature parameters, a target picture including depth

information of the picture to be processed; and the picture to be processed and the target picture are used for determining a loss function of the three-dimensional auto-encoder.

[11] Preferably, the decoder includes a three-dimensional feature extraction module and an upsampling module; the three-dimensional feature extraction module is used for obtaining the depth information of the picture to be processed according to the spatial feature parameters; and the upsampling module is used for generating the target picture according to the depth information and the spatial feature parameters processed by the three-dimensional feature extraction module.

[12] Preferably, the three-dimensional feature extraction module includes a mapping network, a two-dimensional generative adversarial network generator, and a neural network renderer; and the spatial feature parameters include a camera parameter and a hidden scalar;

[13] the mapping network is used for processing the hidden scalar to correspond to specifications of the picture to be processed, and outputting the processed hidden scalar to the two-dimensional generative adversarial network generator and the upsampling module;

[14] the two-dimensional generative adversarial network generator is used for generating a feature-deformed feature map according to the processed spatial feature parameters, and inputting the feature-deformed feature map into the neural network renderer; and

[15] the neural network renderer is used for processing the feature-deformed feature map and the camera parameter to output the depth information of the picture to be processed.

[16] Preferably, the neural network renderer includes a three-plane decoder and a voxel rendering module;

[17] the three-plane decoder is used for processing the feature-deformed feature map to output voxel information to the voxel rendering module; and

[18] the voxel rendering module is used for outputting the depth information of the picture to be processed according to the voxel information and the camera parameter.

[19] Preferably, the three-plane decoder includes three fully connected layers connected in sequence; and the mapping network includes three fully connected layers connected in sequence.

[20] Preferably, the encoder includes a downsampling convolutional neural network; and the downsampling convolutional neural network is used for downsampling the picture to be processed.

[21] Preferably, the encoder further includes a fully connected layer; the downsampling convolutional neural network includes a convolutional layer and a pooling layer; the convolutional layer, the pooling layer and the fully connected layer are connected in sequence; and the convolutional layer is implemented based on an attention mechanism and a ReLU activation function.

[22] Preferably, the three-dimensional feature extraction module is obtained based on a feature extraction module of a three-dimensional generative adversarial network, and the feature extraction module is used for extracting three-dimensional scenario information of the picture to be processed.

[23] A training method for a three-dimensional auto-encoder is further provided, wherein the training method is used for training the three-dimensional auto-encoder; and the training method includes:

[24] locking parameters of a decoder of the three-dimensional auto-encoder, and training parameters of an encoder of the three-dimensional auto-encoder using a first training data set until convergence; and the first training data set is used for training and obtaining the three-dimensional generative adversarial network; and

[25] unlocking the parameters of the encoder on, and continuing to train the parameters of the encoder and the decoder until convergence.

[26] A training method for a three-dimensional vision module is further provided, including the following steps:

[27] processing video frames using the above three-dimensional auto-encoder to obtain three-dimensional spatial features corresponding to the video frames;

[28] dividing the video frames associated with the three-dimensional spatial feature into a plurality of video patches;

[29] assigning position vectors and time vectors to the plurality of video patches, respectively;

[30] screening and filtering out a first preset proportion of video frames randomly according to the time vectors;

[31] selecting a second preset proportion of video patches randomly from the remaining video frames as training data, inputting the training data into an encoder of a three-dimensional vision module, and inputting the outputted video patches plus tokens corresponding to the first preset proportion of video frames filtered out through the screening into a decoder of the

three-dimensional vision module; the video patches are arranged in an order according to the time vectors and the position vectors; and

[32]training the three-dimensional vision module based on the output data of the decoder of the three-dimensional vision module and the video frames until convergence.

[33]An electronic device is further provided, including a memory, a processor and a computer program stored in the memory and executable on the processor; the processor is configured to implement the above training method for a three-dimensional vision module when executing the computer program.

[34]A computer-readable storage medium is further provided; the storage medium includes a computer program stored thereon, which when executed by the processor, is configured to implement the above training method for a three-dimensional vision module.

[35]The present invention has the following positive progressive effects: by providing a three-dimensional auto-encoder and a training method therefor, a training method for a three-dimensional vision module, an electronic device, and a storage medium, perception of three-dimensional scenario features of an image can be implemented by means of self-supervision; moreover, more scenarios can be covered by using massive training sets, training is performed without a labeled data set, training can be performed by using a large amount of unlabeled original data, which saves the labor and cost, and a trained model can be conveniently transplanted to other scenarios. Inductive bias of a three-dimensional spatial relationship is added to a neural network design, so that interactive information with an environment included in an action can be perceived and the accuracy of recognition can be improved.

BRIEF DESCRIPTION OF THE DRAWINGS

[36]FIG. 1 is a structure diagram of a three-dimensional auto-encoder according to Embodiment 1 of the present invention.

[37]FIG. 2 is a structure diagram of a decoder of the three-dimensional auto-encoder according to Embodiment 1 of the present invention.

[38]FIG. 3 is a structure diagram of a encoder of the three-dimensional auto-encoder according to Embodiment 1 of the present invention.

[39]FIG. 4 is a flow chart of a training method for a three-dimensional auto-encoder according to Embodiment 2 of the present invention.

[40]FIG. 5 is a flow chart of a training method for a three-dimensional vision module according to Embodiment 3 of the present invention.

[41]FIG. 6 is a processing flow chart of perception of three-dimensional scenario features of video frames in Embodiment 3 of the present invention.

[42]FIG. 7 is a structure diagram of an encoder of a three-dimensional vision module in Embodiment 3 of the present invention.

[43]FIG. 8 is a structure diagram of a decoder of a three-dimensional vision module in Embodiment 3 of the present invention.

[44]FIG. 9 is a structure block diagram of an electronic device according to Embodiment 4 of the present invention.

DETAILED DESCRIPTIONS OF THE EMBODIMENTS

[45]The present invention will be further described below by ways of embodiments, but is not therefore limited to the scope of the described embodiments.

[46]Embodiment 1

[47]By reference to FIGs. 1-3, a three-dimensional auto-encoder is specifically provided, including an encoder and a decoder; the encoder is used for extracting spatial feature parameters of an input picture to be processed and outputting the spatial feature parameters to the decoder; the decoder is used for outputting, according to the spatial feature parameters, a target picture including depth information of the picture to be processed; and the picture to be processed and the target picture are used for determining a loss function of the three-dimensional auto-encoder.

[48]In the embodiment, a three-dimensional auto-encoder constructed based on an encoder-decoder mode is provided, as shown in FIG. 1, which includes an encoder and a decoder. As a preferred implementation mode, the three-dimensional feature extraction module is obtained based on a feature extraction module of a three-dimensional generative adversarial network, and the feature extraction module is used for extracting three-dimensional scenario information of the picture to be processed. The three-dimensional generative adversarial network is a GANs capable of perceiving 3D structures,

implementation of which is based on introduction of inductive bias capable of perceiving 3D structures in a generative network, and a neural network rendering engine capable of providing image results with consistent perspectives, and feature maps corresponding to 3D structures from random features can be outputted. The inductive bias capable of perceiving 3D structures can generally be modeled with an explicit voxel grid or can be modeled with an implicit neural network.

[49]As a preferred implementation mode, by reference to FIG. 2, the decoder includes a three-dimensional feature extraction module and an upsampling module; the three-dimensional feature extraction module is used for obtaining the depth information of the picture to be processed according to the spatial feature parameters; and the upsampling module is used for generating the target picture according to the depth information and the spatial feature parameters processed by the three-dimensional feature extraction module. Upsampling, i.e., a technique for enlarging an image to a higher resolution, is usually implemented through inverse convolution, uppooling method, bilinear interpolation and the like. In the embodiment, the image is enlarged to the same specification as the image to be processed through processing by the upsampling module, so that further comparison can be performed for calculation of a loss function.

[50]Preferably, the three-dimensional feature extraction module includes a mapping network, a two-dimensional generative adversarial network generator, and a neural network renderer; the spatial feature parameters include a camera parameter and a hidden scalar; the mapping network is used for processing the hidden scalar to correspond to the specifications of the picture to be processed, and outputting the processed hidden scalar to the two-dimensional generative adversarial network generator and the upsampling module; the two-dimensional generative adversarial network generator is used for generating a feature-deformed feature map according to the processed spatial feature parameters, and inputting the feature-deformed feature map into the neural network renderer; and the neural network renderer is used for processing the feature-deformed feature map and the camera parameter to output the depth information of the picture to be processed.

[51]The neural network renderer includes a three-plane decoder and a voxel rendering module; the three-plane decoder is used for processing the feature-deformed feature map to output voxel information to the voxel rendering module; and the voxel rendering module is used for outputting the depth information of the picture to be processed according to the voxel information and the camera parameter. Optionally, the three-plane decoder includes

three fully connected layers connected in sequence; and the mapping network may likewise include three fully connected layers connected in sequence.

[52] As a preferred implementation mode, by reference to FIG. 3, the encoder includes a downsampling convolutional neural network; and the downsampling convolutional neural network is used for downsampling the picture to be processed. The main objective of downsampling, i.e., shrinking an image, is to make the image conform to a size of a display area and generate a thumbnail of the corresponding image. This is usually implemented by means of a pooling layer or a convolutional layer in the convolutional neural network, and a difference is that a convolutional process results in a smaller image for the purpose of extracting a feature, while pooling downsampling is to reduce a dimension of a feature. Preferably, the encoder further includes a fully connected layer; the downsampling convolutional neural network includes a convolutional layer and a pooling layer; the convolutional layer, the pooling layer and the fully connected layer are connected in sequence; and the convolutional layer is implemented based on an attention mechanism and a ReLU activation function.

[53] The three-dimensional auto-encoder in the embodiment makes full use of a scenario feature perception advantage of 3DGANs to process an image, so as to obtain video frames containing scenario features that can be used for subsequent implementation by means of self-supervision; more scenarios can be covered by using massive training sets, training is performed without a labeled data set, training can be performed by using a large amount of unlabeled original data, which saves the labor and cost, and a trained model can be conveniently transplanted to other scenarios. Inductive bias of a three-dimensional spatial relationship is added to a neural network design, so that interactive information with an environment included in an action can be perceived and the accuracy of recognition can be improved.

[54] Embodiment 2

[55] By reference to FIG. 4, a training method for a three-dimensional auto-encoder is specifically provided in the embodiment; the training method is used for training the three-dimensional auto-encoder in Embodiment 1; and the training method includes:

[56] S0. A first training data set (e.g., 70,000 RGB pictures) is used to train three-dimensional generative adversarial networks (3DGANs), and a generator of the trained 3DGANs is extracted as a decoder of a three-dimensional auto-encoder.

[57]S1. Parameters of the decoder of the three-dimensional auto-encoder are locked, and parameters of an encoder of the three-dimensional auto-encoder are trained using a first training data set until convergence; the first training data set is used for training and obtaining the three-dimensional generative adversarial network.

[58]S2. The parameters of the encoder are unlocked, and the parameters of the encoder and the decoder is continued to be trained until convergence.

[59]In the embodiment, the first data training set may use more than 70,000 three-primary-color (RGB) pictures, which are likewise used for training the three-dimensional auto-encoder.

[60]In the training method for a three-dimensional auto-encoder of the embodiment, a training method is designed for a composition principle of the three-dimensional auto-encoder, i.e., firstly training the encoder, and then training the three-dimensional auto-encoder as a whole, so as to be able to obtain a compliant three-dimensional auto-encoder used for processing images using the scenario feature perception advantage of 3DGANs, and obtain video frames containing scenario features that can be used for subsequent implementation by means of self-supervision; more scenarios can be covered by using massive training sets, training is performed without a labeled data set, training can be performed by using a large amount of unlabeled original data, which saves the labor and cost, and a trained model can be conveniently transplanted to other scenarios. Inductive bias of a three-dimensional spatial relationship is added to a neural network design, so that interactive information with an environment included in an action can be perceived and the accuracy of recognition can be improved.

[61]Embodiment 3

[62]By reference to FIG. 5, a training method for a three-dimensional vision module is specifically provided, including the following steps:

[63]S101. Video frames are processed using the above three-dimensional auto-encoder to obtain three-dimensional spatial features corresponding to the video frames;

[64]S102. The video frames associated with the three-dimensional spatial feature are divided into a plurality of video patches;

[65]S103. Position vectors and time vectors are assigned to the plurality of video patches, respectively;

[66]S104. A first preset proportion of video frames is screened and filtered out randomly according to the time vector;

[67]S105. A second preset proportion of video patches is selected randomly from the remaining video frames as training data, the training data are inputted into an encoder of the three-dimensional vision module, and the outputted video patches plus tokens corresponding to the first preset proportion of video frames filtered out through the screening are inputted into a decoder of the three-dimensional vision module; the video patches are arranged in an order according to the time vector and the position vector; and

[68]S106. The three-dimensional vision module is trained based on the output data of the decoder of the three-dimensional vision module and the video frames until convergence.

[69]In the embodiment, the training data set may be performed using 4 million 10-second-level short video files. Based on the above steps, it is known to those skilled in the art that perception of three-dimensional scenarios of a video frame image can be performed using a processing flow as shown in FIG. 6 to improve an accuracy of action recognition. Specifically, the three-dimensional adversarial generative network is trained, a feature extraction module thereof is taken as the decoder of the three-dimensional auto-encoder, and the three-dimensional auto-encoder is trained to process the video frames and used for pre-training of the three-dimensional vision module 3DViT.

[70]The encoder and decoder of the three-dimensional vision module of the embodiment are shown in FIGs. 7 and 8, respectively, where a three-dimensional video block (3DTubelet) is converted into tokens, and Token Embedding outputs an Tubelet Hidden Representation after combining a position feature and performing spatial code conversion, and combining a temporal feature and performing temporal code conversion, respectively.

[71]It is understood that after the three-dimensional vision module processes pre-training, fine-tuning training can be performed using other small data sets. Fine-tuning means freezing some of convolutional layers of a pre-trained model (usually most of the convolutional layers close to the input) or even training the remaining convolutional layers (usually some of convolutional layers close to the output) and the fully connected layers without freezing any of the network layers, so as to make use of known network structures and known network parameters and fine-tune parameters of several layers before the last layer, and this effectively makes use of a generalization ability of a deep neural network while eliminating the need for a model with complex design and time-consuming training.

[72]The training method for a three-dimensional vision module of the embodiment can achieve perception of three-dimensional scenario features of an image by means of self-supervision; more scenarios can be covered by using massive training sets, training is performed without a labeled data set, training can be performed by using a large amount of unlabeled original data, which saves the labor and cost, and a trained model can be conveniently transplanted to other scenarios. Inductive bias of a three-dimensional spatial relationship is added to a neural network design, so that interactive information with an environment included in an action can be perceived and the accuracy of recognition can be improved.

[73]Embodiment 4

[74]By reference to FIG. 9, an electronic device 30 is provided in the embodiment, including a processor 31, a memory 32, and a computer program stored in the memory 32 and executable on the processor 31; the processor 31 is configured to implement the training method for a three-dimensional self-encoder in Embodiment 2 and/or the training method for a three-dimensional vision module in Embodiment 3 when executing a program. The electronic device 30 shown in FIG. 9 is merely an example and should not impose any limitation on the functionality and application scope of the embodiments of the present invention.

[75]The electronic device 30 may be represented in a form of a general-purpose computing device; for example, the electronic device 30 may be a server device. Components of the electronic device 30 may include, but are not limited to, at least one processor 31 described above, at least one memory 32 described above, and buses 33 connecting different system components (including the memory 32 and the processor 31).

[76]The buses 33 include a data bus, an address bus, and a control bus.

[77]The memory 32 may include a volatile memory, such as a random access memory (RAM) 321 and/or a cache memory 322, and may further include a read-only memory (ROM) 323.

[78]The memory 32 may also include a program/utility 325 having a set (at least one) of program modules 324, such program modules 324 include, but are not limited to, an operating system, one or more applications, other program modules, and program data, and each or some combination of these examples may include implementation of a network environment.

[79]The processor 31 is configured to perform various functional applications and data processing by running computer programs stored in the memory 32, such as the training method for a three-dimensional auto-encoder in Embodiment 2 of the present invention and/or the training method for a three-dimensional vision module in Embodiment 3.

[80]The electronic device 30 may also communicate with one or more external devices 34 (e.g., a keyboard, a pointing device and the like). Such communication may be performed through an input/output (I/O) interface 35. Moreover, the model-generated device 30 may also communicate with one or more networks (such as a local area network (LAN), a wide area network (WAN), and/or a public network, such as the Internet) through a network adapter 36. The network adapter 36 is configured to communicate with other modules of the model-generated device 30 through the bus 33. Other hardware and/or software modules may be used in conjunction with the model-generated device 30, including but not limited to microcode, device drivers, redundancy processors, external disk drive arrays, RAID (disk array) systems, tape drives, data backup storage systems and the like.

[81]It should be noted that although some units/modules or sub-units/modules of the electronic device are referred to in the detailed description above, this division is merely exemplary, rather than mandatory. In fact, according to the implementation mode of the present invention, the features and functions of two or more units/modules described above may be embodied in a single unit/module. On the contrary, the features and functions of the single unit/module described above may be further divided to be embodied by more than one unit/module.

[82]Embodiment 5

[83]A computer-readable storage medium is provided in the embodiment; the storage medium includes a computer program stored thereon, which when executed by the processor, is configured to implement the training method for a three-dimensional auto-encoder in Embodiment 2 and/or the training method for a three-dimensional vision module in Embodiment 3.

[84]Further, the readable storage medium may be used more specifically, and may include, but is not limited to, a portable disk, a hard disk, a random access memory, a read-only memory, an erasable programmable read-only memory, an optical storage device, a magnetic storage device, or any suitable combination of the above.

[85] In possible implementation modes, the present invention may also be implemented as a form of a program product, including a program code which, when the program product is run on a terminal device, is used to allow the terminal device to execute and implement the training method for a three-dimensional auto-encoder in Embodiment 2 and/or the training method for a three-dimensional vision module in Embodiment 3.

[86] Further, the program code used for executing the present invention may be written in any combination of one or more programming languages, and the program code may be fully executed on a user device, partially executed on the user device, executed as a stand-alone software package, partially executed on the user device and partially executed on a remote device, or fully executed on the remote device.

[87] Although detailed descriptions of the embodiments of the present invention are described above, it should be understood by those skilled in the art that this is only illustrative and the protection scope of the present invention is limited by appended claims. Those skilled in the art may make a variety of changes or modifications to these embodiments without departing from the principle and substance of the present invention, but all these changes and modifications should fall into the protection scope of the present invention.

CLAIMS

What is claimed is:

1. A three-dimensional auto-encoder, comprising an encoder and a decoder; the encoder is used for extracting spatial feature parameters of an input picture to be processed and outputting the spatial feature parameters to the decoder; the decoder is used for outputting, according to the spatial feature parameters, a target picture comprising depth information of the picture to be processed; and the picture to be processed and the target picture are used for determining a loss function of the three-dimensional auto-encoder.

2. The three-dimensional auto-encoder according to claim 1, wherein the decoder comprises a three-dimensional feature extraction module and an upsampling module; the three-dimensional feature extraction module is used for obtaining the depth information of the picture to be processed according to the spatial feature parameters; and the upsampling module is used for generating the target picture according to the depth information and the spatial feature parameters processed by the three-dimensional feature extraction module.

3. The three-dimensional auto-encoder according to claim 2, wherein the three-dimensional feature extraction module comprises a mapping network, a two-dimensional generative adversarial network generator, and a neural network renderer; and the spatial feature parameters comprise a camera parameter and a hidden scalar;

the mapping network is used for processing the hidden scalar to correspond to specifications of the picture to be processed, and outputting the processed hidden scalar to the two-dimensional generative adversarial network generator and the upsampling module;

the two-dimensional generative adversarial network generator is used for generating a feature-deformed feature map according to the processed spatial feature parameters, and inputting the feature-deformed feature map into the neural network renderer; and

the neural network renderer is used for processing the feature-deformed feature map and the camera parameter to output the depth information of the picture to be processed.

4. The three-dimensional auto-encoder according to claim 3, wherein the neural network renderer comprises a three-plane decoder and a voxel rendering module;

the three-plane decoder is used for processing the feature-deformed feature map to output

voxel information to the voxel rendering module; and

the voxel rendering module is used for outputting the depth information of the picture to be processed according to the voxel information and the camera parameter.

5. The three-dimensional auto-encoder according to claim 4, wherein the three-plane decoder comprises three fully connected layers connected in sequence; and the mapping network comprises three fully connected layers connected in sequence.

6. The three-dimensional auto-encoder according to claim 2, wherein the encoder comprises a downsampling convolutional neural network; and the downsampling convolutional neural network is used for downsampling the picture to be processed.

7. The three-dimensional auto-encoder according to claim 6, wherein the encoder further comprises a fully connected layer; the downsampling convolutional neural network comprises a convolutional layer and a pooling layer; the convolutional layer, the pooling layer and the fully connected layer are connected in sequence; and the convolutional layer is implemented based on an attention mechanism and a ReLU activation function.

8. The three-dimensional auto-encoder according to any of claims 2-7, wherein the three-dimensional feature extraction module is obtained based on a feature extraction module of a three-dimensional generative adversarial network, and the feature extraction module is used for extracting three-dimensional scenario information of the picture to be processed.

9. A training method for a three-dimensional auto-encoder, wherein the training method is used for training the three-dimensional auto-encoder according to claim 8; and the training method comprises:

locking parameters of a decoder of the three-dimensional auto-encoder, and training parameters of an encoder of the three-dimensional auto-encoder using a first training data set until convergence; and the first training data set is used for training and obtaining the three-dimensional generative adversarial network; and

unlocking the parameters of the encoder, and continuing to train the parameters of the encoder and the decoder until convergence.

10. A training method for a three-dimensional vision module, comprising the following

steps:

processing video frames using the three-dimensional auto-encoder according to claim 8 to obtain three-dimensional spatial features corresponding to the video frames;

dividing the video frames associated with the three-dimensional spatial feature into a plurality of several video patches;

assigning position vectors and time vectors to the plurality of video patches, respectively;

screening and filtering out a first preset proportion of video frames randomly according to the time vectors;

selecting a second preset proportion of video patches randomly from the remaining video frames as training data, inputting the training data into an encoder of a three-dimensional vision module, and inputting the outputted video patches plus tokens corresponding to the first preset proportion of video frames filtered out through the screening into a decoder of the three-dimensional vision module; wherein the video patches are arranged in an order according to the time vectors and the position vectors; and

training the three-dimensional vision module based on the output data of the decoder of the three-dimensional vision module and the video frames until convergence.

11. An electronic device, comprising a memory, a processor and a computer program stored in the memory and executable on the processor, wherein the processor is configured to implement the training method for a three-dimensional auto-encoder according to claim 9 and/or the training method for a three-dimensional vision module according to claim 10 when executing the computer program.

12. A computer-readable storage medium, wherein the storage medium comprises a computer program stored thereon, which when being executed by the processor, is configured to implement the training method for a three-dimensional auto-encoder according to claim 9 and/or the training method for a three-dimensional vision module according to claim 10.