Research Article

Jiangtao Wang*

Development and research of deep neural network fusion computer vision technology

https://doi.org/10.1515/jisys-2022-0264
received November 12, 2022; accepted May 22, 2023

Abstract: Deep learning (DL) has revolutionized advanced digital picture processing, enabling significant advancements in computer vision (CV). However, it is important to note that older CV techniques, developed prior to the emergence of DL, still hold value and relevance. Particularly in the realm of more complex, three-dimensional (3D) data such as video and 3D models, CV and multimedia retrieval remain at the forefront of technological advancements. We provide critical insights into the progress made in developing higher-dimensional qualities through the application of DL, and also discuss the advantages and strategies employed in DL. With the widespread use of 3D sensor data and 3D modeling, the analysis and representation of the world in three dimensions have become commonplace. This progress has been facilitated by the development of additional sensors, driven by advancements in areas such as 3D gaming and self-driving vehicles. These advancements have enabled researchers to create feature description models that surpass traditional two-dimensional approaches. This study reveals the current state of advanced digital picture processing, highlighting the role of DL in pushing the boundaries of CV and multimedia retrieval in handling complex, 3D data.

Keywords: deep neural network, computer vision technology, 3D

1 Introduction

Search engine optimization, sentiment analysis, or item informal groups are examples of applications for machine learning (ML) technologies. Commodities such as smartphones and cameras increasingly come with this. ML algorithms can be used for image identification, language to clearly indicate, curiosity identification for news items, comments, as well as other items, as well as the proper selection of search rankings. Deep learning (DL) is being employed in such types of programs as it gains popularity. Due to this restriction, conventional ML techniques were not able to interpret organic information while it was in its original form. Technologies for feature identification and ML have been created over time with rigorous design and deep subject-matter expertise. For instance, building an internal model or extracted features appropriate for the training component, that is, typically a classification, for a new dataset such as a picture's input image needs years of meticulous design and substantial subject expertise [1,2]. Whenever given raw data, machines are capable of producing the models required for identification or categorization on their own. To do this, a group of approaches known as transfer learning can be applied. In reality, ML systems with several stages of description are realized by constructing simple yet complex components that effectively determine the depiction place at a single layer (beginning with the raw input) into a depiction at a slightly more ideological level. It is possible to master complicated tasks by making a lot of small tweaks over time. We can decrease the quantity of knowledge that is currently needed by enhancing the number of descriptions in classification techniques. In the domain of artificial intelligence (AI) and ML, end-to-end training is a method in which the

* **Corresponding author: Jiangtao Wang**, School of Network Communication, Zhejiang Yuexiu University, Shaoxing, 312000, China, e-mail: 20171034@zyufl.edu.cn

system learns every stage from the input feature stage through the eventual expected output. This supervised neural method trains all the various components concurrently rather than progressively. There are many methods for representing pictures, and the initial surface is constantly made up of learned representations that portray the existence or nonattendance of corners in particular viewpoints or areas, while the second layer is generally made up of learning algorithms that represent the existence or non-availability of corners in all directions as well as positions. For the second level, it is routine to identify border configurations that are detectable even when the edges' positions are slightly changed. The following levels can identify the greater permutations that were generated whenever an object's elements are combined in the third step. Since such characteristics are not determined by individuals, any form of information can be learned using a general-purpose method.

The development of deep neural network (DNN) fusion computer vision (CV) technology comes under the scope of neural network (NN) fusion for CV as it involves the combination of multiple DNN architectures, such as Convolutional Neural Networks (CNNs), recurrent neural networks (RNNs), and transformer networks, to enhance the accuracy and robustness of CV systems. NN fusion is an ML technique that integrates the strengths of different NN models to improve the performance of a given task. The development and research of DNN fusion for CV applies this technique to advance the field of CV by creating more powerful models that can effectively interpret and understand visual data.

1.1 DL

Unsupervised deep neural systems for dimension compression were originally identified as a distinct topic of ML study in 2006. When it won the ImageNet competition in 2012, findings show the large advantage it held over the conveyor. An automatic deep CNN is used to categorize and retrieve images [3]. Kernel flows in two dimensions in two-dimensional (2D) CNN. Information from a 2D CNN's inputs and results are highly complex and are utilized primarily with visual information. Kernel flows in three axes in three-dimensional (3D) CNN. The findings indicate that the learning rate is enhanced by using Thermal Encode on the large datasets by drastically lowering the number of periods or iterations required. The results demonstrated that the efficiency of an NN network is impacted by altering the depiction of the input information. DL consequently had become a well-liked research area in almost each intellectual pursuit in the decades that followed. There have been substantial advancements in the fields of text categorization, natural language synthesis, and image and voice identification. In order to develop a network model on a representation with less characteristics, the complexity of the feature space can be reduced to a tolerable level.

In order to solve the most challenging issues in digital image analysis, DL is widely used. Because of large databases as well as powerful computation, DL experts have managed to move above what was originally believed conceivable (DL). We have been capable of overcoming obstacles that seemed intractable at the time compared to advances in science and technology. The process of classifying photos provides a list of this. Computer simulations have allowed academics to accurately represent the intricate structures present in huge amounts of information.

This is facilitated by the use of numerous processing layers. "Deep learning" encompasses NN models, multilayer reinforcement learning, or a variety of methods for gaining knowledge attribute values. ML is included in transfer learning. The vast quantity of challenging data from diverse resources can be connected to the rising popularity in ML, in contrast to its demonstrated ability to outperform previous state-of-the-art methods in a variety of applications for a basic comprehension of DL. Creating statistical formulas that precisely explain the event under study is the first step in descriptive statistic. In order to perform investigation, an investigator collects information, formulates hypotheses, and then verifies those hypotheses utilizing real statistics from a method or procedure. Therefore, this must be accomplished [4]. If researchers and engineers ignore challenging, obscure, or counterintuitive procedures, bad outcomes may emerge [5]. A subset of analysis tools called predictive analytics employs historical information along with data analysis, information retrieval, or CV to forecast prospective results. Utilizing trends in this knowledge, businesses use

predictive modeling to spot dangers and possibilities. In order to accurately predict the outcome of an event, predictive analysis relies heavily on identifying the underlying laws that control it. It is possible to produce new patterns rather than analyze difficulties by providing the system with a large number of training patterns (a collection of inputs for which desired outcomes are known). As a result of this paradigm shift, traditional programming is now obsolete. DL, a subset of ML that heavily relies on artificial neural networks (ANNs), is a subset of ANNs.

1.2 Advantages of DL

DL distinguishes itself from other ML algorithms because of its end-to-end training and representation-based training. In many cases, end-to-end DL training can be useful. Because the model is so flexible, it may essentially "encode" data, which is why this is the case. Making forecasts for the future using existing and past training datasets requires the application of a variety of mathematical approaches, in addition to data gathering, simulation models, DL, and intelligent systems. If you are employing neural machine translation, you do not need any human input to build the model. In comparison to normal feature-based statistical machine translation, this is a substantial advantage. All data representations can be learned with DL (text and image). As a result, data from various media can be processed. In order to identify the most appropriate event photos, for example, you can match the query (word) against the images.

Progress in DL and advancements in device capabilities, such as central processing unit (CPU), memory, battery, image sensor quality and optics, have accelerated the adoption of vision-based applications. Conventional methods of CV are less accurate in a variety of applications, including image classification, semantic segmentation, object identification, and semantic segmentation and object identification. The data and outcomes were extensively analyzed in standard CV. An effective technique might be combined with the statistically computable information that can be taken from a picture to achieve the intended outcome, according to the thorough study. With large databases, DL may uncover intricate underlying patterns and become more accurate as the amount of information set increases. Additionally, greater technologically difficult, DL methods required a powerful graphics processing unit (GPU). With today's huge video data, there is a need for NNs that are trained rather than coded, which decreases the amount of professional analysis and fine-tuning required for applications that utilize this technology. DL methods are more flexible than CV approaches, which are frequently more application-specific, because they can retrain CNN models and frameworks with a custom dataset [6]. A huge sample is necessary for the deep training system to function well. However, by enhancing the information we have, we could enable the model to function better. The algorithm benefits from being able to generalize to many imagery kinds.

1.3 CV

Many areas of business and management science are undergoing major upheavals [7]. In both academics and industry, CV is becoming increasingly popular. The foundation of traditional computer safety is the frequently cited classification of potential threats, which comprises stealing, secrecy, authenticity, or reliability. These four sorts of diversified threats will, in general, be applicable to vital infrastructure. For a variety of user needs, CV algorithms have already been shown to be effective in the context [7]. The ability to do more difficult support activities will grow in tandem with the development of the underlying knowledge, as is often the case when knowledge is transferred from theoretical to practical fields. While visual intelligence had remained largely steady over the previous decade, the impact of the DL paradigm has pushed it somewhat higher in the last 5 years or so. These scholars had previously dominated this field, but it is now used in a wide range of fields, including image processing, speech recognition, medical imaging, and self-driving vehicles. In the late 1980s, NNs were first used to map inputs to outputs in order to recognize handwritten handwriting automatically.

1.4 Object localization and recognition

The topic of object localization and recognition in CV has evolved greatly in recent years as a result of DL. There are numerous trained models for this task; therefore, it takes little effort to construct an image or video recognition system that can detect the majority of things in an image or video, even if there are several overlapping objects and various backdrops present. Recent DL-based architectures are capable of not only recognizing a large number of objects in a scene, but also accurately determining their boundaries and relationships to one another. For instance, deep structured learning may be used to discover correlations based on features, geometry, and labeling [8], as well as physics and inferences about the abstract properties of the entire system [9].

2 DL methods and developments

2.1 CNNs

Based on the visual system models developed by Hubel, CNNs were developed (1962). When neurons with the same parameters are applied to patches of a previous layer at various locations, translational invariance is gained. The CNN is translation invariant due to linguistic parallelism. If we convert the signals, the CNN will continue to be capable of determining the category that the information corresponds because it is invariant to interpretation. The pooling operation leads to longitudinal directionality. This is explained by the local connections between neurons and hierarchically organized image transformations. A series of various image segmentations at differing stages of segmented information is known as a hierarchical edge detection, and the categorizations at finer degrees of specificity can be created by simply merging areas from market segments at higher information levels. For the first time, this computational paradigm may be found in Neocognitron (Fukushima [10]). For each layer, there is a certain function assigned to it. To convert input data into a one-dimensional (1D) feature vector, CNNs use their final, fully linked layers to convert the volume of neuron activation. From data acquired, CNN can determine an individual's movement, including whether they are seated, moving, leaping, etc. This information has two aspects. Time-steps make up a first component, while the second is the velocity rates along three dimensions. Many CV applications have benefited from CNNs, such as face and object detection, robotic vision, and self-driving cars.

(i) *Convolutional layers.* By converging the image and the intermediate feature maps, the convolutional layers of a CNN build a diverse set of feature maps. For example, Szegedy et al. [11] and Boureau et al. [12] recommend using convolutional layers rather than fully connected layers to obtain faster learning times [12,13].

(ii) *Pooling layers.* To reduce the size of the subsequent convolutional layer's input volume, the spatial dimensions of the prior layers' input volumes are pooled (width and height). Imagine that an object is being examined by a convolutional neural system to determine its information. The nine images that are included in a filtering with a 3×3 pixel resolution will be reduced to one picture in the output nodes. Consequently, the production will decrease as the step, or action, is lengthened. The size of the extracted features is reduced by convolution layer. As a result, it lessens the quantity of system calculation as well as the variety of parameters that must be learned. The data point created by a convolutional gradient information convolution layer describes the characteristics that are available in a certain area. The pooling layer has no effect on the volume's depth due to its continuous nature. As a result of the reduction in size, this layer also performs downsampling, which is referred to as "downsampling." This process is known as downsampling. Then it makes up the initial portion of connectivity that completely utilizes inversion.

(iii) *Fully connected layers.* Fully connected layers of the NN carry out network-level reasoning following several convolutional and pooling layers. They will thereafter be able to access the network. Due to their interconnection with the activity of all the neurons in the previous layer, the neurons in this layer are referred to as "fully interconnected." Because matrix multiplication with bias offset can be used in conjunction with matrix multiplication, its activation can be determined. A fully connected layer provides a biased variable after multiplying the data by only a weight matrix. Each or even more completely linked levels come after the multilayer (or down-sampling) levels. As the title indicates, every synapse in a level that is fully linked has connections to every cell in the level above it. All of the 2D feature maps are combined into a single 1D feature vector.

In the design of CNNs, for example, receptive fields, coupled weights, spatial subsampling, and other concepts are used. Neighboring units in the previous layer's receptive field provide input to convolutional layers. These basic visual elements such as edges and corners may now be extracted from sensory data by neurons. Convolutional layers discover higher-order features by merging lower-order traits into one feature. A 3D CNN was developed by Baccouche et al. [14] using this technique, which is another example of its application. Based on the output of the 3D CNN network, they build an RNN-long short-term memory (LSTM) network to handle long-term activities [15]. RNN extensions that expand the storage are classified as LSTM systems. Construction pieces for an RNN's layers are called LSTM. By giving data strength training, LSTMs enable RNNs to accept additional knowledge, remember it, or provide it some significance to affect the result. Convolutions are carried out via a 3D filtering in a 3D CNN. In contrast to a 2D CNN, which only allows for sliding in two aspects, the operating system can move in three axes. A convolutional layer functions the same as any normal level would: it accepts input, modifies it in a certain manner, and afterward transmits the modified data to the subsequent stage. Fully connected levels' inputs and outcomes are referred to as activation functions and extensive array, respectively. When an LSTM network is stacked on top of a CNN network, RNNs are formed (LRCN). End-to-end training and an ImageNet pretrained CNN have been used to improve the model originally built [16].

This model's generalization, the stacking LSTM, includes numerous hidden LSTM levels, each having a number of computer memory. The model becomes deeper as a result of the layered LSTM convolutional nodes, better appropriately qualifying the method as ML. Therefore, to appropriately respond to your query, let me list the following three aspects of transfer learning that make it unique: a composing hierarchy. The interpretation of the similarity of the Kurdish languages utilizing a scatterplot as well as a nonmetric multivariate visualization methodology is a novel approach that is suggested. On the Kurdish language categorization, the 1D CNN approach can produce forecasts with an overall accuracy of 95.53% [17].

2.2 An AI explosion

During the past few decades, our capacity to identify objects has increased dramatically. CNNs have made tremendous progress in the domains of ML and resume writing. Increasing computing power and data availability to DL models have enabled this rapid progress. AI research has seen a meteoric rise in interest. There has been a fresh wave of cutting-edge visual recognition research prompted by the ImageNet Large Scale Visual Recognition Challenge [18] A sizable graphical collection created to be used in studies on visual machine vision technology is called the ImageNet project. The initiative has manually labeled more than 14 million photographs to identify the things they depict, and, at minimum, one million of those photographs also include connected components. Community creativity has been bolstered by open-source AI research tools that are available to the entire population. Using digital libraries, sophisticated mapping functions can be taught without the need to build a model manually. For high-level features, this is particularly difficult, as a lack of resilience might be caused by a lack of appropriate specification or by changing conditions [21].

2.3 Deep belief networks (DBNs) and deep Boltzmann machines (DBMs)

Restricted Boltzmann machines (RBMs), including deep belief networks (DBNs) and deep Boltzmann machines (DBMs), incorporate a learning component known as the RBM learning module. RBM NNs are generated [20] At lower levels, there is a direct connection between two layers of DBNs. DBMs are used to connect the various network layers. The second form of deep model, called DBM, is constructed using RBM as a component. In comparison to the DBN, the DBM's top two layers consist of an undirected network, while the bottom layers consist of an undirected graphical model. Even and odd-numbered units are conditionally independent of one another in a DBM with many levels of concealed units. A system with additional hidden layers and purposeless interconnections among the terminals is referred to as a DBM. DBMs systematically extract attributes from raw information or encode characteristics from one level as unknown parameters for the subsequent level. However, DBMs have long been recognized as a challenging framework for inference. For a long time, the DBM has been regarded as an especially difficult framework for inferencing. An easier-to-use model could be produced by selecting more complete links between visible and concealed elements. DBMs employ an stochastic maximum likelihood-based technique [21] to optimize the lower limit on the likelihood during network training by simultaneously training all layers of a given unsupervised model rather than optimizing the likelihood directly. System collapse would be near-impossible if many units fall into catastrophic local minima at the same time [22]. It has been suggested that instead of pre-training the DBM layers, we stack RBMs and train each one to mimic the output of the one before it, followed by a final joint fine tuning.

2.4 DL for high-dimensional data

An algorithm taxonomy for high-dimensional data is presented in this section, with descriptions of the algorithms included. Procedures can be classified based on the degree of generality they achieve. A massive computational challenge arises when DNNs are trained on high-dimensional information lacking geometric features. It suggests a network topology with a massive NN, which dramatically raises the number of parameters and frequently renders retraining impossible. Labeling flattening is a basic technique for accelerating neuromorphic learning. CNNs, convolutional auto encoders (CAEs), and other low-dimensional DL algorithms have been modified from their original setups in order to be used with more complicated data sets. Expanding physical dimensions and modalities are two different kinds of methods. No lower dimensional data have been adapted to this model because it was built for high-dimensional data. All DL methods for 2D and 3D (images) are either CNNs or their derivatives, such as CAE, and this holds true for both 2D and 3D.

2.5 NNs and DL

The implementation of DL, on the other hand, is not an easy process, as it is just a more sophisticated algorithm than shallow ML algorithms. In order to offer faster and more precise results than ever before, DNNs require graphic-processing capabilities. Deep neural systems' several levels make it possible for modeling to acquire complex properties faster or to handle increasingly demanding computing tasks, namely, carry out numerous complex functions at once. Google's developed self-driving automobile and Apple's facial recognition system exemplify current worldwide applications of deep learning. Additionally, widely used personal assistants like Siri and Cortana further showcase the practical implementation of deep learning techniques. DL is also offered in Amazon Go locations, as well as other locations. Generalizability is the first of three DL capabilities to consider. Generalizability is defined by the machine's capacity to produce accurate estimations on unformed input. The second quality is a DL framework's ability to quickly adapt to changing circumstances. We can determine a machine's expressibility by examining its ability to make valid generalizations. The third and final

criterion to consider is expressibility. Among the criteria to evaluate are interpretability, appraised quality, and transferability; inertness; ill-disposed soundness and security; and transferability [23].

2.6 Computing hardware for DL

Computer architectures must be redesigned to meet the increased demands of DL. Some of the constraints include the need for speed in order to reduce training durations, the need for low-energy consumption when implementing DL on mobile devices, and the need for massive memory needs for DNNs. There are many types of computer architectures that can be utilized for ML applications, including a computer's CPU and its GPU [24]. The CPU is a general purpose device that can handle a wide range of activities. For graphic design or artificial learning activities, a GPU (graphics processing) is a specialized functional unit with improved arithmetic device. GPUs are capable of handling several operations at once. As a result, training procedures can be distributed, which greatly speeds up ML activities. With GPUs, you may add several cores with lower material requirements while compromising performance or energy. To meet the demands of individual data-link applications, field-programmable gate arrays and application-specific integrated circuits can be customized (DL applications). An efficient and effective system known as a field programmable gate array comprises a computer's hardware components with user-programmable connectors to tailor performance for a particular purpose. A premade semiconductor device called a gateway matrix contains the majority of its pixels with no predefined purpose. Metal coatings can be used to interconnect such devices as well as create conventional NAND or NOR logical operations. Training DNNs can benefit greatly from GPUs, a form of parallel computing architecture that can run at speeds many orders of magnitude faster than a computer's central processor. A huge number of "neurons" in NNs enables the processing of massive amounts of data at the same time [25, 26]. With TOPS/Watt values average 30–80 times higher than GPU and CPU, a Google-developed TPU surpassed them both in normal data center workloads. There are a number of strategies that can help in the creation of hardware for DL, including compression, acceleration, and regularization.

2.7 Current challenges and future

DL has recently gained attention as a critical area of study for the advancement of AI. DL may readily be utilized to power existing AI applications such as picture and speech recognition, text processing, and Natural Language Processing. DL algorithms, which are advancing in sophistication as ANNs get more complicated, are increasingly emulating the functioning of human brains. It is vital to keep an eye out for vast amounts of data, large amounts of information, and large amounts of information in DL systems. Hyperparameter optimization and overfitting are two words used in NNs. NNs are essentially a black box that requires high-end technology to operate efficiently and offer little flexibility or multitasking.

It has been surpassed in popularity by rigorously supervised learning, which has renewed interest in DL as a result of its success. A machine vision technology called object recognition makes it easier to identify items in pictures and movies. One of the main results of DL or computer training systems is entity identification. Convolutional neural systems are the predominant structure utilized for image identification and identification applications (CNNs). Despite the fact that our review did not spend much attention to unsupervised learning, it is probable that it will become more relevant in the future. In comparison to animals, humans and the great majority of other creatures acquire the most of their talents through observation rather than being taught the names of individual objects. It is an active method that sequentially samples the optic array in an intelligent, task-specific manner, with the fovea being narrow and high resolution and the surround being large and low resolution, a process known as progressive sampling. A significant chunk of future vision development is likely to come from end-to-end trained systems that combine ConvNets and RNNs and employ reinforcement learning to select where to look next. When it comes to classification, deep-learning and

reinforcement-learning systems are still in their infancy, yet they have already exceeded passive vision systems in this discipline. According to industry analysts, DL is expected to have a significant impact on the field of natural language processing in the next years. We believe that by training RNNs to focus on a particular aspect of a sentence or the totality of a document, they can improve their accuracy significantly. The future of AI is in the development of representation learning and advanced reasoning systems. New paradigms are necessary to replace rule-based manipulation of symbolic expressions with operations on huge vectors for speech and handwriting recognition, which have been used successfully for a long period of time.

3 DL and its working process in recent developments

Each of these difficulties will be overcome as DL improves in the realm of CV and its associated problems such as categorization reorganization, identification, language processing, and video processing. The classification reduces the material to a significantly reasonable fraction and offers the framework for looking up actual expertise. Recent model-based research has seen a substantial evolution of modern CV models based on CNNs . The principal applications of a CNN, which comprises one or more consecutive levels, are image enhancement, categorization, localization, or other automatically associated information. In essence, a convolutional is a filtering that is dragged over the data. There are nine intermediary outputting algorithms altogether, and nine feature map images as a result. Weights of pre-trained DL configurations can now be adjusted easily to generate a range of DL configurations. For instance, ImageNet can be used to categorize photographs in a variety of ways. While object detection and segmentation are challenging problems to solve, they require creative ways to overcome their complexity. In contrast to object detection, which is concerned with learning the things and building a rectangular bounding box around them, segmentation is concerned with locating the individual pixels that correspond to each object. As one of the most diverse characteristics of image categorization, it is defined by the fact that a single image may contain a diverse variety of items and people of varying sizes and shapes. DL is necessary, but it also presents considerable hurdles, according to Nick Reed, academy director at the Transport Research Laboratory. In-depth training is essential, but it also causes significant obstacles, as he points out. Uncertainty persists about DL's origins.

DL-based alternatives to typical CV algorithms have gained popularity in recent years. A growing number of studies have focused on using DL to handle CV problems involving text, audio, images, and even graphs since 2012. History and digital library applications were also examined throughout this research project. A section on current events and topics was also agreed upon. Following a literature review, Rusu and Cousins (2015) conducted an in-depth assessment of NNs and DL [27]. Check out this collection of deep-learning-related review articles for additional information. They are all reporting on 1D and 2D data-based accomplishments (i.e., text, sound, and images). Studying several strategies based on DNN architectures and aimed at solving 3D compute vision challenges was crucial to filling a gap in the literature.

Since its inception, DL has shown great promise as a paradigm for automating decision-making processes. This is because of the close connection between DL and the way the human brain processes information. Figure 1 illustrates the traditional ML methods. In an ML scenario, feature extraction and classification appear to be carried out independently, necessitating a complex design and a significant amount of important

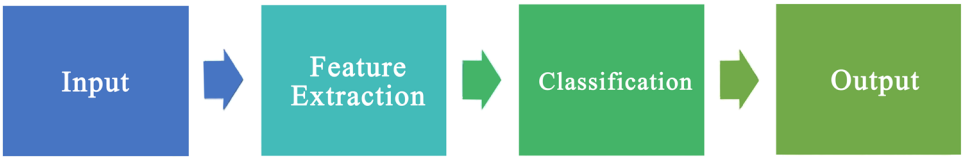

Figure 1: Traditional ML flow.

mathematics on the user's part. Even when precisely following the instructions, the system was not always efficient and performed badly in real-world applications.

DL and NNs are essentially indistinguishable in terms of how they work. For both supervised and unsupervised learning, DL is a useful tool. It has previously been said that DL is capable of solving complex CV problems that ML was unable to. Training and testing were the two main components of the process.

First, there is a lengthy period of data classification and feature definition in the training stages. When the training model encounters similar data, it compares and keeps these characteristics in order to execute proper reasoning and conclusion. In order to build a system, we first divide it into three categories: Classification model, Verification information, and Test samples. A typical DL training program has the following stages: If you have a series of binary True/False questions, an ANN can be used to answer them. Use the data bar to calculate the numerical values. Classify data based on the responses you received. Identify the data and the data source.

During testing, it is necessary to classify fresh and unexposed data, and then draw conclusions utilizing past expertise. To build on the foundation of existing data, traditional ML methods use data that have already been loaded into the computer. As an ML analyst, the analyst generates ML instructions and fixes machine errors. The overtraining effect that is common in DL is eliminated by this strategy. This is known as supervised learning in the context of ML, where the analyst offers examples and training data to help the system make appropriate conclusions. The system gets overfitted and fails to generalize successfully to updated information whenever it memorizes the disturbance or models the learning algorithm excessively accurately. A system will not be capable of carrying out the categorization or forecasting activities that it was designed for if it cannot generalize successfully to updated information. For massive NN models, overfitting can be easily controlled using regularization techniques such value decaying. Utilizing early halting with washout as well as a weighting restriction is a contemporary suggestion for regularization. The computer can perform many tasks with the help of standard ML, but it cannot do so without human supervision. ML vs. DL: what is the difference? While DL requires a vast amount of unlabeled training data in order to make clear findings, ML requires only a small bit of input from the analyst to draw conclusions. The assumption is that analysts will be able to reliably recognize features in ML, but DL creates new features on its own. This is a major difference between ML and DL. DL solves the problem from start to finish, whereas ML breaks down tasks into smaller components and then combines and finishes them into a single output. As a result, the training time for DL is much higher than for ML. It is easier to understand ML's conclusions than DL's. For as long as it is possible, DL implies that the machine makes its own useful decisions. In a multi-tiered structure of assistance, that is a conceptual method for continual improvement, all tiers of the education sector employ data-based problem-solving as well as decision-making to assist learning. A multi-tiered system of endorse assists educational institutions and constituencies in assigning funds via the adherence and long-term sustainability of the implementation of educational performance and behavioral aspirations, thereby accelerating the achievement of any scholar to accomplish and/or surpass competency. As a summary, DL applications adopt a multi-tiered approach, which includes choosing the most important features to study.

4 Non-deep state-of-the-art systems addressing 3D CV tasks

This section discusses 3D segmentation, retrieval, and recognition using simplified approaches that are summarized for ease of comprehension in Tables 1 and 2.

4.1 Software and datasets

Given the growing availability of 3D data (from Kinects, laser scanners, and other devices), it is vital that software for processing 3D data (from Kinects, laser scanners, and other devices) is efficient and trustworthy. This section discusses how to use open-source libraries to process and exploit 3D data. Additionally, it

Table 1: Non-deep 3D segmentation systems

Method	Type	Input	Performance
Schnabel et al. [28]	Model fitting	3D point clouds	Good performance on primitive shape detection and decomposition
Nüchter et al. [29]	Model	3D laser scans	Good results using simulated and real data
Douillard et al. [30]	Clustering	Dense and sparse 3D point clouds	Tested on four hand-labeled point clouds using two novel metrics
Huang et al. [31]	Graph-based	3D point clouds	Two indoor scans of a laboratory environment, outdoor scans of the campus of Jacobs University, visual evaluation
Matsuzaki and Komorita [32]	Aupervoxeli?	Voxelized point clouds	Better oversegmentations than the state of the art (Achanta et al. [2012], Weikersdorfer et al. [2012]) in terms of undersegmentation error, and equivalent to the best performing method (Achanta et al. [2012]) in boundary recall using NYU Depth Dataset V2
Aijazi et al. [33]	Supervoxels	Sparse LIDAR point, clouds	Overall segmentation accuracy: 87% using 3D Urban Data Challenge dataset and 3D datasets of Blaise Pascal University
Stuckler and Behnke [34]	Probabilistic	RGB-D data, stereo image sequences	78.5% correctly labeled mesh faces on NYU depth dataset (indoor) and 77.05% recall on K1TTI odometry dataset (outdoor) versus 67.5 and 65.4% of Ladickv et al. [2009], respectively

describes several newly introduced DL software libraries and implementations of DL-based systems for 3D CV. Additionally, there is a collection of testable 3D datasets.

4.1.1 DL software libraries

As DL technologies gain popularity, a number of software libraries have been developed to aid in their implementation and assessment. Numerous organizations, both academic and commercial, are developing and funding DL software. Caffe, Theano, and Torch are likely to be the most familiar open-source frameworks to you if you are looking for the most popular open-source frameworks, which includes the three libraries mentioned earlier. LeNet was trained on the MNIST datasets [35], with the purpose of training two CNNs for digit classification and image recognition. Google and Microsoft announced the availability of new DL frameworks only lately. It was created in 2011 by the Google Brain project researchers. It was introduced in 2011 and has since grown in popularity. Despite the fact that the core was created in C++, it includes C++ and Python front-ends. From mobile devices to large-scale distributed systems, heterogeneous systems are all capable of performing Tensor-Flow calculations.

4.1.2 Libraries for 3D processing

You may use the Point Cloud Library (PCL) for commercial and research purposes [35]. PCL is a large-scale open project for 2D/3D image and point cloud processing that was initiated by Rusu and Cousins. The first official release, which was made available in 2011, was distributed under the BSD license. PCL is a cross-platform application created in C++. PCL has been partitioned into distinct libraries to make it easier for programmers to work with it. PCL 1.7.2 has 16 modules as of this writing, including common, features filters, geometry, io, kdtree keypoints, out-of-core recognition, registration sample consent, consensus search, segmentation surface, and visualization.

4.2 3D datasets

The 3D reconstruction approach can be used to create 3D representations of objects using only a single or a large number of images. A group of academics have devised a method for learning-based 3D reconstruction of

Table 2: "Nondeep" 3D object retrieval systems

Method	Input	Descriptor	Matching	Performance
Chen et al. [36]	2D views	Light Field Descriptor (LFD)	LI-distance	State-of-the-art performance on a dataset with >10k 3D objects
Kazhdan et al. [37]	3D models	Spherical harmonic representation	L2-distance	Improved performance on a dataset of household objects
Zioulis et al. [38]	3D models	Spherical trace transform (STT)	LI-distance	Outperformed service oriented architecture (SoA) methods (e.g., LFD [Chen et al. 2003]) on two datasets (PSB, W'eb-dataset) in terms of precision, recall
Mademlis et al. [39]	Volumetric represent.	3D shape impact descriptor	Dimension descriptor, Multidimensioned descriptor	Competitive versus, for example, Gaussian Euclidean distance transform, LFD on PSB, ESB. ITI
Daras and Axenopoulos [40]	2D views, 2D images or sketches	Compact multi-view descriptor (CMVD)	LI-distance	Outperformed SoA methods (e.g., LFD [Chen et al. 2003], BoW) on ITI database, PSB, ESB
Gao et al. [41]	2D views	Zernike moments	Distance learning	Good results on three datasets (ETH, NTU, and SI1REC 2010)

generic photographs. On the other hand, learning-based approaches have a lot of promise, but they are still in their infancy. As a result, a review of the study and application of these methodologies in the literature is necessary to enhance them further.

4.3 Implementations of DL techniques for 3D data

As mentioned previously, there are numerous DL libraries available for developing systems based on DL and NNs. If you are interested in a certain activity, it may be prudent to begin with a state-of-the-art method to familiarize yourself with the techniques used. Using experimental methods to enhance the efficiency and efficacy of existing algorithms or to develop new ones can result in significant advances in both. To foster collaboration in this area, a number of researchers have shared their testing code with the scientific community. The availability of open-source implementations of both ancient and current algorithms simplifies the process of comparing old and new algorithms.

Large 3D datasets are now more accessible than ever before, owing to the evolution of quicker 3D sensors. Numerous 3D datasets have been generated and released in recent years to assist academics working in the subject of 3D processing in becoming more productive. The Point Cloud Data (.pcd) format is the most often used file type for 3D objects or scenes contained in these datasets. Polygon File Format (.ply), Wavefront File Format (.obj), Object File Format (.off), and Object File Format (.obj) are the most often used file formats (.off). The Stanford Graphics Lab developed the Polygon File Format (.ply) in the mid-1990s to save graphical objects in a collection of polygons. It made its debut in the mid-1990s. Binary and ASCII are the two supported formats. Each ply file contains a description of the characteristics of a single object. The file provides information on the object's vertices and faces, as well as its color and normal direction. The file's structure is explained in a header at the file's beginning. Wavefront Technologies invented the .objfile format for storing geometric definitions, which is currently extensively used. Typically, an obj file is ASCII-encoded and contains geometric vertices and textures, vertex normals, and polygonal faces. A second material file (.mtl) can be used in conjunction with an obj file to record information about the face color. Numerous attempts have been made in the past to address the difficulty of 3D reconstruction, utilizing traditional CV and ML techniques, among others. Additionally, we believe this is the first time we have seen such a thorough analysis of DL for 3D reconstruction.

These 3D models are used to perform a variety of tasks associated with object comprehension, including detection and categorization, shape understanding, and other analogous tasks. It is possible that some of these collections, such as CAD models, contain 3D photos or scans of real-world artifacts. Apart from that, additional datasets are used for a number of purposes [42–45].

The datasets in Table 3 are frequently used to evaluate 3D CV methods such as 3D object identification, 3D scene segmentation, and 3D shape retrieval.

They are frequently used to evaluate a large number of algorithmic approaches concurrently. Along with the dataset's title and brief description, the data collection equipment used, the data format, and a download link are all included. Additionally, Guo's report on other publicly available 3D datasets and a brief overview of some essential 3D acquisition processes are presented in Table 3 [45].

5 Discussion

The development of DNN fusion technology involves several crucial steps to enhance the accuracy and robustness of visual recognition systems. Initially, the data must be carefully curated, preprocessed, and augmented to increase diversity. Next, multiple DNNs are trained to focus on specific aspects of the visual recognition problem. These individual models are then combined using techniques such as averaging or attention mechanisms to create a unified system. Finally, the fused model is evaluated on a validation set and fine-tuned if necessary. The benefits of this technology include improved accuracy and robustness in

Table 3: Indicative 3D datasets

Dataset	Description	Capturing device	Format	Link
Desk 3D	Six objects and over 850 test scenes	Kinect	Polygon File Format (.ply)	https://sites,google.com/site/ujwalbonde/pulications/downloads.
ACCV3D	Video sequences of 15 textureless objects, each with over 1,100 test frames	Kinect	Polygon File Format (.ply)	http://campar.in.tum.de/Main/StefanHinterstoisser.
NYU-Depth V2	~408,000 unlabeled RGB-D images from 464 indoor scenes (1,449 densely labeled frames)	Kinect	RGB and depth images	http://cs.nyu.edu/siberman/datasets/nyu_depth_v2.html.
RGB-D Object Dataset	300 instances of household objects from 51 categories, 250,000 RGB-D images in total	Kinect	Point Cloud data (PCD)	http://rgbd-dataset.cs.washington.edu/dataset/.
RGB-D Scenes Dataset	22 annotated video sequences of natural scenes containing objects from the GRB-D Object Dataset	Kinect	Point Cloud Data (PCD) or Polygon File Format(.ply)	http://rgbd-dataset.cs.washington.edu/dataset.
Cornell RGB-D	24 labeled office scene point clouds and 28 labeled home scene point clouds	Kinect	Point Cloud Data (PCD)	http://pr.cs.cornell.edu/sceneunderstanding/data/data.php.
Bologna Kinect	This dataset is composed of six models and 16 scenes	Kinect	Polygon File Format (.Ply)	http://vision.deis.unibo.it/research/78-cvlab/80-shot.
SUN3D	Full 3D spaces with semantics: RGB-D images, object segmentations, camera poses, and point clouds registered into a global coordinate frame. In total, 415 sequences from 254 different spaces in 41 buildings.	ASUS Xtion PRO Live	Polygon File Format (.ply)	http://sun3d.cs.princetion.edu/.
A large dataset of Object Scans	Over 10,000 dedicated RGB-D scans of individual objects along with 398 reconstructed models from nine categories	Prime Sense Carmine	Polygon File Format (.ply)	http://redwood-data.org/3dscan/.
Princeton Model Net 40	12,311 unique objects from 40 categories	CAD	Object File Format(.off)	http://modelnet.cs.princeton.edu/.
McGill 3D Shape Bench-mark	19 object categories each with 20–30 3D models (457 models in total)	CAD	Voxelized (im). Mesh (.ply) and Object File Format (.off)	http://www.cim.mcgill.ca/shape/bench.Mark/.
Princeton Shape Bench Mark	1,814 objects from 161 categories		Object File Format (.off)	http://shape.cs.prinecton.edu/benchmark/.
NTU 3D Model Bench Mark	1,833 3D models from which 549 (mainly vehicles and household items) were categorized into 47 categories		Wave front file format (.obj)	http://3d.csie.ntu.edu.tw/dynamic/benchmark/index.html.
TOSCA high resolution	80 object models from nine categories	Synthetic meshes	MATLAB (.mat) and ASCII text files	http://tosca.cs.technion.acil/book/resources_data.html.
Bologna Stanford	45 scenes each containing a subset of six models	Synthetic	Polygon File Format (.ply)	http://vision.deis.unibo.it/research/78-cvlab/80-shot.
Sydney Urban Objects	Scans of 631 urban objects from 26 categories	Velodyne LIDAR	Ascii.csv or binary	http://www.acfr.usyd.edu.au/papers/Sydney Urban Objects Dataset.shtml.
UZH	Scans of 40 academic offices, with two to four scans per office	Faro LINDAR	Polygon File Format (.ply)	http://www.ifi.uzh.ch/vmml/publiacations/ObjDetandClas.html.
Indian Pines	A 145 × 145 pixel image with 220 spectral bands	AVIRIS sensor	.tif file	http://purr.purdue.edu.publications/1947/1.
Pavia University	610 × 340 image with 115	ROSIS sensor	MATLAB(.mat)	http://www.ehu.eus/ccwintco/indx.php?title = Hyperspectral _Remote_sensing_scenes.

challenging scenarios, enhanced interpretability, and reduced risk of overfitting. DL has been widely implemented in CV to address numerous challenges such as image classification, object detection, semantic segmentation, and face recognition. Various DL models such as ResNet, VGG, Inception, and MobileNet are commonly used for image classification, while models such as Faster R-CNN, YOLO, and SSD are often utilized for object detection. Semantic segmentation is accomplished using models such as U-Net, SegNet, and DeepLab, while face recognition is achieved with models such as FaceNet, DeepFace, and VGGFace. DL models using techniques such as object detection, semantic segmentation, and optical flow are used in autonomous driving to perceive the environment, make decisions, and control the vehicle. As the accessibility of large-scale image datasets and powerful computing resources continues to increase, we can anticipate witnessing further groundbreaking applications of DL in CV.

The field of DL has brought significant advancements to CV, but it still faces several challenges that must be addressed to enhance its effectiveness. These challenges include the need for large amounts of labeled data, the risk of overfitting, difficulties in interpreting the models' predictions, the high demand for computational resources, domain-specific challenges, vulnerability to adversarial attacks, and difficulty in generalizing to new scenarios. Tackling these challenges requires continuous research and innovation to overcome the limitations of DL and further improve its applications in CV.

6 Conclusion

As a result of DL, we could expect a significant increase in the study into how to better reflect human behavior and cognition. For people with disabilities, adapting their assistive technology is essential. The relationship between eye movements and cognitive processes could be examined to provide insight into memory recall, cognitive load, interest, domain knowledge, problem solving, desire to learn, and reasoning strategy use because eye gaze has been extensively studied in interactive intelligent systems as a cue for inferring users' internal states and establishing priors about users' intent. Using eye gaze as a cue for determining users' internal emotions and prioritizing their intentions has been extensively studied in interactive intelligent systems.

Early research suggests that DL may be able to outperform traditional feature-based approaches in terms of recognition accuracy (i.e., object classification, recognition and detection, semantic segmentation, and human action classification). While this is true, the implementation time for specialized feature-based solutions is far shorter. With regard to object detection, Georgette et al. [46] demonstrated that their results are comparable to those obtained by DL technology. The IDT technique has been shown to supplement DL features [45], significantly improving a system's overall performance using human action recognition as an example. Despite early research suggesting that data compression and the use of massively parallel systems outperformed raw processing of high-dimensional data, we are currently seeing the reverse trend in our data analysis. HAR outperforms 2D projection techniques to object detection, according to Brock et al. and Carreira and Zisserman [47,48]. Fusion across several processing layers and stages appears to outperform all other techniques. Although significant progress has been made, there is still a lot of room for development in this area. Most datasets have additional dimensions, and no single approach or solution will work for all of them. There is no such thing as a "one size fits all" when it comes to dealing with these dimensions. The time–space distinction nonetheless remains unresolved despite much research into video comprehension. Furthermore, it is not known which raw data format, such as point clouds, 3D meshes, or voxelized data, would be appropriate for 3D static model.

Funding information: This study did not receive any funding in any form.

Author contributions: Jiangtao Wang contributed to the design and methodology of this study, the assessment of the outcomes, and the writing of the manuscript.

Conflict of interest: The author declares that there is no conflict of interest regarding the publication of this article.

Code availability: Not applicable.

Data availability statement: The data used to support the findings of this study are available from the corresponding author upon request.

References

[1] Rezaeianjouybari B, Shang Y. Deep learning for prognostics and health management: State of the art, challenges, and opportunities. Measurement. 2020;163:107929.

[2] Salakhutdinov R, Hinton G. Deep Boltzmann machines. In Proceedings of the International Conference on Artificial Intelligence and Statistics. Vol. 24; 2009. p. 448–55.

[3] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. NIPS'12 Proc 25th Int Conf Neural Inf Process Syst. 2012;1:1097–105 2

[4] Bonaccorso G. Mach learning algorithms popular algorithms for data science and machine learning. Vol. 4, 2nd edn. Birmingham, UK: Packt Publishing Ltd; 2017. p. 56–67.

[5] Mahony NO, Murphy T, Panduru K, Riordan D, Walsh J. Improving controller performance in a powder blending process using predictive control. In: 2017 28th Irish Signals and Systems Conference (ISSC). IEEE; 2017. p. 1–6.

[6] O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, et al. Deep learning vs traditional computer vision. In Science and Information Conference. Cham: Springer; 2019, April. p. 128–44.

[7] Leo M, Medioni G, Trivedi M, Kanade T, Farinella G. Computer vision for assistive technologies. Computer Vis Image Underst. 2017;154(Supplement C):1–15.

[8] Zhu Y, Jiang S. Deep structured learning for visual relationship detection. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18); 2018.

[9] Battaglia P, Pascanu R, Lai M, Rezende DJ. Interaction networks for learning about objects, relations and physics. Adv Neural Inf Process Syst. 2016;110:4502–10.

[10] Fukushima K. Recent advances in the deep CNN neocognitron. IEICE Nonlinear Theory Appl. 2019;10:304–21.

[11] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15). Boston, Mass, USA; June 2015. p. 1–9.

[12] Boureau YL, Ponce J, LeCun Y. A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the ICML; 2010.

[13] Ioannidou A, Chatzilari E, Nikolopoulos S, Kompatsiaris. I. Deep learning advances in computer vision with 3D data. ACM Comput Surv. 2017;50(2):1–38.

[14] Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A. Sequential deep learning for human action recognition. Interntional Workshop Hum Behav Underst. 2011;11:29–39.

[15] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the CVPR. IEEE; 2015. p. 2625–34.

[16] Ghafoor KJ, Rawf KMH, Abdulrahman AO, Taher SH. Kurdish dialect recognition using 1D CNN. ARO-The Sci J Koya Univ. 2021;9(2):10–4.

[17] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. Dec. 2015;115(3):211–52.

[18] Molleda J, Usamentiaga R, García DF, Bulnes FG, Espina A, Dieye B. An improved 3D imaging system for dimensional quality inspection of rolled products in the metal industry. Comput Ind. Dec. 2013;64(9):1186–1200.

[19] Han S. Pool J. Tran J. Dally W. Learning both weights and connections for efficient neural network. Adv Neural Inf Process Syst. 2015;12:1135–43.

[20] Younes L. On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. Stoch Stoch Rep. 1999;65(3–4):177–228.

[21] Salakhutdinov R, Larochelle H. Efficient learning of deep Boltzmann machines. In Proceedings of the AISTATS; 2010.

[22] Patel P, Thakkar A. The upsurge of deep learning for computer vision applications. Int J Electr Comput Eng. 2020;10(1):538.

[23] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2015;61:85–117.

[24] Kim H, Nam H, Jung W, Lee J. Performance analysis of CNN frameworks for GPUs. In 2017 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS); 2017. p. 5564.

[25] Lawrence J, Malmsten J, Rybka A, Sabol DA, Triplin K. Comparing TensorFlow deep learning performance using CPUs, GPUs, local PCs and cloud. Student-Faculty Res. Pleasantville, New York: Day, CSIS, Pace Univ; 2017.

[26] Georgiou T, Liu Y, Chen W, Lew M. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. Int J Multimed Inf Retr. 2019;9:135–70.
[27] Rusu RB, Cousins S. 3D is here: Point Cloud Library (PCL). In IEEE International Conference on Robotics and Automation (ICRA'11); 2011. p. 1–4.
[28] Schnabel R, Wahl R, Klein R. Efficient RANSAC for point cloud shape detection. Comput Graph Forum. 2007;26:214–26.
[29] Nüchter A, Gutev S, Borrmann D, Elseberg J. Skyline-based registration of 3D laser scans. Geo-spatial Inf Sci. 2011;14:85–90.
[30] Douillard B, Underwood J, Kuntz N, Vlaskine V, Quadros A, et al. On the segmentation of 3D LIDAR point clouds. IEEE Int Conf Robot Autom. 2011;8(15):1–10.
[31] Huang QD, Dong XY, Chen DD, Zhou H, Zhang WM, Yu NH. Shape-invariant 3D adversarial point clouds. IEEE Conference on Computer Vision and Pattern Recognition. Vol. 18, 2022. p. 15314–23.
[32] Matsuzaki K, Komorita S. Efficient deep super-resolution of voxelized point cloud in geometry compression. IEEE Sens J 23:1328–42.
[33] Aijazi AK, Checchin P, Trassoudaine L. Segmentation based classification of 3D urban point clouds: a super-voxel based approach with evaluation. Remote Sens. 2013;5:1624–50.
[34] Stuckler J, Behnke S. Efficient dense rigid-body motion segmentation and estimation in RGB-D video. Int J Comput Vis. 2015;113:233–45.
[35] LeCun Y, Bengio Y, Hinton GE. Deep learning. Nature. 2015;521(2015):436–44.
[36] Chen DY, Tian XP, Shen YT, Ouhyoung M. On visual similarity based 3D model retrieval. In Proc. Eurographics. Vol. 4, 2003. p. 223–32.
[37] Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3D shape descriptors. Proc. Symposium of Geometry Processing. Vol. 2, 2003. p. 4–6.
[38] Zioulis N, Karakottas A, Zarpalas D, Daras P. Omni-depth: dense depth estimation for indoors spherical panoramas. European Conference on Computer Vision (ECCV). Vol. 3, 2018. p. 448–65.
[39] Mademlis A, Daras P, Tzovaras D, Strintzis MG. 3D object retrieval using the 3D shape impact descriptor. Pattern Recognit. 2009;42:2447–59.
[40] Daras P, Axenopoulos A. A 3D shape retrieval framework supporting multimodal queries. Int'l J Comput Vis. 2010;89:229–47.
[41] Gao SY, Zhao MY, Zhang L, Zou Y. Improved algorithm about subpixel edge detection of image based on Zernike orthogonal moments. Acta Automatica Sin. 2008;34:1163–8.
[42] Shilane P, Min P, Kazhdan M, Funkhouser T. The princeton shape benchmark. In: Shape Modeling Applications, 2004. Proceedings. IEEE; 2004. p. 167–78.
[43] Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, et al. 3D shapenets: A deep representation for volumetric shapes. In: Proceedings of the CVPR. IEEE; 2015. p. 1912–20.
[44] Liu Y, Guo Y, Georgiou T, Lew MS. Fusion that matters: convolutional fusion networks for visual recognition. Multimed Tools Appl. 2018;77:1–28.
[45] Wang H, Kläser A, Schmid C, Liu CL. Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis. 2013;103:60–79.
[46] Georgette A, Yaakov S, Christian H. Age-related disintegration in functional connectivity: Evidence from Reference Ability Neural Network (RANN) cohort. Neuropsychologia. 2021;156:107856.
[47] Brock A, Lim T, Ritchie J, Weston N. Generative and discriminative voxel modeling with convolutional neural networks; 2016.
[48] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the CVPR. IEEE; 2017. p. 4724–33.