International Journal of Multimedia Information Retrieval (2020) 9:81–101
https://doi.org/10.1007/s13735-019-00190-x

REGULAR PAPER

A study on deep learning spatiotemporal models and feature extraction techniques for video understanding

M. Suresha[1] · S. Kuppa[1] · D. S. Raghukumar[1]

Received: 20 May 2019 / Revised: 7 December 2019 / Accepted: 21 December 2019 / Published online: 24 January 2020

Abstract
Video understanding requires abundant semantic information. Substantial progress has been made on deep learning models in the image, text, and audio domains, and notable efforts have been recently dedicated to the design of deep networks in the video domain. We discuss the state-of-the-art convolutional neural network (CNN) and its pipelines for the exploration of video features, various fusion strategies, and their performances; we also discuss the limitations of CNN for long-term motion cues and the use of sequential learning models such as long short-term memory to overcome these limitations. In addition, we address various multi-model approaches for extracting important cues and score fusion techniques from hybrid deep learning frameworks. Then, we highlight future plans in this domain, recent trends, and substantial challenges for video understanding. This survey's objectives are to study the plethora of approaches that have been developed for solving video understanding problems, to comprehensively study spatiotemporal cues, to explore the various models that are available for solving these problems and to identify the most promising approaches.

Keywords Spatiotemporal · Deep learning · Video understanding · Computer vision · Survey

1 Introduction

The amount of digital multimedia content, such as image, text, audio and video content, is growing exponentially. Video has become a common communication medium or transmission module between Internet users with the proliferation of sensors such as productive mobile devices. Recently, video understanding has been extensively studied by the research community. Significant developments have been recently realized in the design of various robust features for video understanding, which are expected to correspond to class variations. For instance, suppose someone utilizes a handcrafted image-based feature set, such as the SIFT [52,108], to explore the spatial features of a video. In addition to this static frame information, motion provides profound importance cues for video understanding. For this purpose, the most popular feature is dense trajectory [5,63,94], which is used to capture the densely sampled local frame patches. Therefore, spatial and temporal motion patterns are more important for video understanding tasks, as illustrated in Fig. 1.

Conventional video representation methods are motivated from image analysis domain, which can be extended into temporal dimension video data [6,43,72,73,94,95,110], and this focuses on exploring powerful spatiotemporal features using some handcrafted feature extracting techniques. Recently, the immense growth of convolutional neural network (CNN) [75] in the spatial domain [20] and more attempts drawn-out into temporal domain [37] from this CNN [37,85,107,110] using optical flow frames. As a result, video understanding may become easy through solving a different set of problems like video content classification, action prediction, motion recognition, etc. Nowadays, most of the video understanding problems solved by these types of deep neural network (DNNs) and spatiotemporal cues play a significant role. In general, classical and modern machine vision approaches are used to extract these features. Especially in modern machine vision, CNN or state-of-the-art pre-trained networks [20,26,77,86] are used to explore spatial features

✉ M. Suresha
sureshm@kuvempu.ac.in

S. Kuppa
kuppa1993@gmail.com

D. S. Raghukumar
rg12.clk@gmail.com

[1] Department of Computer Science, Kuvempu University, Shimoga, Karnataka 577451, India

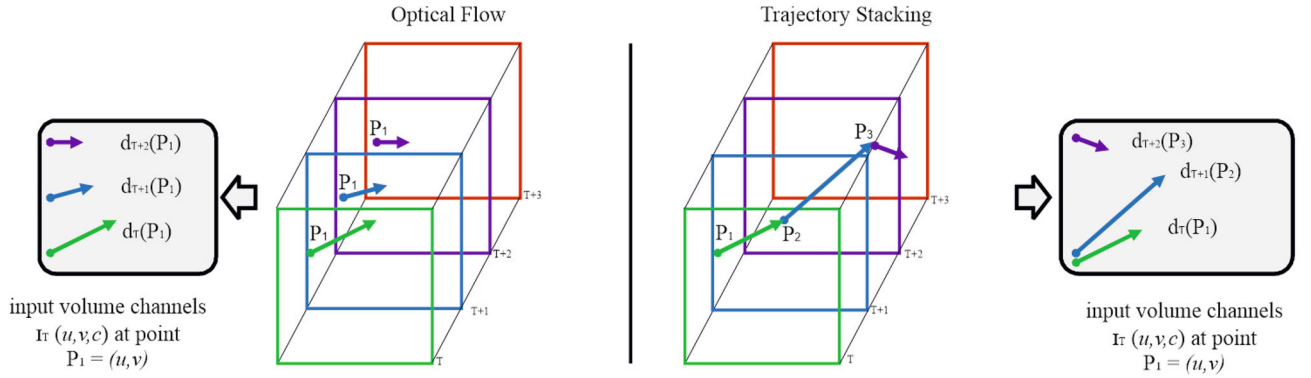

Fig. 1 Representing the motion by using optical flow and dense trajectory stacking

and FlowNet [31,32] networks are used to explore temporal features. FlowNets are derived from a simple CNN model; it explores optical flow frames from raw temporal video data and provides plenty of features for video understanding.

State-of-the-art deep learning models are used to explore important features for video understanding. Various techniques are used to explore these features, such as fusion-based two-stream models [78] and hybrid multi-models [33,101,105]. In two-stream networks, 2-CNN pipelines are used to model spatiotemporal patterns with stacked optical flow frames as input to two consecutive spatiotemporal streams. This two-stream CNN model is extended to exploit long-term motion cues using a recurrent neural network (RNN) with a long short-term memory (LSTM) layer [19,29,66,79] in the temporal domain. Multi-model hybrid frameworks exploit spatiotemporal patterns via hierarchical approaches. In this case, features are fused at various levels: Jiang et al.'s [33] regularized features are fused at video level and sequential LSTM and CNN multi-models fuse the long-term and short-term motion cues at the top level (Fig. 5) by using stacked optical flow and spatial frames as input. From these observations, we are motivated to address the challenges in and develop approaches for extracting spatial and temporal cues via two strategies. The first strategy is to apply two types of techniques for exploring spatiotemporal cues: classical vision approaches will be used to identify spatiotemporal patterns, and modern machine vision approaches will be used to identify the same patterns. The second strategy is to use a hybrid framework approach that is based on feature fusion techniques. In the hybrid framework approach, state-of-the-art feature fusion models act on the spatiotemporal patterns.

To the best of our knowledge, this is the first time an effort has been made to study the various approaches for the extraction of spatial and temporal patterns. Therefore, it focuses on deep learning models that extract in-depth features by using state-of-the-art modalities. The main contributions of this survey are to discuss the current trends, major strengths, weaknesses, and challenges of the following:

1. Classical versus modern machine vision techniques for spatiotemporal cue modelling.
2. State-of-the-art deep learning approaches for extracting spatiotemporal patterns for video understanding (including two-stream and hybrid multi-models).
3. Context in real-world problems.

The remainder of the paper is organized as follows: Sect. 1.1 presents the related work. Section 2 describes the classical and modern machine vision technologies for extracting in-depth video features in detail. Section 3 discusses the state-of-the-art models that are used to evaluate in-depth features with various modalities. Section 4 describes societal, industrial, and research problems. Section 5 explores the current and future research plans. Section 6 summarizes the strengths and trends in this research area. The challenges and weakness in the extraction of spatiotemporal patterns and the challenges that are encountered with deep video understanding models are discussed in Sect. 7. Finally, the conclusions of this work are presented in Sect. 8.

1.1 Literature review

Video understanding is a longstanding active research area in multimedia and computer vision. It addresses various problems, such as action recognition, video classification, and captioning. Successful systems depend heavily on video cues, and most recent work concentrated mainly on the design of robust and discriminating descriptors. Therefore, the substantial advances in the spatial domain, which have been extended to time period video data, namely temporal data, motivate many video representation methods. The modelling of spatiotemporal cues is an active research area in video understanding. Lowe et al. [52] and [108] utilized classical

image-based descriptors such as SIFT to capture the local still image spatial cues in videos. Local spatial cues for visual features and global visual cues such as motion are also highly important. For modelling these cues, a simplified feature extraction approach was introduced, namely dense trajectory [94], which densely sampled the local temporal patches with time and evaluated multiple features based on trajectories.

In conventional handcrafted features, the performance is limited by the processing of natural data in raw form [44]. Rapid progress has been made in learning robust features from raw data with the advances in deep neural networks (DNNs). In [31,32], state-of-the-art deep networks for evaluating the optical flow from raw video data were introduced. The simple CNN architecture directly learns to explore the optical flow from raw video data. Instead of using this approach, Qiu et al. [67] evaluated the dependencies between spatiotemporal patterns by using both CNN and RNN. Among deep feature extraction models with various neural structures, convolutional neural networks (CNNs) have realized substantial success on various problems, such as image classification [71], image object localization [61] and speech recognition [22]. In video understanding, Sun et al. [106] and Karpathy et al. [37] applied CNN on the temporal domain with stacked video frames over time and improved this work [3–6] by introducing two-stream CNN approaches, which use two CNNs with static images in a spatial net and optical flow images in a temporal net to acquire the spatial and motion patterns, respectively. Using these patterns, CNN-based features are applied to classical and modern machine vision approaches to solve various problems in video understanding.

These two-stream CNNs [18,66,78,109] were not used to explore long-term motion cues of videos because they focus only on short-term motion: they use the pre-computed stacked optical flow frames as input; hence, this type of optical flow is discarded [101] by the learning process. In addition, they are insufficient for highly complex semantic contents [11,23] and can only be effectively used to identify temporally ordered short-term actions. To address these limitations, Wu et al. [105] proposed a hybrid multi-model deep learning framework that can learn feature relationships via the common two-stream CNN-based approach and utilize a regularized feature fusion network at the video level. It can model not only the short-term motion patterns but also the long-term motion patterns using the recurrent neural network (RNN) approach with a long short-term memory (LSTM) layer that accepts video sequences as inputs.

According to a survey [99,106], multi-model frameworks fuse local and global features. The CNN calculates a fusion score based on spatial and long-term optical flow cues, which provides promising results and improves the accuracy of video understanding. Tremendous broad endeavours have been made in the study of human action and complex events. However, the works that are discussed above are not reliable; they focus only on combining several descriptors via normal fusion strategies and ignore the inter-feature and inter-class semantic relationships. In [34,104], a regularized deep neural network (rDNN) was modelled that both learns the feature relationships and explores the class relationships. Most of these works used regularized feature fusion strategies such as early fusion and late fusion, which use spatial and optical flow correlations as inputs to a CNN and are combined with the traditional visual features, such as dense trajectories. These feature-based multi-class cues are essential inputs into deep neural networks (DNNs).

After the fusion of the CNN features, the modelling of the interrelationships among spatiotemporal structures at multiple abstract levels remains ambiguous. First, the spatial stream fails if two frames share the same background. Second, two actions can be confused in a short clip of the temporal stream but distinguished in a long clip. For instance, consider Pull-ups and Rope Climbing in the UCF101 data set. These two classes could be moving in the same direction in a short clip. However, if we expand the time period, it can be easily determined that the man in Pull-ups is moving up and down, whereas the man in Rope Climbing is moving straight upwards. For solving this problem, Wang et al. [101] proposed a spatiotemporal pyramid network that utilizes a hierarchical fusion strategy to distinguish actions and to support the modelling of long-term temporal fusion and a visual attention mechanism by introducing a spatiotemporal compact bilinear model, as illustrated in Fig. 6. This model combines spatiotemporal features at multiple abstract levels. First, optical flow features are pooled into a spatiotemporal compact bilinear layer over time. Second, the spatiotemporal attention module identifies the salient region and fuses the features from the spatiotemporal compact bilinear model and the attention module. Influenced by the effective presentation of spatiotemporal compact bilinear (STCB) model for long-term optical flows, features are modelled without loss of abstract information. In [84], Sun et al. applied this approach [101] to optical flow extraction.

Many video features are explored in state-of-the-art problems such as image classification, semantic segmentation, object detection, action recognition, simultaneous localization and mapping (SLAM), object tracking, and video classification, captioning and understanding. The commonly considered video features are space–time, spatiotemporal, global motion, dense motion, and optical flow. These visual features are modelled via various approaches, such as classical and modern machine vision techniques (as listed in Table 1), and this modelling process is challenging. Via this approach, various sets of methods have been developed by the multimedia community. First, in classical vision research, shallow encoding techniques such as HOG and HOF are used to explore high-dimensional

Table 1 Spatiotemporal feature modelling methods

Classical vision	Modern machine vision
Histogram of gradient (HOG)	Convolutional neural network (CNN)
Histogram of optical flow (HOF)	Recurrent neural network (RNN)
Space–time interest points (STIP)	Long short-term memory (LSTM)
Scale-invariant feature transform (SIFT)	Generative adversarial neural network (GAN)
Dense trajectory	Regularized feature fusion models

space–time interest points (STIPs) [52,108], which are used to model the spatiotemporal and dynamic motion patterns, and these features are pooled into a bag of features (BoF) representation and combined with an SVM classifier. Instead of using these local video features to model the motion, the model could utilize dense point trajectories [5,63,94] for depth estimation of motion patterns and could densely sample local patches from each frame at various scales for computation of the motion from the optical flow. The motion boundary histogram (MBH), which is a gradient-based feature that is computed separately from the vertical and horizontal components of the optical flow, realizes the same performance as the trajectory-based approach.

Furthermore, to improve the trajectory-based approach, a camera is used, or the global motion is calculated using a variety of benchmarks and coupled with quantization techniques such as Fisher vector encoding. However, these representations only focus on modelling local motion patterns within short time snippets. These feature encoding methods discard the temporal motion information of video sequences.

In the modelling of spatiotemporal features by using classical vision techniques, various challenges are encountered. The classical vision models that are discussed above are not efficient or effective in modelling rich spatiotemporal information; for example, the improved version of the dense trajectory (IDT) [94] incurs a high computational cost and does not substantially outperform other methods in exploring expected abstract features. Modelling these features using modern machine vision approaches yields superior results compared to classical vision approaches. Modern machine approaches use familiar deep neural networks, such as convolutional neural networks (CNNs) [75] and recurrent neural networks (RNNs) [19]. The CNN input of the model is formed by stacking optical flow displacement fields between several consecutive frames. Figure 1 illustrates two motion representation methods: trajectory stacking and optical flow stacking. CNN explicitly describes the motion between video frames, which facilitates recognition as the network need not estimate the motion implicitly. In addition, from a base set of configurations, optical flow stacking, bidirectional optical flow and mean flow subtraction methods can be evaluated, whereas motion stream CNN explores the short-term motion patterns only. Due to many computations for long-term motions, the initial input to subsequent layers becomes negligible, which results in the vanishing gradient problem in the CNN. To address this problem by using the recurrent neural network (RNN) model, the long short-term memory (LSTM) [29,66,79] network utilizes the information of the previous frame to solve for the uninterrupted gradient flow, which facilitates back-propagation, increases the stability, prevents the occurrence of vanishing and exploding gradients in the model and identifies the long-term motion patterns efficiently. However, the modelling of the dynamic movement patterns without losing motion movements that correspond to the spatial features is essential; hence, the dynamic motion patterns are evaluated by extracting more abstract features from the spatial and temporal patterns. Then, it correlates short-term, long-term and spatial patterns by combining network layers such as CNN and RNN-LSTM and using various fusion strategies at various levels. Via this approach, well-known models are derived, such as two-stream convolutional neural networks [78], hybrid multi-models [105,116] and regularized feature fusion models [33] at the video level. The core mechanism of these models is the mutual extraction of spatiotemporal features, which are input into various levels. The coupled features are transferred to multiple hierarchical network architectures. Then, abstract features such as long-term and short-term temporal motion dynamics are modelled to obtain next-level predictions. However, these multi-models and fusion techniques do not jointly learn both spatial and temporal features, and CNN+LSTM was not resolved well in the temporal dynamic modelling problem. In [29,66,77], a recurrent neural network with long short-term memory (LSTM) is used to model the temporal evolution. The RNN structure facilitates the exploration of temporal dynamics from dense frames with consideration of the time order. However, it has only realized similar performance to temporal pooling [78]. This might be attributed to the training difficulty and the occurrence of gradient vanishing for long videos. To address these limitations and the previously discussed problems that are encountered when using spatiotemporal mutual attention models, such as actions sharing a similar background and action confusion for short snippets, in

the modelling of pyramid architectures [102], temporal segment networks (TSNs) [29] and temporal–spatial mapping (TSM), the network jointly learns spatiotemporal feature representations and can explore the information of dense frame features.

Based on this literature review, we discuss the evolution of spatiotemporal cues from an image domain to the temporal domain in video data and discuss various traditional and deep learning methods that are used to extract spatiotemporal cues. In addition, we study the importance of spatiotemporal cues and the performances of various fusion strategies for video understanding. Table 2 presents an overview of the features that are used in the state-of-the-art deep learning models and deep neural networks (DNNs) that are used to evaluate the video understanding task and its result. Then, we review various deep learning techniques for exploring spatiotemporal patterns, identify their limitations for modelling long-term motions and discuss how to overcome these challenges by using long short-term memory (LSTM) and hybrid deep learning frameworks.

2 Spatiotemporal features

Spatiotemporal feature exploration is of paramount importance for video understanding. The spatiotemporal features are motivated by the two-class hypothesis findings in neuroscience, as indicated by the visual cortex of the human brain, which contains two pathways. The ventral stream performs object acknowledgement (spatial), and the dorsal stream perceives the movement (temporal) [21,40]. From this information, in the early stage, handcrafted representations have been explored using popular feature descriptors such as space–time interest points (STIPs) [42], SIFT-3D [65], spatiotemporal SIFT [47], HOG and 3D histogram of gradients [62] and spatial cues such as texture have been created by other representations, such as histogram of optical flow (HOF). A successful handcrafted feature is dense trajectory [94], and its updated version [63] focuses on motion information. However, the available handcrafted feature exploration methods [7] are not suitable for this task. Recently, deep neural network models have shown substantial potential for the derivation of robust features from raw data on various tasks, such as image classification and object detection. The design of a powerful feature representation method is an important topic in the multimedia community; hence, researchers have attempted to apply deep learning techniques to the video domain. Here, we are categorizing two perspectives for extracting spatiotemporal cues: first, classical vision and modern machine vision approaches for extracting spatial cues; second, temporal optical flow features extracted from classical vision and modern machine vision approaches.

2.1 Spatial domain

Spatial feature extraction is performed on still video frames and describes the relative spatial area of an object and its spatial semantic relationships with other objects in each video frame. It is highly useful in differentiating the static and dynamic contexts. In the dynamic context, a moving object performs an activity while moving along the field of view, and in the static context, a moving object performs an activity while remaining in place. The spatial information includes the position of the moving object; thus, in considering spatial information regarding the activity of a moving object, one is expected to differentiate between the static and dynamic contexts. Here, we study two approaches for spatial feature extraction.

2.1.1 Classical vision

In classical feature extraction, [42,47,62,65] defined various handcrafted spatial feature extraction techniques. These methods capture frame-wise spatial cues using space–time interest point detectors, such as corner detectors, edge detectors, the Harris detector and the Hessian detector, or sampling methods, such as motion-based adaptive sampling and dense sampling for each frame.

2.1.2 Modern machine vision

Deep neural networks (DNNs) have largely driven the advances in image recognition, object detection, and classification, which outperform all large families of classical vision techniques on image data and have been extended to video data. Convolutional neural networks (CNNs) in particular yield promising results. For instance, a simple extension is to stack multiple frames over time as inputs to a CNN for spatial feature learning. Otherwise, pre-trained networks are used to exploit the many pre-trained annotated images in the spatial network, such as images from the ImageNet data set [20] and other data sets [26,77,86]. Using these spatial features, spatial domain problems are easily evaluated via various transfer learning and feature fusion techniques.

2.2 Temporal domain

Video semantics typically do not arise in segregation, and a class of interest can be conveniently recognized from its semantic contextual relationship (see Sect. 1.1 5th Pa.). The temporal features recognize the action or movement of an object from motion in the form of an optical flow [3]. An optical flow is a motion pattern of an object in stacked

Table 2 Overview of state-of-the-art deep learning models and different features used for video understanding

Features	Model	Method	Data set	Task	Result
Spatial–temporal optical flow and dense trajectory	3DCNN and pooling [91]	CNN	Sports1M	Action recognition	85.2
			UCF101	Action recognition	85.2
			ASLAN	Labelling of action similarity	78.3
			YUPENN	Scene classification	98.1
			UMD	Scene classification	87.7
			Object	Object recognition	22.3
Spatial and stacked motion optical flow (long-term temporal)	Regularized feature fusion [33,105]	CNN and LSTM	UCF-101	Video classification	91.3
			CCV	Video classification	83.5
Spatiotemporal geometric and kinematic signals	Part-based graph convolutions [89]	CNN and LSTM	NTURGB+D	Action recognition	87.5
			HDM05	Action recognition	88.17
Spatial and temporal	Time information fusion [37]	CNN	UCF-101	Action recognition	65.4
			Sports-1M	Action recognition	80.2
Spatial, optical flow, motion vector and residual	Discriminative motion cue [76]	CNN-ResNet	Kinetics	Action recognition	65.42
			UCF-101	Action recognition	71.8
			HMDB-51	Action recognition	92.3
Spatial and optical flow	Two-stream fusion and inception [103]	I3D CNN	LRW	Action recognition (Lip reading)	82.11
Spatial and optical flow	Spatiotemporal compact bilinear (STCB) [101]	CNN and LSTM	UCF-101	Action recognition	94.16
			HMDB-51	Action recognition	68.9
Spatial and optical flow	Two-stream feature extraction and Self-attentive action classification [117]	CNN-ResNet	THUMOS14	Action recognition	55.9
			ActivityNet1.3	Action detection	33.1
Spatial and temporal super-resolution	Two-stream and super-resolution [114]	CNN	UCF-101	Action recognition	87.58
			HMDB-51	Action recognition	65.82
Spatial and optical flow	Spatial-temporal attention and Static motion collaborative [66]	CNN-ResNet50	THUMOS-14	Video classification	84.7
			UCF-101	Video classification	94
			UCF-50	Video classification	95.7
			HMDB-51	Video classification	68.7
Spatiotemporal	Deep spatiotemporal fully Convolutional network [67]	2D FCN and ConvLSTM	CamVid	Video semantic	30.2
			A2D	segmentation	68
Stacked spatial frames	Deep bidirectional [93]	CNN and LSTM	UCF-101	Action recognition	91.21
			HMDB-51	Action recognition	87.64
			YouTube	Action recognition	92.84
Spatial-temporal and trajectory features	Regularized DNN [34]	CNN	Hollywood2	Video classification	65.9
			CCV	Video classification	72.9
			FCVID	Video classification	75.4

video frames, and it is applied in structuring the motion, video compression and stabilization. The modelling of the short-term and long-term temporal optical flow cues for video understanding is a crucial task; hence, the motion features are estimated via various metrics based on the optical flow of stacked video frames for the exploration of optical flow features. The multimedia community has derived various classical descriptors, such as dense trajectory, HOF [100] and its extended version, namely histogram of oriented optical flow (HOOF) [9]. To explore the long-term optical flow features, graphical models such as Bayesian networks (BNs), hidden Markov models (HMMs) and conditional random fields (CRFs) are popularly used. In the deep learning era, various robust models have been developed for exploring optical flow features using modern machine vision approaches such as CNN and RNN. In a deep neural network, CNN has been successfully applied for optical flow estimation. Various available deep neural networks push the envelope in terms of accuracy by extracting optical flow features effectively. These networks include FlowNet2 [32], LiteFlowNet [31] and EV-FlowNet [119]. Here, we study both classical and modern machine vision techniques that are used for temporal feature extraction.

2.2.1 Classical vision

In classical vision, the temporal feature extraction methods are divided into optical flow and dense optical flow methods. An optical flow is a motion pattern of an object in consecutive video frames. An optical flow operates under two basic assumptions. First, the pixel intensity of the object is constant between the consecutive stacked frames. Second, adjacent pixels' motions are similar.

For instance, a pixel $I(x, y, t)$ in the first frame moves by distance *(dx,dy)* in the next frame after a time interval of dt. The object pixels are of equal intensity among frames and is as expressed in Eq. 1:

$$I(x, y, t) = I(x + d_x, y + d_y, t + dt). \tag{1}$$

A Taylor series approximation is applied in the RHS, similar terms are removed, and both sides are divided by dt to yield the following equation:

$$f_x u + f_y v + f_t = 0, \tag{2}$$

where

$$f_x = \frac{\partial f}{\partial x}; f_y = \frac{\partial f}{\partial y} \tag{3}$$

$$u = \frac{dx}{dt}; v = \frac{dy}{dt}. \tag{4}$$

Equation 2 is an equation for computing the optical flow, which depends on the image gradients *fx* and *fy* and time gradient *ft*, in which *(u, v)* is unknown. It is impossible to solve a single equation with two unknown factors. Various methods can be used to address this problem. One of the methods is the Lucas–Kanade (LK) method. Based on the second assumption, the LK method considers *3X3* chunk around the points with similar motion:

$$(f_x, f_y, f_t). \tag{5}$$

Then, solve

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i {f_{x_i}}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i {f_{y_i}}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix}. \tag{6}$$

Equation 6 is used to calculate the optical flow, along with the scale. It computes only the optical flow for a sparse feature set. The dense optical flow provides the optical flow for the entire region in the frame. Inspired by this, a dense trajectory is evaluated [63,94] by computing dense points from every frame and the motion-based optical flow field is computed from the displacement vector.

2.2.2 Modern machine vision

In modern machine vision, various FlowNets are used for optical flow estimation. State-of-the-art traditional methods are also applied. FlowNets mainly focus on small displacements and natural data. FlowNets have not been utilized in variable procedures. They are mainly applied to massive clustered temporal data using state-of-the-art deep learning models, which are commonly known as convolutional neural networks (CNNs), for the evaluation of optical flows. Ilg et al. [32] proposed FlowNet 2.0, which learns the optical flow directly from raw data. This leads to improved performance on real-world problems of video understanding. Figure 2 presents a schematic diagram of the FlowNet networks. It presents three primary outcomes: first, it focuses on training data and presenting data during training; second, a stacked architecture, namely motion stacked difference image (MSDI), is developed [96] for warping the next image with a transitional optical flow; and third, a small group of displacements are introduced into an additional intra-network model for the exploration of small-scale motions. Hence, it realizes additional improvements in terms of image quality and performance and is on par with the previously described state-of-the-art methods while maintaining reiterative frame rates. Hui et al. proposed an updated version, namely LiteFlowNet [31], in which higher speed of the network is realized by reducing the network model size, and a pyramidal architecture is implemented to avoid cue loss by generating pixel-level flow to the sub-pixel level.

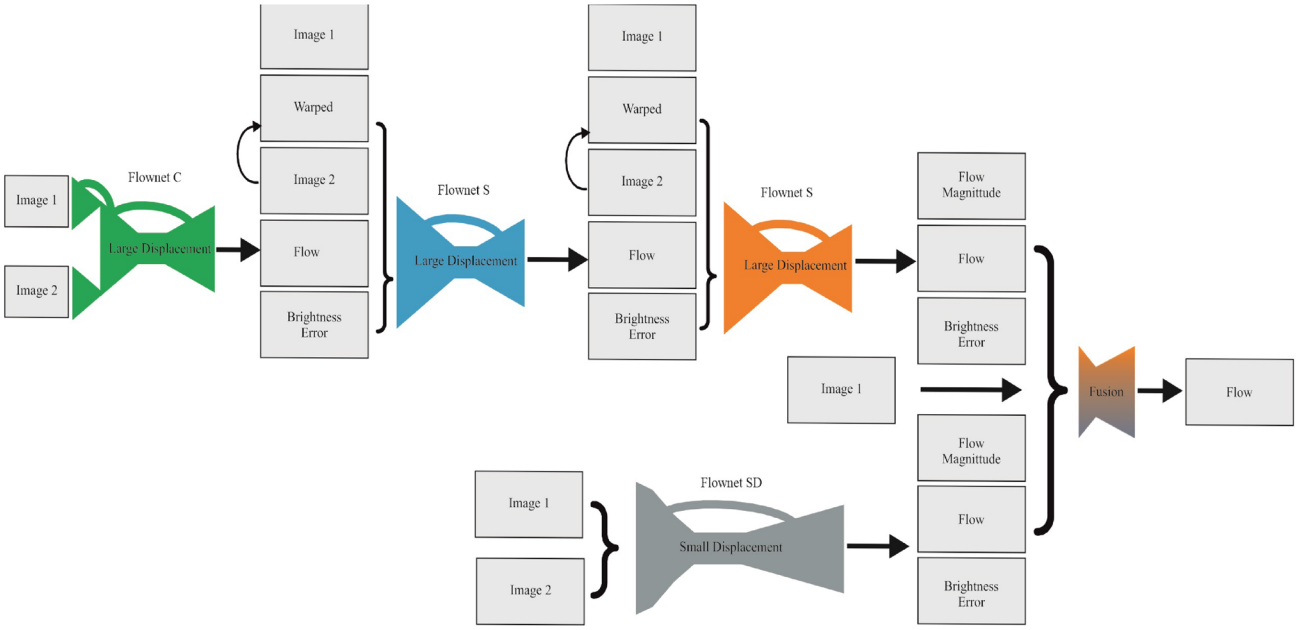

Fig. 2 Schematic diagram of the FlowNet architecture for computing the large-displacement optical flow with the combination of multiple FlowNets (braces denote concatenations of inputs)

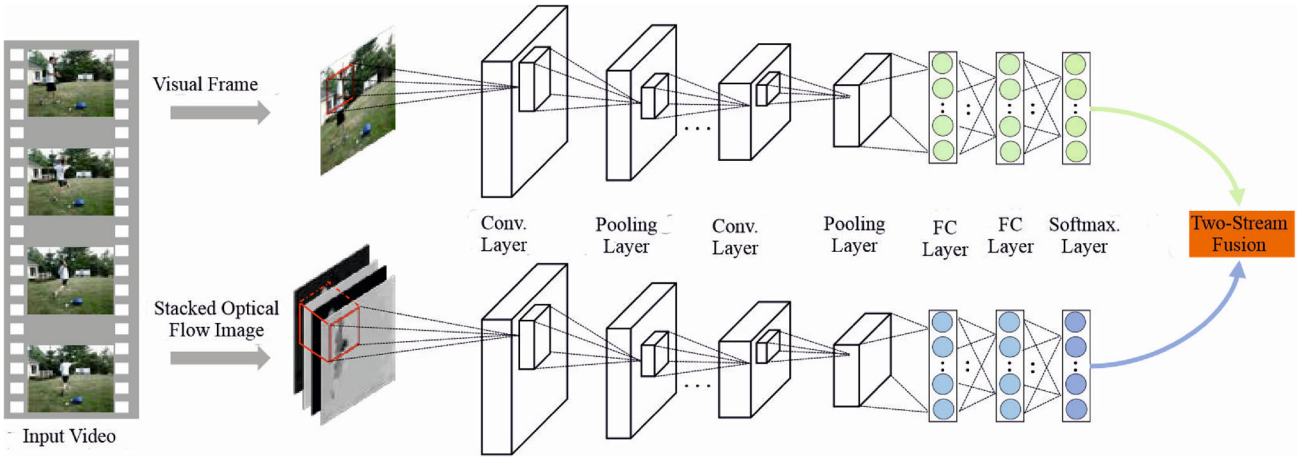

Fig. 3 Pipeline architecture of two-stream CNN

Using these networks, various video understanding problems can be addressed, such as motion segmentation, video classification and captioning, action recognition, simultaneous localization aggregation and mapping (SLAM) and 3D reconstruction.

3 Proposed models and methods for feature extraction

Alternative class models and techniques have been proposed in deep video understanding for exploring video signs, which depend on various neural designs such as two-stream and multi-model hybrid systems. Two-stream systems primarily focus on the combination of procedures on spatiotemporal features from stacked optical flows and still frames. Hybrid multi-models concentrate on abstract features such as long-term motion cues.

3.1 Two-stream approach

According to findings in neuroscience [21,40] (see Sect. 3), the two-stream neural network was developed using spatial–temporal features and fusion strategies; Fig. 3 presents an outline of the two-stream CNN. First, Simonyan et al. [78] implemented the two-stream CNN for action recognition and stacked dense optical flow frames for motion estimation. The proposed network contains two CNN phases: first, spa-

tial features are extracted from still video frames; second, motion cues are extracted from stacked optical flow frames. Each stream's scores are combined via score fusion techniques. This network is extended by improving the score fusion technique [18] at the convolutional layers without the loss of abstract features instead of introducing a softmax layer by implementing temporal and spatial fusion at an early stage; this network is also used for video classification [109] tasks. In [29], T-CNN was proposed, which extracts activity-based 3D convolutional features by dividing a video clip into segments of equal length; it is regarded as a tube proposal (T-CNN). T-CNN employs spatiotemporal movement detection and network flow. The optical-flow-guided weighted mean-squared-error loss for the spatially oriented SR (SoSR) network has been proposed for improving the super-resolution (SR) [114] in the spatial part. To avoid temporal disjoints between frames, the Siamese network is used for temporally oriented SR (ToSR), which highlights the temporal stability between successive frames. Spatiotemporal feature fusion realizes excellent performance in video understanding because it rapidly updates with new feature sets. Wang et al. [93] proposed a CNN for describing an additional feature class, namely global accumulated motion patterns; this approach is named the motion stacked different image (MSDI) approach. The network model contains three input streams, namely spatial, local temporal and global temporal streams, and is called the three-stream CNN model.

In the two-stream architecture, modelling a spatial cue based on CNN is a simple task and is more efficient because the displacement values correspond to the moving scene points at the same spatial position across several frames. The modelling of temporal cues focuses only on short-term motion that is computed in very short frame windows. Furthermore, the network is improved by modelling a long-term temporal cue that was introduced in the recurrent neural network (RNN) model. The long short-term memory (LSTM) layers [83,118] model the temporal gestures within and between the actions using bidirectional LSTM, followed by CNN. Figure 4 illustrates the CNN and LSTM architecture. Long-term cues and maps of the spatial patterns and their temporal relationships are explored via fusion of the outputs from both CNN and LSTM. A deep fusion framework [19] exploits various fusion techniques in CNN spatial features and LSTM temporal dynamics patterns. DNN captures the still and motion information separately, and two main limitations are encountered. First, spatiotemporal attention coexistence relationships are avoided. Second, strong static and motion data complementarity coexist in the video. To address these problems, Peng et al. [66] introduced two approaches: the spatiotemporal attention model, which extracts discriminative features by jointly modelling spatiotemporal cues in video, and the static motion collaborative model, which collaboratively trains both features.

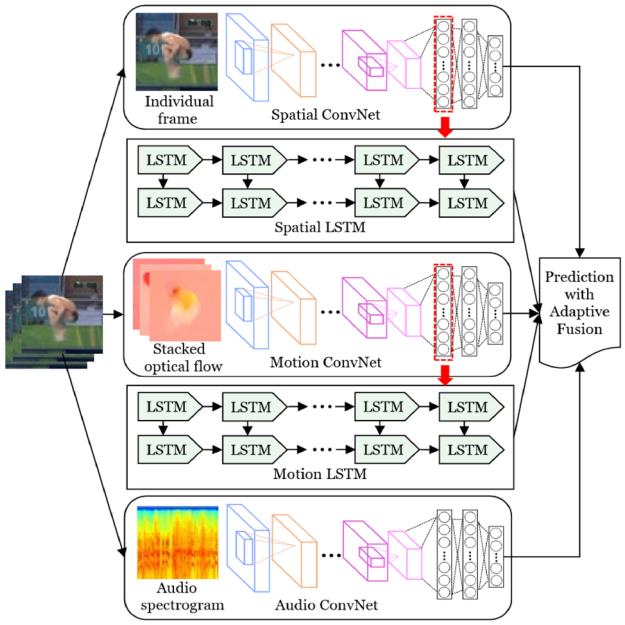

Fig. 4 Two-stream network with long short-term memory (LSTM)

3.2 Hybrid multi-model approach

The previous works that are discussed above have a significant limitation in the modelling of the long-term cues. To address this limitation, the multimedia community introduced hybrid deep learning frameworks for video understanding. In [33,105], a framework for harnessing spatial, short-term and long-term cues was proposed. Additionally, the spatial cues and short-term motion cues were combined using regularized fusion networks and sequence-based LSTM as a classification strategy. Figure 5 presents an overview of the multi-model framework. Spatial and short-term features are extracted from the two-stream CNN and input into the long short-term memory (LSTM) layer for long-term feature modelling. Local spatial and global motion features at the video level are extracted and pooled by a regularized feature fusion network [102]. Afterwards, all outputs from the feature fusion network and sequence-based LSTM are fused to yield the final predictions.

The key strategy is to fuse the spatial and temporal features, which has the following limitations: if two videos share similar backgrounds at the spatial levels and for short snippets of motion in the temporal region, the modelling of the spatiotemporal cues with multi-level abstractions is uncertain. Some CNN frameworks are not specifically designed for video understanding, and it is difficult to fully exploit spatiotemporal features. To address this problem, Wang et al. [101] introduced numerous abstraction levels for spatiotemporal features, and a pyramid architecture network fuses the spatiotemporal features hierarchically. It learns huge global video cues by using multiple-path temporal sub-networks,

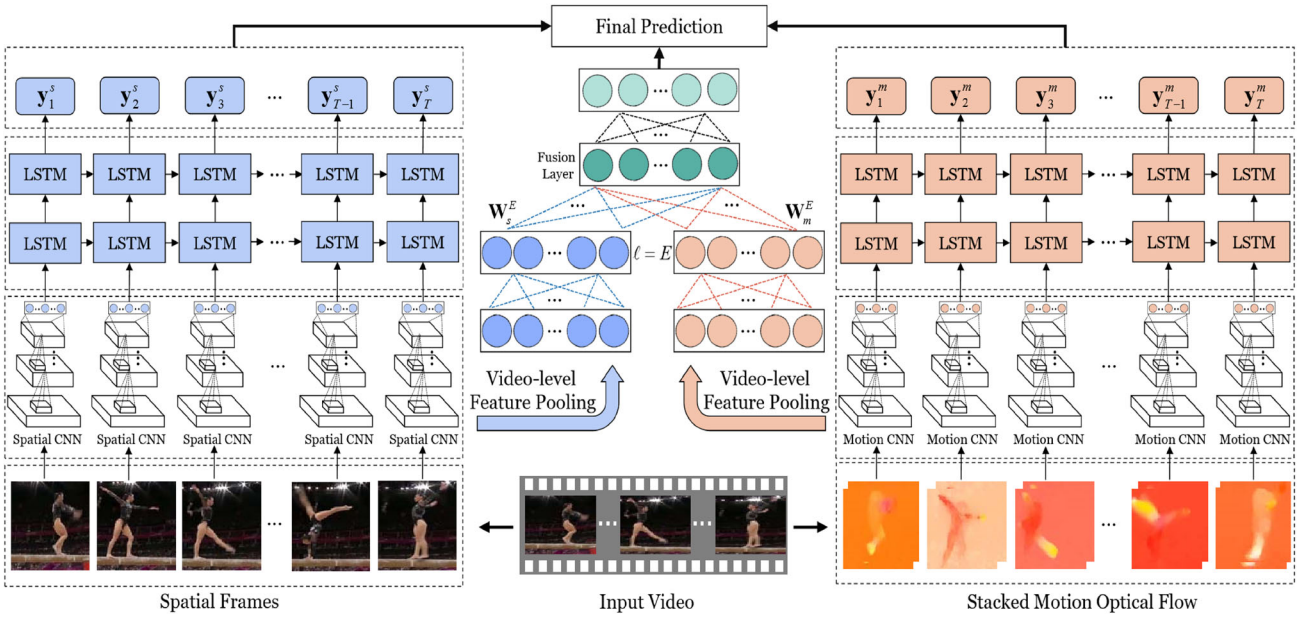

Fig. 5 Overview of the multi-model framework

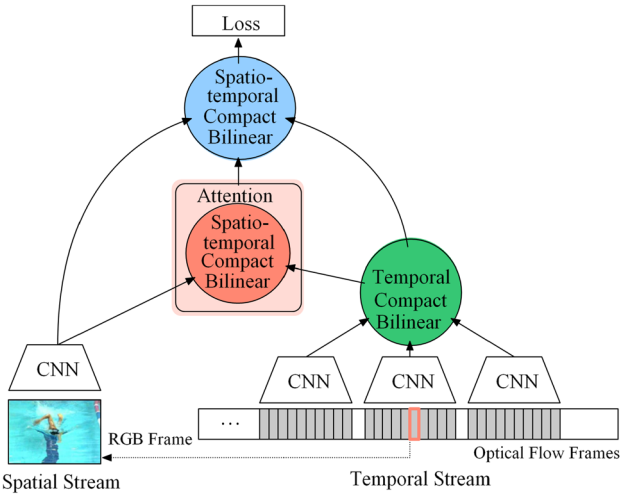

Fig. 6 Overview of the spatiotemporal pyramid network

which are used to test optical flow frames in a long-term sequence, and it explores multiple fusion methods for the fusion of the temporal patterns effectively. A satisfactory fusion technique should maximally preserve the spatiotemporal cues while maximizing their interaction. Here, we introduce a new spatiotemporal compact bilinear (STCB) model and a spatiotemporal attention model for the preservation of spatiotemporal cues. In STCB fusion (Fig. 6), the spatiotemporal information is fused via the element-wise sum and concatenation method. Bilinear fusion enables all the spatiotemporal cues in various dimensions to be iterated together via a multiplicative approach.

4 Real-world problems

A solution that addresses various sets of problems in diverse domains is called a standard solution. Video understanding is used to resolve various sets of problems in, for example, industry, academics, society and research, which are discussed here in detail.

In human intelligence and sensitive systems, visual understanding is extensive; however, there is no satisfactory replacement for this system. By solving this video understanding problem, deep neural networks resolve various sets of problems for humankind. Well-known problems from the societal and industrial perspectives include the advancement in security systems [111] and autonomous computing [24], among others. For example, in the security domain, the tracing of the security event alerts during an incident or even pre-incident alerts is attractive in security applications, and the identification of suspicious activity [92] at various levels in different scenarios facilitates crime prevention [41] and the realization of the smart and secure city [12] concept. In addition, severe problems in the real world could be addressed in elder care [49], such as detecting if someone has fallen or has taken his or her medicine, and devices that help blind people could be developed.

One of the major advances in public security surveillance systems is the enhancement of autonomous surveillance systems [27]. Intelligent cameras are currently operating in single-scenario restricted platforms to understand the entire video content in a different scenario. This approach will be extended to all scenarios by using transfer learning techniques, and various problems will be solved in the

surveillance domain, such as abnormality detection [8] or abnormal behaviours in the workplace, accident prevention in highways [41], anomalous activity in public places [60] and crowded scene understanding [82]. In traffic management, traditional intelligent analysis systems are highly accurate in classifying vehicles or non-vehicles by using simple attributes; however, video understanding systems annotate and classify objects such as human and intelligent traffic signs, which facilitates the automation of traffic management systems. In sports, decisions are taken in unique cases by using deep spatial-temporal features.

Climate change [15] has become an active area of research; AI could help fight climate change [70] and save our planet and humanity from imminent peril. Computer vision and deep learning techniques have made valuable contributions to climate science research. Spatiotemporal models capture real-time data from the environment, and the data are processed via video understanding techniques. For instance, many countries around the world have little or no information on their energy utilization or greenhouse gas emissions, which is a significant hurdle in the implementation and design of mitigation procedures. Spatial–temporal models can extract footprints of buildings and their characteristics from satellite imaginary, which are input into deep learning algorithms that can evaluate city energy utilization. The same models could also identify buildings that should improve their efficiency. Via this type of deep vision approach, more accurate estimation of energy consumption is possible. Another example is the tracking of deforestation. Deforestation produces approximately 10% of the global greenhouse gas emissions. Prevention and tracking are typically tedious tasks and are conducted on the ground. Satellite imaginary understanding using deep vision models facilitates the automatic analysis of the loss of tree cover on a larger scale.

Video understanding delivers promising performance on challenges in industry 4.0. It contributes to the solution of the most familiar problems, such as autonomous vehicles or self-driving cars [90], robotics [46] or manipulation, targeted advertisement using crowd understanding [82], video captioning [58] and film certification processes in the entertainment industry, and various [13] automated vision-based IoT systems or consumer applications. In addition, it is useful in special cases with conditional scenarios such as autonomous invigilation, library and metro station monitoring by using video understanding approaches and intelligent security surveillance systems.

Especially in academic and research areas, it outperforms many types of handcrafted computational methods [52,108] and models in computer vision. One of the major contributions in the theoretical purview is the extension of the stiffer classical vision techniques to modern machine vision techniques for the extraction of abstract multimedia features. Various problems have been addressed via this approach, such as human action recognition (HAR) [48,56,98], video classification [33], video captioning [106] and abnormality detection [36]. Human action recognition has been an actively studied research topic for a long time; deep learning and spatiotemporal models are used to realize HAR via state-of-the-art methods with promising accuracy. From these results, additional categories of applications and research avenues will emerge in the future, for example, virtual assistant systems and vision-based consumer applications that use human actions and gestures. Video classification and captioning are also active topics in this area; the exploration of valuable information from videos based on human requirements is challenging. Because videos contain huge amounts of multidimensional raw data, they are difficult to analyse and categorize. Advanced deep-learning-based video understanding methods are used to overcome this difficulty. Such advances will create new research avenues and provide significant results for the future.

5 Advanced perspective deep spatiotemporal models

In most video understanding systems, two major features were popularly deployed: spatial and temporal cues [14,50,105,115]. In the spatial domain, [37,80,110] neural models are trained on huge annotated data sets [71] or pre-trained networks are utilized, and multiple stacked optical flow images are used for training in the temporal domain [31,32]. Due to this problem, neural models may become computationally expensive because they require long computation times [41,59] for training on many still images and require large storage areas; hence, the development of a useful video understanding model for the end-level users has proven difficult. The major challenges in the deep learning community are to reduce the amount of input data, the number of layers [17] and the required computational power without losing abstract features and accuracy. For reducing the length of a video while maintaining low storage and computational requirements, temporal context is provided. However, this approach only performs well on short contextual problems; in long-term motion estimation problems, one can expect the loss of abstract information. For example, for safer automobile transportation vehicles (self-driving cars), understanding based on short contexts is challenging due to the loss of visual features.

The deep feature extraction techniques that are discussed above and stacked optical flow frames [18,19] are typically input into the temporal domain and static visual frames into the spatial domain. Here, a raw set of video frames are accepted as input, and there is no high-level standard knowledge model for processing a raw set of video frames. The

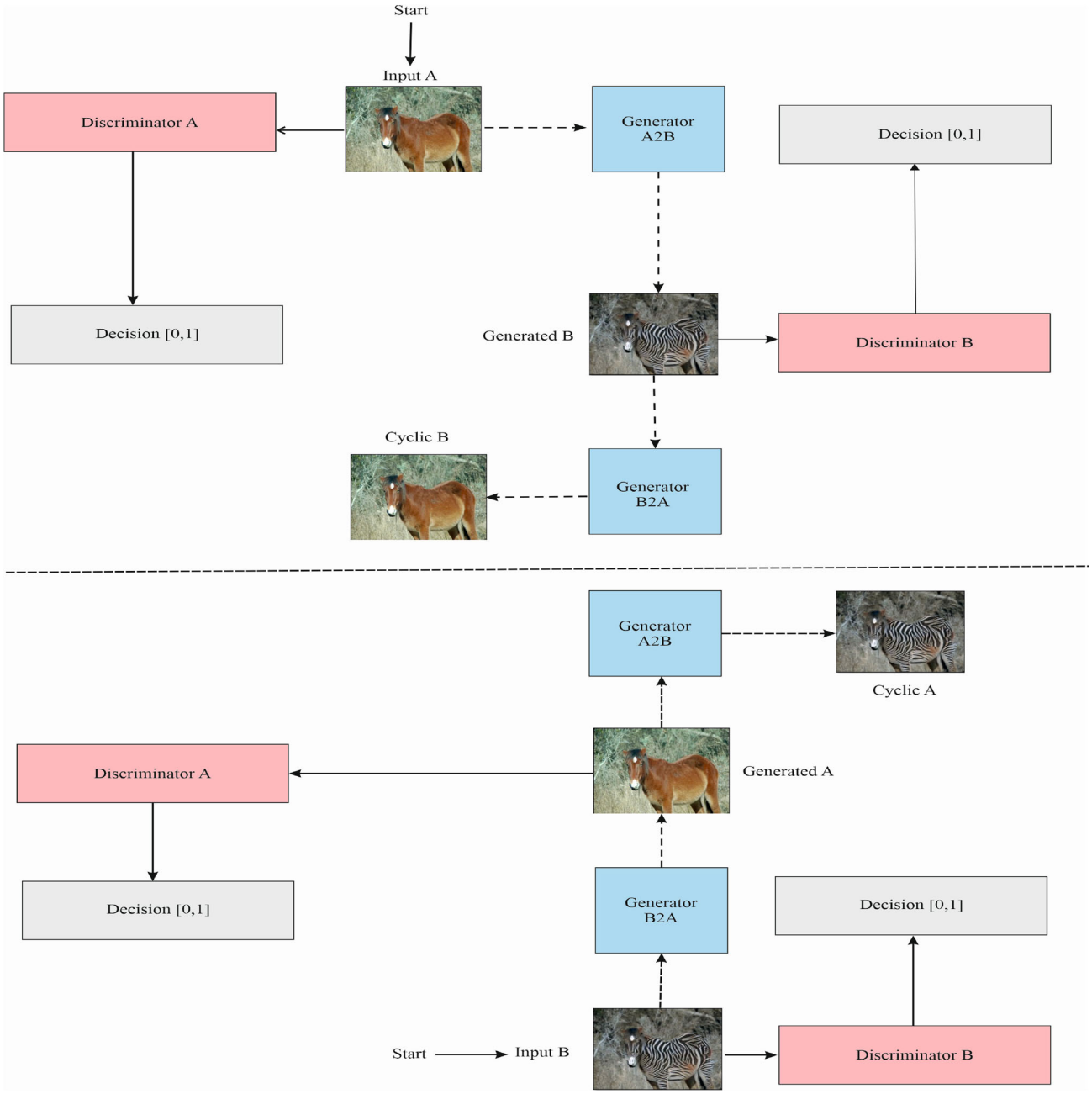

Fig. 7 Image synthesis example from cyclic GAN

requirement for some deep models at the input level avoids unwanted results such as duplicate or redundant frames. In addition, deep models require outcast action extraction on a spatial level and key action removal, key frame extraction and frame synthesis techniques on a temporal level. The improvement in the standard model via the incorporation of the techniques that are discussed above will not only influence the network complexity but also improve the computational performance of the model.

Now, it is possible to design a standard input model for video processing by using the aforementioned techniques, which is an active area of research. In the spatial domain, spatial synthesis techniques are used to clear the temporal moment and to extract suitable cues. Figure 7 illustrates the spatial image synthesis [30,112,120] techniques, which remove the region of interest of an image and extend the temporal image synthesis to sequential frames. For instance, a region of interest in the spatial domain is a spatial relationship between the objects that are used to classify the scenario; however, in many cases, the objects are unclear due to temporal action occlusion, which causes incorrect predictions in scenario classification. To avoid this type of problem, the

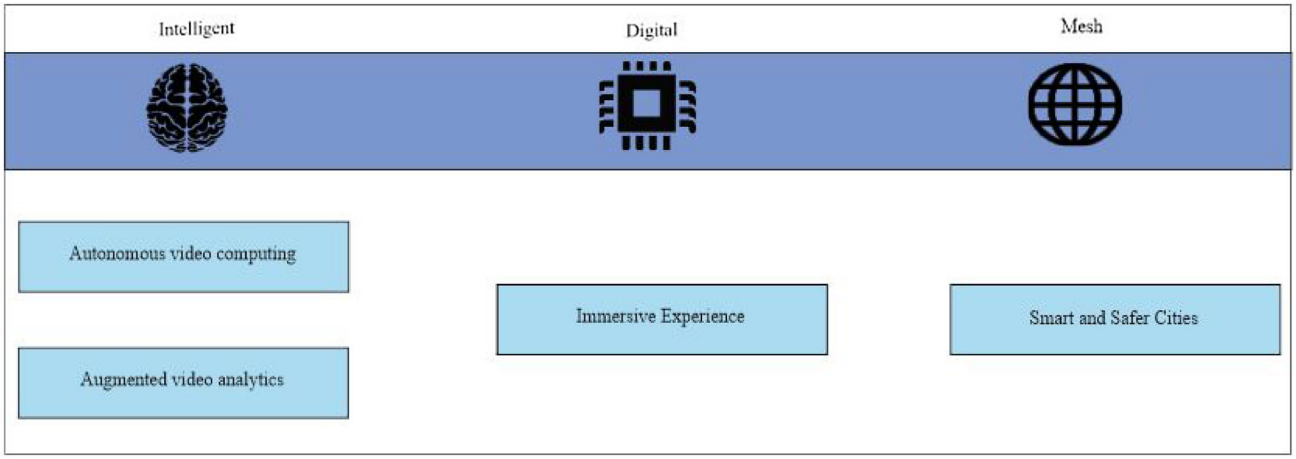

Fig. 8 Strategic trends in video understanding

deep learning community typically uses multiple representations of frames for training in deep networks. This causes the deep computational model to accept a large amount of data as input, increases the number of hidden layers and incurs a huge computational cost. The deep learning community mitigates these shortcomings by extracting the temporal motions from still video frames to reduce the maximum representations at the training level, reduces the computation time by decreasing the number of hidden layers and employs a deep network to improve the computational performance. Training on all the frames of the temporal domain is not required; instead, the best key action frames can be selected using recurrent neural models, which reduces the number of frames. For instance, prior to training the neural networks, the most appropriate targeted action is selected from the stacked optical flows. This can be realized using decision-based recurrent neural network models.

In the future, we must address various problems to enhance the model accuracy and to overcome current deep learning weaknesses. We should aim at satisfying the LLL objective [55]: less data, less layers and less computational power. By introducing super frames, super actions and visual attention, Wang et al. [102] extraction methods could be proposed by using various recurrent neural network (RNN) models and various bandit confidence bound methodologies. For instance, via the super frame extraction method, we can identify the best and worst spatial frames in a long video sequence according to their similarities and temporal motion values. Consider the metro station scenario in the Avenue data set [53], which is classified as a metro station by removing temporal motions, as classification is a difficult task. If this will be resolved, we will expect promising results, and it will be possible to classify the content with less prior knowledge. In super action, the targeted action is identified with the help of previous temporal occurrence frames via the estimation of similarity and dissimilarity information. It helps reproduce the targeted action and reduce the length of the sequence of optical flow frames in the temporal domain. One of the major future advances is transfer learning [87], in which pretrained models are portable to other sets of problems. For instance, a model for an individual who is performing an activity can be transferred to a scenario in which an animal performs a similar activity. This type of advancement, which is known as transfer learning, will be significant for video understanding and artificial intelligence (AI). In the future, we will observe how machines conduct transfer learning. This will be a promising research avenue.

6 Summary of strengths and trends

If someone wants to comprehend why something is happening, motion provides substantial knowledge, and it cannot be captured in a single frame. The practical use of spatiotemporal cues and deep neural networks provides important benefits that may lead to robust approaches for annotating, searching and mining video footage. Here, we discuss major strengths and trends in this area of research. The strengths are summarized in terms of feature extraction and feature mechanisms, unstructured data, unsupervised learning and its accuracy level (Fig. 8).

6.1 Strengths

Deep-learning-based video understanding has several core strengths compared to classical vision approaches: it can be applied to various areas of research, it can be used to solve a unique set of research problems, and it enhances several applications.

6.1.1 Feature engineering

In video understanding, one of the major strengths is the use of deep-learning-based spatiotemporal models and feature

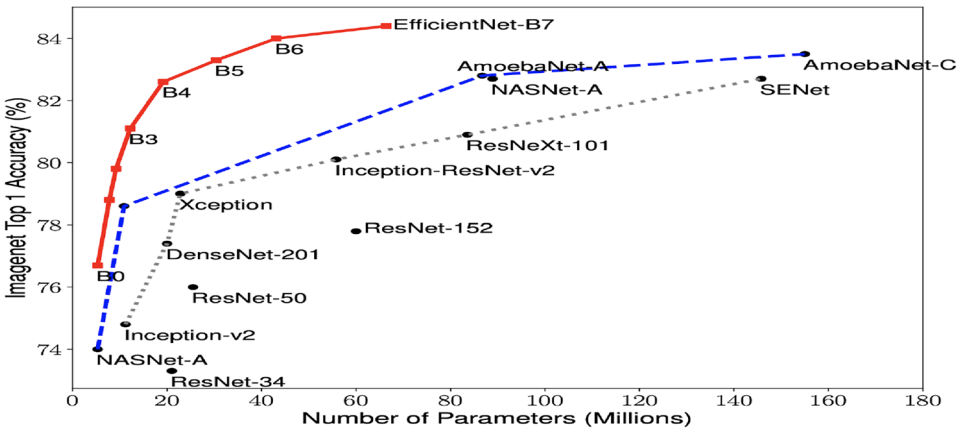

Fig. 9 Recent accuracy results of EfficientNet-B0

engineering techniques [44]. Feature engineering is a process of feature extraction from raw data, which requires a better way of describing underlying problems. One of the main advantages of spatiotemporal models over the traditional handcrafted machine learning algorithms is the extraction of features via state-of-the-art features engineering models [25,34]. A DNN-based spatiotemporal model inspects the data and extracts various classes of extensive feature sets that combine and correlate the data to enable quick learning without explicit instructions. Hence, the multimedia community can sometimes save months of work. In addition, the spatiotemporal model can uncover new, more complex and more abstract features that humans may miss.

6.1.2 Unstructured data

According to a recent survey from Gartner, 80% of data are unstructured, and most data are in audio, video and pdf forms, among other forms [59] (Table 3). However, in this unstructured space, the maximum area is covered by video data [80,82]. By using handcrafted algorithms, the data are difficult to analyse; hence, some data will be unutilized. This situation is where the deep spatiotemporal model can help. It can extract and be trained using various abstract features, and it can infer experiences that are pertinent to the motivation behind its preparation. For instance, a spatiotemporal model can uncover any existing relation between objects, actions and activities for the prediction of video understanding tasks.

6.1.3 Unsupervised feature learning

The evaluation of the value of features is one of the difficult problems that are encountered using traditional approaches because data labelling is expensive and tedious. Sometimes, labelling is simple but time-consuming. For instance, labelling manually a cat or a dog is easy, but an algorithm must compare thousands of examples to label objects. In addition, it is challenging to extend this to the video domain for action prediction or temporal movement identification. In spatiotemporal models, well-labelled data [106] are obsolete because the models learn without guidelines. Other traditional machine learning methods are not nearly as successful as this type of learning.

6.1.4 Efficient deliver of high-quality result

Using modern machine vision techniques, the extraction of abstract features can uncover various semantic relationships [33] between objects and their motions, which improves the quality of the results in video understanding. Via the effective use of a spatiotemporal model and deep learning techniques, problems can be solved efficiently. Once trained correctly, a neural brain can perform multiple repetitive routine jobs within a short period. The quality of the results never diminishes unless the training data include raw data that do not represent the considered problem. For example, Fig. 9 presents the recent result of EfficientNet-B0 [88], which realizes 84.4% top 1 and 97.1% top 5 accuracies.

6.2 Trends

Multimedia transformation domains continually update their models, and huge changes will require a set of new approaches that consider future trends. These models amplify consistent changes at an ever-expanding velocity. Several important trends are:

- Develop new mindsets and practices that grasp never-ending change;
- Embrace the "Ceaseless Next" system for making progress in a world that is continuously moving forward; and
- Position their domains to encourage interminable integration, innovation and delivery.

Table 3 Diversity of structured and unstructured data

	Structured data	Unstructured data
Characteristics	• Predefined data models	• No predefined data models
	• Predefined data models	• No predefined data models
	• Usually text only	• May be text, images, audio, video and other formats
	• Easy to search	• Difficult to search models
Resides in	• Relational databases	• Applications
	• Data warehouses	• No SQL databases
		• Data warehouses
		• Data lakes
Generated by	Human or machines	Human or machines
Typical application	• Airline reservation system	• Word processing
	• Inventory control	• Presentation software
	• CRM system	• Email clients
	• ERP system	• Tools for viewing or editing media
Examples	• Dates	• Audio files
	• Phone numbers	• Image files
	• Social security numbers	• Video files
	• Credit card numbers	• Surveillance imagery
	• Customer names	• Reports
	• Addresses	• Email messages
	• Product names and numbers	• Text files
	• Transaction information	

A boom in AI has been sparked and opens up a new frontier for research, which is, in part, due to the substantial success that has been realized in the detection, recognition, classification and captioning of the contents of static images and temporal video sequences by training deep neural networks on large data sets. These advances have been applied to every service, application and Internet of things (IoT) object to integrate an intelligent aspect, augment an application process or automate the human activities. The recent survey [1,39,57,90] recommends that enterprise architecture and technology innovation leaders explore the use of AI-driven autonomous physical devices.

6.2.1 Autonomous video computing

Video-understanding-based autonomous computing automates functions that are performed by humans. Its advanced programming modules and neural techniques provide valuable results regarding their surroundings and people. Autonomous computing uses various sets of operating conditions in several environments and multiple stages. In this discussion, an overview of autonomous computing is presented.

Autonomous computing in the physical and virtual world

In autonomous computing, various physical devices act in the real world, such as robots, drones, autonomous vehicles and AI-based IoT devices. Typically, they use video-understanding-powered applications. Autonomous video computing objects operate in various environments, such as sea, air and land, but every physical device focuses on operations that are directly or indirectly related to humans. For example, robots and AI-based IoT devices operate independently of humans. Drones are operated by humans and autonomous vehicles have human passengers.

Furthermore, video understanding applications improve virtually any physical device. Although virtual computing focuses on well-scoped needs, it typically automates routine human activities. It explores creative business environments and helps seek opportunities for the replacement of manual and semi-automated tasks with intelligent computing—for example, crime prevention through autonomous patrolling robots, advanced agriculture and safer automobile transportation. In communication, video understanding impacts upcoming industry 4.0 ecosystems such as front-line workers, traditional workplaces and industrial environments via the use of virtual assistants or independent agents [35]. This autonomous video computing spectrum spans from semi-autonomous to fully autonomous for a specified task or a defined context with various expediency levels, such as human-assisted, partial, conditional, high and fully autonomous. For example, a self-controlled vacuum cleaner has limited autonomy and intelligence, while a drone identifies obstacles and flies [45].

Collaborative industry 4.0 Video understanding promotes the collaborative industry as it proliferates from independent, intelligent models to a group of collaborative intelligent models, and it will work together with multiple devices either with human input or independently. For example, a drone that examines a large field and determines whether it is ready for harvesting is an "autonomous harvester." In the cargo delivery system, autonomous vehicles transport packages to their destinations; the final delivery will be conducted by robots or drones that are onboard the vehicle. A significant contributor to the army examines the targeted area for a drone attack or defends army targets [74]. In all these cases, spatiotemporal cues and video understanding play significant roles. Other examples of recent developments are as follows:

- Intel used swarm drones for the Winter Olympics opening ceremony in 2018 [4];
- Autonomous police vehicles deploy drones for surveillance [5]; and
- Car manufacturers merge scenarios by using communicating vehicles to optimize traffic flows [28].

6.2.2 Augmented video analytics

In video data analysis, traditional handcrafted methods are used to explore feasible pattern combinations. Manually explored features have more noise and suffer from the loss of some abstract features. Not enough data are available for scientists to complete the video analytics task. Augmented data analysis resolves this problem by using deep-learning-based analytics methods to empower the data scientist, which transform machine learning to high-level analysis of the development, sharing and consumption of analytics models. The following are related research trends:

- Maximize the value of data science efforts by empowering citizen data scientists;
- Augmented analytics is the future of data and analytics; and
- Market guide for process mining.

6.2.3 Immersive experience

Video understanding has strengths in virtual [97], augmented and mixed reality applications. Combinations of these models shift both perception and interaction models towards future immersive performance. Human–computer interaction [115], virtual assistance [35] using human action recognition and gesture-based vision-controlled system research boost the performance. These trends promote multiple research and industrial domains, such as digital design, edge computing, IoT, quantum computing and intelligent consumer application design.

6.2.4 Smart and safer cities

Video understanding makes a state-of-the-art contribution to smarter and safer cities [12]. In our public domain, video cameras [111] are commonly used as security devices in various scenarios, such as traffic management [41] and security surveillance systems. One of the significant applications of video understanding is in surveillance systems. Our sense of security depends on the eyes that watch over us and on their ability to understand what they see, thus allowing for the information to be shared in real time. Security systems should never lose information, whether on one site or many, and a robust system is essential to their success. Video-understanding-based intelligent visual management systems can be extended to obtain holistic security management solutions with built-in multi-way integration plug-ins, seamless integration with cognitive video analytics and video-centric situational awareness for incident management capabilities that enable them to safeguard us adequately.

7 Major challenges and weaknesses

7.1 Challenges

Reality factors (space and time) are universal parts of perceptions in diverse domains [2], including earth sciences, epidemiology, social sciences, mobile health, neuroscience, climate science and transportation. They are extended to computer vision in the multimedia domain for video analytics and analysis. A video contains abundant spatiotemporal cues. Mining of these cues is not easy [2], and spatiotemporal cue extraction faces the usual challenges that are faced in event recognition, such as tracking an action throughout a video and localizing the time span when an action occurs. However, there are additional challenges, such as the following:

- Background litter or object obstruction in the video;
- Spatial entanglement in a scene that involves the candidate objects;
- Linkage actions between frames that are due to irregular camera motion; and
- Predicting the optical flow of an action.

However, there is a more fundamental problem to consider with the classical approach to video understanding. We cannot address the problem via a straightforward approach for video understanding. Even object localization techniques require region proposition for classification [69]. This shortcoming can be more severe in the temporal domain and would result in incremental progress, which would render any such approach impossible to use.

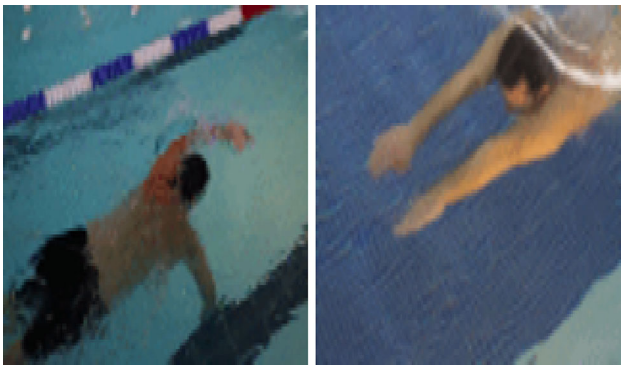

Fig. 10 Long-term motion cue example from a swimming video in UCF-101

For a machine, recognition of a cat or a dog in static images is an intelligent task. However, a more challenging task for a deep learning model is to determine when the dog is walking on the floor and chasing the cat around a kitchen. The next challenge is to teach a machine to understand not only what things are contained in the video but also what is occurring.

In video understanding, significant challenges are encountered when two actions share similar backgrounds [71,101], and the exploration of long-term motion cues is the most difficult challenge in understanding video content [2,80], which includes capturing the spatiotemporal context over frames. Furthermore, the captured spatial cues must compensate for image movement. Reliable spatial object detection is insufficient, as the motion cues carry better information. There are local and global contexts regarding motion cues that should be captured for powerful predictions. For instance, consider the video representation in Fig. 10. A satisfactory image classifier can recognize the human and the water body in both videos but is universal in nature, and the subsequent temporal action differentiates the front crawl from the breaststroke.

7.2 Weakness

7.2.1 The need for large amounts of data

"How much data is sufficient for training a deep learning algorithm?" This is a common question in deep learning. Unfortunately, there is no straightforward answer, but according to the data science community, for the exploration of extensive abstract features, a large amount of data are required [17,78]. In this case, for the deep neural network architecture, the required amount of data for training will be much larger compared to the traditional algorithms [68] because the deep learning algorithm is a two-part task. First, the domain must be identified; then, the problem is solved. When the training starts, the network begins from scratch. We can identify the domain, and the network requires many parameters for tuning and "playing around." In our brains, the neurons behave in a manner similar to a deep learning algorithm. We need many essential experiences with what already we learned. In this way, deep neural networks gain experiences from a vast number of representations, which is directly proportional to the increase in the data scale. In addition, there are no standard benchmark data sets; the popularly known data sets are UCF-101 [81] and SPORTS-1M [37]. Implementing a flexible architecture for SPORTS-1M is more expensive. The high spatial correlation in UCF-101 lengthens the training time. Therefore, the multimedia community needs a new set of standard benchmarks that are based on highly complex problems.

7.2.2 Deep learning models are black boxes

One of the significant limitations of deep learning is that it is difficult to understand the mechanism and how the neural network obtains a solution. It is difficult to see inside of the network to observe how it performs [16]. The deep neural network architecture is embedded with thousands of simulated neurons, which are similar to neurons in the human brain. They are organized into dozens or even hundreds of intricately interconnected hidden layers. Together, they can form a highly complex web, where inputs sent from one layer to the next layer until an overall output is produced. Additionally, back-propagation [22] changes the values and the evaluations of each neuron such that the network learns to build the desired output accurately. Even though the network produces valuable results, in the "thinking" process, the lack of transparency renders it difficult to predict when failures might occur [64]. Hence, such networks are unacceptable for domains where verification of the procedure is essential [10,54]. One example is the medical domain. Consider as an example a deep learning algorithm that was applied to patient records. After training, it was able to detect some illnesses better than human doctors were able to. This is a satisfactory result. However, for a tool to be useful in medicine, the accuracy of its prediction must be proven to justify a change in someone's treatment. Without interpretation, it is difficult to gain the trust of patients or to understand why missteps in diagnosis can be made [38].

7.2.3 Overfitting of the model

Overfitting refers to modelling the "training data" too well using deep neural networks or over-training the model [51, 113]. Overfitting occurs when a network learns in detail, and common noise occurs in the training data, which adversely affects the overall performance of the model in various scenarios. In a neural network, overfitting is a common problem, especially in state-of-the-art networks, which often have an exponentially large number of parameters and noise. When the model is overtrained, the accuracy stops improving after

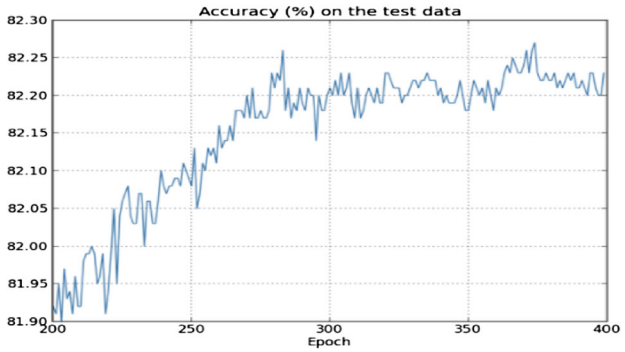

Fig. 11 Example of overfitting of the model at the training level

a threshold number of epochs (Fig. 11). In the example, the accuracy changes at 275th epoch from 82.15 to 82.25%; hence, the model was overtrained at the 275th epoch level.

8 Conclusions

In this paper, an extensive survey has been conducted on spatiotemporal features in video data, deep learning approaches for extracting those features, state-of-the-art deep neural networks that address various problems such as long-term and short-term motions cues and various hybrid frameworks for feature extraction procedures. This study explores real-world video understanding problems, such as societal, industrial, research and academic problems. In addition, current and future perspective research avenues are investigated, such as action removal in the spatial region, super action prediction and critical action prediction in the temporal domain, for the realization of the deep learning LLL (less data, less layers and less computational power) goals. We discussed the major strengths and trends of video understanding, and we identified several challenges and weakness that are faced by the multimedia community in the extraction of spatiotemporal patterns using deep neural networks.

Acknowledgements We would like to thank Editor for helpful suggestions, and the anonymous reviewers for constructive comments.

References

1. Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, Hasan M, Van Essen BC, Awwal AAS, Asari VK (2019) A state-of-the-art survey on deep learning theory and architectures. Electronics 8(3):292
2. Atluri G, Karpatne A, Kumar V (2018) Spatio-temporal data mining: a survey of problems and methods. ACM Comput Surv: CSUR 51(4):83
3. Baker S, Scharstein D, Lewis JP, Roth S, Black MJ, Szeliski R (2011) A database and evaluation methodology for optical flow. Int J Comput Vis 92(1):1–31
4. Barrett B (2018) Inside the olympics opening ceremony world-record drone show. In: wired. https://www.wired.com/story/olympics-opening-ceremony-drone-show/
5. Bhorge SB, Manthalkar RR (2018) Three-dimensional spatio-temporal trajectory descriptor for human action recognition. Int J Multimed Inf Retr 7(3):197–205
6. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 3:257–267
7. Burghouts GJ, Schutte K (2013) Spatio-temporal layout of human actions for improved bag-of-words action detection. Pattern Recogn Lett 34(15):1861–1869
8. Chalapathy R, Chawla S (2019) Deep learning for anomaly detection: a survey. arXiv:1901.03407
9. Chaudhry R, Ravichandran A, Hager G, Vidal R (2009) June. Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 1932–1939
10. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, Naessens JM, Larson DW, Liu H (2019) Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med 2(1):1–5
11. Chen K, Kovvuri R, Gao J, Nevatia R (2018) MSRC: multimodal spatial regression with semantic context for phrase grounding. Int J Multimed Inf Retr 7(1):17–28
12. Cocchia A (2014) Smart and digital city: a systematic literature review. In: Dameri RP, Rosenthal-Sabroux C (eds) Smart city. Progress in IS. Springer, Cham, pp 13–43. https://doi.org/10.1007/978-3-319-06160-3_2
13. Deldjoo Y, Elahi M, Quadrana M, Cremonesi P (2018) Using visual features based on MPEG-7 and deep learning for movie recommendation. Int J Multimed Inf Retr 7(4):207–219
14. Du Y, Yuan C, Li B, Zhao L, Li Y, Hu W (2018) Interaction-aware spatio-temporal pyramid attention networks for action classification. In: Proceedings of the European conference on computer vision (ECCV), pp 373–389
15. Evensen D (2019) The rhetorical limitations of the #FridaysForFuture movement. Nat Clim Chang 9:428–430. https://doi.org/10.1038/s41558-019-0481-1
16. Fan J, Ma C, Zhong Y (2019) A selective overview of deep learning. arXiv:1904.05526
17. Federal Highway Administration (2015) Video analytics research projects. U.S Department of Transportation. 16 p
18. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1933–1941
19. Gammulle H, Denman S, Sridharan S, Fookes C (2017) March. Two stream lstm: a deep fusion framework for human action recognition. In: 2017 IEEE Winter conference on applications of computer vision (WACV). IEEE, pp 177–186
20. Gonzalez TF (2007) Handbook of approximation algorithms and metaheuristics. Chapman and Hall, London
21. Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends Neurosci 15(1):20–25
22. Guiming D, Xia W, Guangyan W, Yan Z, Dan L (2016) Speech recognition based on convolutional neural networks. In: 2016 IEEE international conference on signal and image processing (ICSIP). IEEE, pp 708–711
23. Guo Y, Liu Y, Georgiou T, Lew MS (2018) A review of semantic segmentation using deep neural networks. Int J Multimed Inf Retr 7(2):87–93
24. Hatcher WG, Yu W (2018) A survey of deep learning: platforms, applications and emerging research trends. IEEE Access 6:24411–24432

25. He D, Li F, Zhao Q, Long X, Fu Y, Wen S (2018) Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. arXiv:1806.10319
26. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
27. Hoang VD, Hoang DH, Hieu CL (2018) Action recognition based on sequential 2D-CNN for surveillance systems. In: IECON 2018-44th annual conference of the IEEE industrial electronics society. IEEE, pp 3225–3230
28. Honda (2018) Cooperative merge. In: Honda news. http://www.multivu.com/players/English/7988331-honda-ces-cooperative-mobility-ecosystem/
29. Hou R, Chen C, Shah M (2017) Tube convolutional neural network (T-CNN) for action detection in videos. In: Proceedings of the IEEE international conference on computer vision, pp 5822–5831
30. Huang H, Yu PS, Wang C (2018) An introduction to image synthesis with generative adversarial nets. arXiv:1803.04469
31. Hui TW, Tang X, Change Loy C (2018) Liteflownet: a lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8981–8989
32. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2017) Flownet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2462–2470
33. Jiang YG, Wu Z, Tang J, Li Z, Xue X, Chang SF (2018) Modeling multimodal clues in a hybrid deep learning framework for video classification. IEEE Trans Multimed 20(11):3137–3147
34. Jiang YG, Wu Z, Wang J, Xue X, Chang SF (2017) Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Trans Pattern Anal Mach Intell 40(2):352–364
35. Kahn J (2018) Meet 'Millie' the Avatar. She'd like to sell you a pair of sunglasses. In: Bloomberg. https://www.bloomberg.com/news/articles/2018-12-15/meet-millie-the-avatar-she-d-like-to-sell-you-a-pair-of-sunglasses
36. Kangwei L, Jianhua W, Zhongzhi H (2018) Abnormal event detection and localization using level set based on hybrid features. Signal Image Video Process 12(2):255–261
37. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732
38. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17(1):195
39. Kong Y, Fu Y (2018) Human action recognition and prediction: a survey. arXiv:1806.11230
40. Kruger N, Janssen P, Kalkan S, Lappe M, Leonardis A, Piater J, Rodriguez-Sanchez AJ, Wiskott L (2012) Deep hierarchies in the primate visual cortex: what can we learn for computer vision? IEEE Trans Pattern Anal Mach Intell 35(8):1847–1871
41. Kumaran SK, Dogra DP, Roy PP (2019) Anomaly detection in road traffic using visual surveillance: a survey. arXiv:1901.08292
42. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
43. Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: CVPR—IEEE conference on computer vision & pattern recognition, Jun 2008, Anchorage, USA, pp 1–8
44. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
45. Lenz I, Gemici M, Saxena A (2012) Low-power parallel algorithms for single image based obstacle avoidance in aerial robots. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 772–779
46. Levine S, Pastor P, Krizhevsky A, Ibarz J, Quillen D (2018) Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. Int J Robot Res 37(4–5):421–436
47. Li F, Du J (2012) October. Local spatio-temporal interest point detection for human action recognition. In: 2012 IEEE fifth international conference on advanced computational intelligence (ICACI). IEEE, pp 579–582
48. Li Q, Qiu Z, Yao T, Mei T, Rui Y, Luo J (2017) Learning hierarchical video representation for action recognition. Int J Multimed Inf Retr 6(1):85–98
49. Li X, Pang T, Liu W, Wang T (2017) Fall detection for elderly person care using convolutional neural networks. In: 2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI). IEEE, pp 1–6
50. Liu J, Sun C, Xu X, Xu B, Yu S (2019) A spatial and temporal features mixture model with body parts for video-based person re-identification. Appl Intell 49(9):3436–3446
51. Livni R, Shalev-Shwartz S, Shamir O (2014) On the computational efficiency of training neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems, vol 27. Curran Associates, Inc., pp 855–863. http://papers.nips.cc/paper/5267-on-the-computational-efficiency-of-training-neural-networks.pdf
52. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
53. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision, pp 2720–2727
54. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. Mol Pharm 13(5):1445–1454
55. Marcus G (2018) Deep learning: a critical appraisal. arXiv:1801.00631
56. Melfi R, Kondra S, Petrosino A (2013) Human activity modeling by spatio temporal textural appearance. Pattern Recogn Lett 34(15):1990–1994
57. Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3061–3070
58. Mithun NC, Li J, Metze F, Roy-Chowdhury AK (2019) Joint embeddings with multimodal cues for video-text retrieval. Int J Multimed Inf Retr 8(1):3–18
59. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1):1
60. Naseer S, Saleem Y, Khalid S, Bashir MK, Han J, Iqbal MM, Han K (2018) Enhanced network anomaly detection based on deep neural networks. IEEE Access 6:48231–48246
61. Ouadiay FZ, Bouftaih H, Bouyakhf EH, Himmi MM (2018) Simultaneous object detection and localization using convolutional neural networks. In: 2018 international conference on intelligent systems and computer vision (ISCV). IEEE, pp 1–8
62. Palmer R, West G, Tan T (2012) Scale proportionate histograms of oriented gradients for object detection in co-registered visual and range data. In: 2012 international conference on digital image computing techniques and applications (DICTA). IEEE, pp 1–8
63. Papadopoulos K, Demisse G, Ghorbel E, Antunes M, Aouada D, Ottersten B (2019) Localized trajectories for 2D and 3D action recognition. arXiv:1904.05244
64. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, pp 372–387

65. Peng K, Chen X, Zhou D, Liu Y (2009) 3D reconstruction based on SIFT and Harris feature points. In: 2009 IEEE international conference on robotics and biomimetics (ROBIO). IEEE, pp 960–964
66. Peng Y, Zhao Y, Zhang J (2018) Two-stream collaborative learning with spatial-temporal attention for video classification. IEEE Trans Circuits Syst Video Technol 29(3):773–786
67. Qiu Z, Yao T, Mei T (2017) Learning deep spatio-temporal dependence for semantic video segmentation. IEEE Trans Multimed 20(4):939–949
68. Ray KS, Chakraborty S (2019) Object detection by spatio-temporal analysis and tracking of the detected objects in a video with variable background. J Vis Commun Image Represent 58:662–674
69. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates, Inc., pp 91–99 http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf
70. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, Ross AS, Milojevic-Dupont N, Jaques N, Waldman-Brown A, Luccioni A (2019) Tackling climate change with machine learning. arXiv:1906.05433
71. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
72. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on multimedia. ACM, pp 357–360
73. Sekma M, Mejdoub M, Amar CB (2015) Human action recognition based on multi-layer fisher vector encoding method. Pattern Recogn Lett 65:37–43
74. Seligman L (2016) How swarming drones could change the face of air warfare. In: Def. News. https://www.defensenews.com/2016/05/17/how-swarming-drones-could-change-the-face-of-air-warfare/
75. Sermanet P, Chintala S, LeCun Y (2012) Convolutional neural networks applied to house numbers digit classification. arXiv:1204.3968
76. Shou Z, Lin X, Kalantidis Y, Sevilla-Lara L, Rohrbach M, Chang SF, Yan Z (2019) Dmc-net: generating discriminative motion cues for fast compressed video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1268–1277
77. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
78. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KD (eds) Advances in neural information processing systems, vol 27. Curran Associates, Inc., pp 568–576. http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf
79. Singh B, Marks TK, Jones M, Tuzel O, Shao M (2016) A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1961–1970
80. Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of Big Data challenges and analytical methods. J Bus Res 70:263–286
81. Soomro K, Zamir AR, Shah M (2012) A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision
82. Sreenu G, Durai MS (2019) Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data 6(1):48
83. Sun C, Shetty S, Sukthankar R, Nevatia R (2015) Temporal localization of fine-grained actions in videos by domain transfer from web images. In: Proceedings of the 23rd ACM international conference on multimedia. ACM, pp 371–380
84. Sun D, Yang X, Liu MY, Kautz J (2018) PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8934–8943
85. Sun L, Jia K, Yeung DY, Shi BE (2015) Human action recognition using factorized spatio-temporal convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4597–4605
86. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
87. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: International conference on artificial neural networks. Springer, Cham, pp 270–279
88. Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946
89. Thakkar K, Narayanan PJ (2018) Part-based graph convolutional network for action recognition. arXiv:1809.04983
90. Tian Y, Pei K, Jana S, Ray B (2018) Deeptest: automated testing of deep-neural-network-driven autonomous cars. In: Proceedings of the 40th international conference on software engineering. ACM, pp 303–314
91. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
92. Tripathi RK, Jalal AS, Agrawal SC (2018) Suspicious human activity recognition: a review. Artif Intell Rev 50(2):283–339
93. Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW (2017) Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access 6:1155–1166
94. Wang H, Kläser A, Schmid C, Liu CL (2011) Action recognition by dense trajectories. CVPR. In: IEEE conference on computer vision & pattern recognition, June 2011. Colorado Springs, United States, pp 3169–3176
95. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp 3551–3558
96. Wang L, Ge L, Li R, Fang Y (2017) Three-stream CNNs for action recognition. Pattern Recogn Lett 92:33–40
97. Wang L, Hu W, Tan T (2003) Recent developments in human motion analysis. Pattern Recogn 36(3):585–601
98. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4305–4314
99. Wang P, Li W, Ogunbona P, Wan J, Escalera S (2018) RGB-D-based human motion recognition with deep learning: a survey. Comput Vis Image Underst 171:118–139
100. Wang T, Snoussi H (2012) Histograms of optical flow orientation for visual abnormal events detection. In: 2012 IEEE ninth international conference on advanced video and signal-based surveillance. IEEE, pp 13–18
101. Wang Y, Long M, Wang J, Yu PS (2017) Spatiotemporal pyramid network for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1529–1538

102. Wang Z, Ren J, Zhang D, Sun M, Jiang J (2018) A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing 287:68–83
103. Weng X (2019) On the importance of video action recognition for visual lipreading. arXiv:1903.09616
104. Wu Z, Jiang YG, Wang J, Pu J, Xue X (2014) November. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: Proceedings of the 22nd ACM international conference on multimedia. ACM, pp 167–176
105. Wu Z, Wang X, Jiang YG, Ye H, Xue X (2015) Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd ACM international conference on multimedia. ACM, pp 461–470
106. Wu Z, Yao T, Fu Y, Jiang YG (2016) Deep learning for video classification and captioning. arXiv:1609.06782
107. Xu Z, Yang Y, Hauptmann AG (2015) A discriminative CNN video representation for event detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1798–1807
108. Yao L (2016) Extract the relational information of static features and motion features for human activities recognition in videos. Intell Neurosci 2016:3. https://doi.org/10.1155/2016/1760172
109. Ye H, Wu Z, Zhao RW, Wang X, Jiang YG, Xue X (2015) Evaluating two-stream CNN for video classification. In: Proceedings of the 5th ACM on international conference on multimedia retrieval. ACM, pp 435–442
110. Yuan Y, Zheng X, Lu X (2016) A discriminative representation for human action recognition. Pattern Recogn 59:88–97
111. Zabłocki M, Gościewska K, Frejlichowski D, Hofman R (2014) Intelligent video surveillance systems for public spaces—a survey. J Theor Appl Comput Sci 8(4):13–27
112. Zhan F, Zhu H, Lu S (2019) Spatial fusion gan for image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3653–3662
113. Zhang C, Vinyals O, Munos R, Bengio S (2018) A study on overfitting in deep reinforcement learning. arXiv:1804.06893
114. Zhang H, Liu D, Xiong Z (2019) Two-stream oriented video super-resolution for action recognition. arXiv:1903.05577
115. Zhang J, Feng Z, Su Y, Xing M, Xue W (2019) Riemannian spatio-temporal features of locomotion for individual recognition. Sensors 19(1):56
116. Zhang W, Luo Y, Chen Z, Du Y, Zhu D, Liu P (2019) A robust visual tracking algorithm based on spatial-temporal context hierarchical response fusion. Algorithms 12(1):8
117. Zhang XY, Shi H, Li C, Zheng K, Zhu X, Duan L (2019) Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision. In: Proceedings of the 33rd AAAI conference on artificial intelligence, pp 1–8
118. Zhao R, Ali H, Van der Smagt P (2017) Two-stream RNN/CNN for action recognition in 3D videos. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 4260–4267
119. Zhu AZ, Yuan L, Chaney K, Daniilidis K (2018) EV-FlowNet: self-supervised optical flow estimation for event-based cameras. arXiv:1802.06898
120. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232