electronics

MDPI

Review

Deep Learning Innovations in Video Classification: A Survey on Techniques and Dataset Evaluations

Makara Mao [1], Ahyoung Lee [2] and Min Hong [3,*]

[1] Department of Software Convergence, Soonchunhyang University, Asan-si 31538, Republic of Korea; makaramao@sch.ac.kr
[2] Department of Computer Science, Kennesaw State University, Marietta, GA 30060, USA; alee146@kennesaw.edu
[3] Department of Computer Software Engineering, Soonchunhyang University, Asan-si 31538, Republic of Korea
* Correspondence: mhong@sch.ac.kr

Abstract: Video classification has achieved remarkable success in recent years, driven by advanced deep learning models that automatically categorize video content. This paper provides a comprehensive review of video classification techniques and the datasets used in this field. We summarize key findings from recent research, focusing on network architectures, model evaluation metrics, and parallel processing methods that enhance training speed. Our review includes an in-depth analysis of state-of-the-art deep learning models and hybrid architectures, comparing models to traditional approaches and highlighting their advantages and limitations. Critical challenges such as handling large-scale datasets, improving model robustness, and addressing computational constraints are explored. By evaluating performance metrics, we identify areas where current models excel and where improvements are needed. Additionally, we discuss data augmentation techniques designed to enhance dataset accuracy and address specific challenges in video classification tasks. This survey also examines the evolution of convolutional neural networks (CNNs) in image processing and their adaptation to video classification tasks. We propose future research directions and provide a detailed comparison of existing approaches using the UCF-101 dataset, highlighting progress and ongoing challenges in achieving robust video classification.

Keywords: UCF-101 dataset; video classification; supervised learning; video processing; parallel processing; deep learning

Citation: Mao, M.; Lee, A.; Hong, M. Deep Learning Innovations in Video Classification: A Survey on Techniques and Dataset Evaluations. *Electronics* **2024**, *13*, 2732. https://doi.org/10.3390/electronics13142732

Academic Editor: Hamed Mozaffari

Received: 7 June 2024
Revised: 9 July 2024
Accepted: 10 July 2024
Published: 11 July 2024

1. Introduction

In recent years, video classification, despite its challenges, has become successful within deep learning. This area has gained attention following the advent of deep learning models, which have been proven to be practical tools for automatic video classification. The plethora of online video datasets highlights the importance of accurate video classification tasks. The rapid advancement of digital technologies has led to a substantial increase in the generation and consumption of video content globally. Various streaming and live video platforms, such as YouTube, Youku, TikTok, and others, have become essential for daily entertainment, significantly influencing user behavior and generating considerable revenue through user engagement. For example, YouTube alone reports over 122 million active users engaging with its content daily [1].

Additionally, Youku is one of China's most popular video platforms. Notably, the number of views from external links on Youku is relatively higher than on YouTube [2]. Lastly, TikTok, widely recognized for its short-form videos, has proven to be an exceptionally effective platform for advertising and launching new products [3]. The advancement of deep learning in video classification has further enhanced the capabilities of these platforms, enabling more personalized content delivery and improved user experiences. Based on the description above, video classification tasks play a fundamental role in enhancing

accuracy in action recognition, presenting significant challenges and opportunities for video classification, which has become a challenging and successful area within deep learning in recent years.

Video classification tasks are not just for academic research, but also present significant challenges and opportunities for major companies such as Google, Mountain View, CA, USA, (YouTube); Facebook, Menlo Park, CA, USA, (Meta); Amazon, Seattle, WA, USA, (Amazon Prime Video); Microsoft, Redmond, DC, USA; and Apple, Cupertino, CA, USA. These companies invest in competitions and release datasets to advance video classification technologies across various sectors, including entertainment and media, healthcare, security and surveillance, automotive, and retail. By doing so, they drive innovation and improve applications in these diverse fields. For instance, Google has introduced the YouTube-8M dataset [4], a valuable resource for researchers, containing millions of video features and over 3700 labels. Such initiatives underscore the importance of robust video classification models in various applications, both in industry and academic research.

Additionally, advancements in video analysis have seen the introduction of sophisticated models such as GPT-4 on the Azure platform. GPT-4, a state-of-the-art language model, can be fine-tuned for video analysis tasks by leveraging its extensive natural language understanding capabilities [5]. The Azure platform offers scalable and efficient processing, enabling real-time video analysis applications. Integrating such models with other machine learning frameworks enhances the robustness and versatility of video analysis systems.

Our survey aims to fill these gaps by providing an updated and thorough review of deep learning techniques for video classification, incorporating the latest research trends and methodologies. The following contributions are made:

- **Comprehensive overview of CNN-based models:** We summarize the state-of-the-art CNN models used for image analysis and their applications in video classification, highlighting key architectures such as LeNet-5, AlexNet, VGG-16, VGG-19, Inception, GoogleNet, ResNet, SqueezeNet, Enet, ShuffleNet, and DenseNet. Each model's features, evaluations, and problem-solving capabilities are detailed to provide a foundational understanding of video classification tasks.
- **Deep learning approaches for video classification:** We cover the integration of CNNs and recurrent neural networks (RNNs) for video classification. CNNs capture spatial features within video frames, while RNNs model temporal dependencies, making this combination effective for video understanding and exploring the use of transformer models in conjunction with CNNs to enhance spatial and temporal feature modeling.
- **Uni-modal and multi-modal fusion frameworks:** We compare uni-modal approaches, which utilize a single data modality, with multi-modal approaches that integrate various modalities (text, audio, visual, sensor, and more). Multi-modal methods improve classification accuracy by leveraging the complementary strengths of different data types.
- **Feature extraction and representation techniques:** Effective feature extraction is critical for video classification. We review techniques such as Scale-Invariant Feature Transform (SIFT) and data augmentation methods like random rotation and shift, which have been shown to improve classification accuracy in video datasets.

The rest of the paper is organized as follows: Section 2 covers background studies on video classification. Section 3 details techniques from papers using deep learning models for video classification. Section 4 overviews benchmark datasets, evaluation metrics, and comparisons. Section 5 provides the research discussion, the challenges in video classification, and future directions. Lastly, Section 6 provides the conclusions. Table 1 provides the abbreviations used in this paper.

Table 1. List of important abbreviations.

Abbreviations	Description
3DCLSTM	3D Convolutional Long Short-Term Memory
Adam	Adaptive Momentum
AMD	Asymmetric Masked Distillation
ARTNet	Appearance-and-Relation Network
BIKE	BIdirectional Crossmodal Knowledge Exploration
C3D	3D Convolutional Network
CCS	Cooperative Cross-Stream
CD-UAR	Cross-Dataset UAR
CNN	Convolutional Neural Network
CNNs	Convolutional Neural Networks
CRNNs	Convolutional Recurrent Neural Networks
DB-LSTM	Deep Bidirectional LSTM
DMC-Net	Discriminative Motion Cue
FASTER32	Feature Aggregation for Spatio-Temporal Redundancy (32 Frames)
FPGA	Field-Programmable Gate Array
HalluciNet	Hallucination Network
HAM	Hybrid Attention Model
HFLSTD	Histogram of fuzzy local spatio-temporal descriptors
I3D	Interactive Three-Dimensional
iFDT	Improved fuzzy dense trajectories
LGD-3D Flow	Local Global Diffusion
LSTM	Long Short-Term Memory
LTC	Long-Term Temporal Convolutions
MARS	Motion-Augmented RGB Stream
MiCRObE	Max Calibration mixture of Experts
MLGCN	Multi-Laplacian Graph Convolutional Network
MR Two-Sream R-CNN	Multi-Region Two-Stream R-CNN
MV-CNN	Motion Vector CNN
OmniVec	Omnidirectional Vector
PoTion	Pose moTion
Prob-Distill	Probabilistic Distribution
ReLU	Rectified Linear Unit
Res3D	Residual Three Dimensional
RGB	Red–Green–Blue
RGB-I3D	Red–Green–Blue—Interactive Three Dimensional
RNNs	Recurrent Neural Networks
R-STAN	Residual Spatial–Temporal Attention Network
SGD	Stochastic Gradient Descent
STAM	Space–Time Attention Model
ST-ResNet	Spatio-temporal Residual Network
SVT	Self-supervised Video Transformer
TDD	Trajectory-pooled Deep-convolutional Descriptor
TS-LSTM	Temporal Segment LSTM
TSN	Temporal Segment Network
UAR	Unseen Action Recognition
VidTr-L	Video Transformer—Large
VIMPAC	Video pre-training via Masked token Prediction And Contrastive learning
ZeroI2V	Zero-cost Adaptation Paradigm

2. Background Studies on Video Classification

This section provides a comprehensive description of the background studies on video classification. The review process included summarizing relevant papers, identifying keywords related to the proposed topic, selecting appropriate publishers, analyzing abstract themes, and applying inclusion and exclusion criteria to ensure the quality of the papers in

our survey. Figure 1 provides the selection process used in the current research study to create a video classification survey.

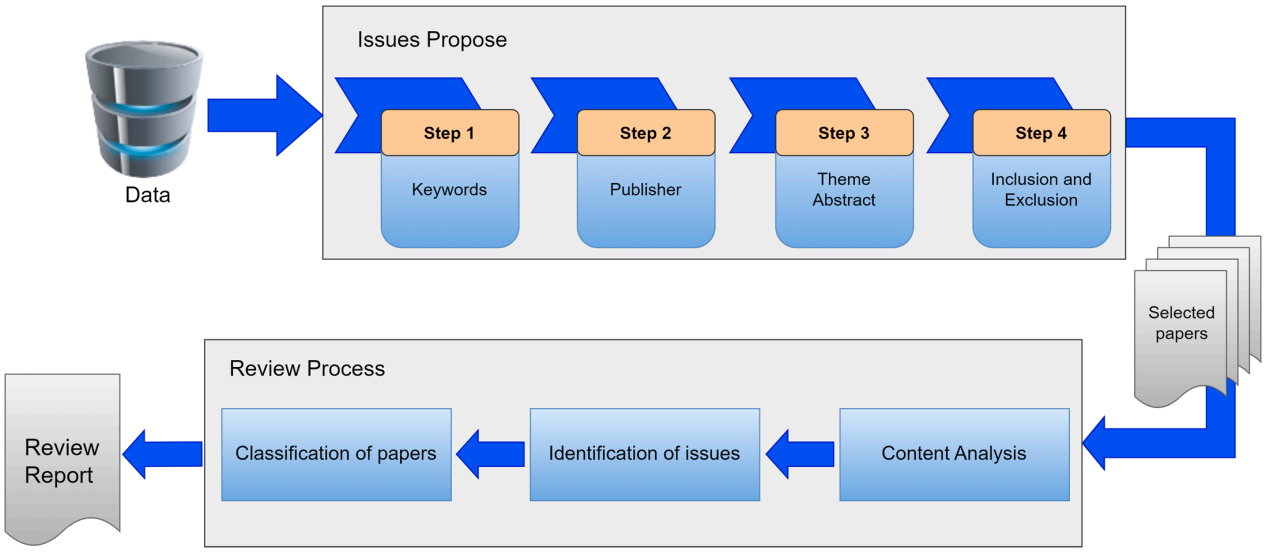

Figure 1. Review process of our paper.

2.1. Relevant Surveys

Several review articles have addressed video classification recently, as summarized in Table 2. The existing works often need more coverage of the most recent state-of-the-art advancements, require closer alignment with current research trends, and have limitations. The following discussion highlights the limitations and critical points of these works.

Table 2. Summary of recent related works on video classification papers.

Ref.	Year	Features	Drawbacks
Anushya [6]	2020	Video classification, tagging, and clustering.	Limited scope and lacks detailed information.
Rani et al. [7]	2020	Classify video content using text, audio, and visual features.	Did not include an analysis of the latest state-of-the-art approaches.
Li et al. [8]	2020	Real-time sports video classification.	Focuses specifically on real sports video classification.
Zuo et al. [9]	2020	Fuzzy local spatio-temporal descriptors for video action recognition.	Uncertainty in pixel voting due to varying numbers of bins.
Islam et at. [10]	2021	Machine learning techniques for classifying video.	Reviews are less focused on deep learning methods.
Ullah et al. [11]	2021	Recognizing human activities with deep learning.	Primarily emphasizes human activity recognition.
Rehman et al. [12]	2022	Detailed review of deep learning strategies for classifying videos.	Places less emphasis on pre-training and foundational model techniques in deep learning for video classification.
This study	2024	Comprehensive techniques for video classification, dataset benchmarking, and deep learning models.	-

Anushya [6] provides a comprehensive review covering methods for video classification, clustering, tagging, and training. However, the study's comprehensiveness, conciseness, coverage of the topic, datasets, and analysis of state-of-the-art approaches could be extended. It is important to keep the audience updated with the latest advancements in video classification by incorporating recent state-of-the-art approaches. Additionally, Rani et al. [7] conducted a recent review on video classification topics, covering some recent

video classification approaches and summarizing descriptions of recent works. An analysis of recent state-of-the-art approaches could be incorporated for further enhancements.

Li et al. [8] conducted a recent systematic review on live sports video classification, encompassing tools, video interaction features, and feature extraction methods. The paper concentrates on the classification of live sports videos. Zuo et al. [9] introduce fuzzy local spatio-temporal descriptors for video action recognition. They aim to tackle the uncertainty in pixel voting arising from the varying numbers of bins in traditional feature extractors. The study highlights a drawback related to the inherent uncertainty in the process of pixel voting concerning the bins in the histogram, which the proposed fuzzy descriptors seek to resolve. In contrast, Islam et al. [10] introduced models for video classification including deep learning and proposed machine learning techniques. Ullah et al. [11] also conducted a recent systematic review on human activity recognition. Furthermore, Rehman et al. [12] review comprehensive deep learning techniques for video classification. The review focuses on the pre-training and main model techniques in deep learning for video classification.

These existing reviews highlight the evolution of video classification research. However, to update recent studies, significant gaps and new trending paradigms will be addressed in our study, particularly in terms of coverage of the latest state-of-the-art advancements, comprehensive deep learning techniques, and dataset evaluations of recent models.

2.2. Evolution of CNNs in Image Processing

The CNN paradigm has evolved significantly and is extensively applied to image-processing tasks, demonstrating its practical relevance and impact. Its inherent parallelism and local connectivity properties make this architecture highly suitable for analog processing systems. Among these architectures, there are time-series approaches and evolved CNNs for image processing, such as LeNet-5, AlexNet, VGG-16, VGG-19, Inception, GoogleNet, ResNet, YOLO, SqueezeNet, Enet, ShuffleNet, DenseNet, MobileNet, and BowNet. A summary of these popular CNN architectures, consistent with the pattern of leaning on networks, is shown in Figure 2, where the extent of the network increases from the left-most (LeNet-5) to the right-most (BowNet). LeNet-5, one of the earliest CNN architectures, has demonstrated the effectiveness of CNNs for image recognition with its simple yet effective structure, paving the way for deeper networks.

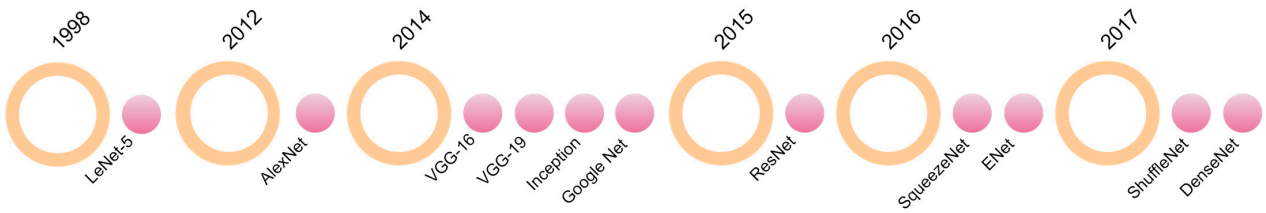

Figure 2. State-of-the-art image recognition using CNN architecture.

AlexNet significantly advanced the field by introducing Rectified Linear Unit (ReLU) activation functions and dropout for regularization, achieving a breakthrough performance in the ImageNet 2012 competition. VGG-16 and VGG-19 further deepened this network architecture, utilizing small 3 $\times$ 3 convolution filters, and achieved high accuracy on large-scale datasets. Inception (GoogleNet) introduced the Inception module to capture multi-scale features efficiently, balancing high performance with reduced computational complexity. ResNet addressed the vanishing gradient problem in deep networks through residual learning, enabling ultra-deep networks to train and set new benchmarks in image classification. SqueezeNet aimed for model compression, achieving AlexNet-level accuracy with 50$\times$ fewer parameters, making it ideal for resource-constrained environments. DenseNet, with its dense connectivity, improved gradient flow and feature reuse, and achieves high performance with efficient parameter usage with the different key features and contributions shown in Table 3.

Table 3. Detailed comparison of the critical features and contributions of each milestone CNN architecture.

Year	Model	Key Features	Contributions
1998	LeNet-5	Five layers, convolutional and pooling layers.	Pioneered CNNs for digit recognition.
2012	AlexNet	Eight layers, ReLU activation, dropout.	Won ImageNet 2012, popularized deep learning.
2014	VGGNet	16–19 layers, small (3×3) convolution filters.	Demonstrated the importance of depth.
2014	GoogLeNet	Inception modules, 22 layers.	Improved computational efficiency.
2015	ResNet	Residual blocks, up to 152 layers.	Enabled training of intense networks.
2015	YOLO	Real-time object detection.	Unified detection and classification, efficient for video analysis.
2016	SqueezeNet	Fire modules, 50× fewer parameters.	Achieved AlexNet-level accuracy with fewer parameters.
2016	ENet	Compound scaling.	Achieved state-of-the-art accuracy with fewer parameters.
2017	ShuffleNet	Point-wise group convolution, channel shuffle.	Efficient computation for mobile devices
2017	DenseNet	Dense connections between layers.	Promoted feature reuse reduced parameters.
2017	MobileNet	Depth-wise separable convolutions.	Optimized for mobile and embedded vision applications.
2019	BowNet	Encoder–decoder structure.	Real-time tracking of tongue contours in ultrasound data.

The transition from 2D CNNs to 3D CNNs marked a significant evolution, particularly relevant to video classification. While 2D CNNs process spatial information in images, 3D CNNs extend this capability to temporal dimensions, capturing motion and changes over time, which are crucial for understanding video data. Integrating CNNs with RNNs and transformer models further enhances video classification by combining spatial feature extraction with temporal dependency modeling, improving performance in capturing complex temporal dynamics in video data. These advancements collectively underscore the evolution of CNNs in tackling the challenges of video classification, with each contributing unique strengths in terms of model architecture, computational efficiency, and classification accuracy.

2.3. Fundamental Techniques in CNN-Based Image Processing

The thread that progresses from the formerly proposed architectures toward the recently proposed architectures depends on the network for image processing. In Table 4, numerous existing papers are detailed based on their descriptions of CNN models for image processing, including the authors' approaches in their works, the models' features, the models' names, evaluations, and the problems solved in the papers.

Table 4. Summary and findings of studies based on CNN models for image processing.

Approach	Features	Model	Evaluations	Problems
Zhang et al. [13]	Modify the logarithmic Rectified Linear Unit (L_ReLU) of the activation function.	LeNet-5	ReLU	The challenges include high hardware requirements, large training sample size, extended training time, slow convergence speed, and low accuracy.
Fu'adah et al. [14]	Automated classification system for images using AlexNet.	AlexNet	Adam, binary cross-entropy	AlexNet architecture was employed to develop an automated object classification system for Alzheimer's disease.

Table 4. *Cont.*

Approach	Features	Model	Evaluations	Problems
Tammina [15]	Classification, regression, and clustering.	VGG-16	Binary cross-entropy	Image classification problem with the restriction of having a small number of training samples per category.
Butt et al. [16]	Street crime snatching and theft detection in video mining.	VGG-19	ReLU, softmax	The meteoric growth of the Internet has made mining and extracting valuable patterns from a large dataset challenging.
Kieffer et al. [17]	Classification.	Inception	Linear, SVM	The task involves retrieving and classifying histopathological images to analyze diagnostic pathology.
Singla et al. [18]	Food image classification.	GoogleNet	Binary classification	Food image classification and recognition are crucial steps for dietary assessment.
Kuttiyappan et al. [19]	Hierarchical network feature extraction.	ResNet	Adam	Improving the cybersecurity of the bank sector by proving malicious attacks using the wrapper step-wise.
Hidayatuloh et al. [20]	Detection and diagnosis of plant diseases.	SqueezeNet	Adam	Identify the types of diseases on the leaves of tomato plants and their healthy leaves.
Li [21]	Image semantic segmentation.	Enet	MioU	Improve the network model of the generative adversarial network.
Chen et al. [22]	Garbage classification.	ShuffleNet	Cross entropy, SGD	Improve the consistency, stability, and sanitary conditions for garbage classification.
Zhang et al. [23]	Multiple features reweight DenseNet.	DenseNet	SGD	Adaptively recalibrating the channel-wise feature and explicitly modeling the interdependence between the features of different convolutional layers.

3. Video Classification

This section presents a comprehensive and significant review of deep learning models for video classification tasks. This section also covers all steps and details the processes involved in video classification tasks, including the modalities of video data, parallel processing approaches, breakthroughs in video classification, and recent state-of-the-art deep learning models for video classification tasks.

A significant amount of research has focused on enhancing video classification and object segmentation accuracy and efficiency in surveillance systems. For instance, Zhao et al. [24] proposed a real-time approach for moving object segmentation and classification from HEVC-compressed surveillance videos. Their paper focuses on utilizing the unique features extracted directly from the HEVC-compressed domain for video surveillance tasks. They introduced a method that involves pre-processing steps such as motion vector interpolation and outlier removal, clustering blocks with nonzero motion vectors into connected foreground regions, and applying object region tracking based on temporal consistency to refine the moving object boundaries. Additionally, they developed a person–vehicle classification model using a bag of spatial–temporal HEVC syntax words to accurately classify moving objects as persons or vehicles and provide solid performance in accurately segmenting and classifying moving persons and vehicles in real-time scenarios.

3.1. Fundamental Deep Learning Architecture for Video Classification

Video classification fundamentally derives from image classification, with individual frames combined to form a video. Significant advancements in video classification have been driven by the development of deep learning architectures designed to automatically learn and extract features from raw video data [25]. These architectures typically rely on supervised learning, where models are trained on labeled datasets to recognize patterns and make accurate predictions. CNNs have become a cornerstone of video classification tasks among the various deep learning models. CNNs are particularly effective at capturing spatial features within individual video frames through their layered structure of convolutional filters [26].

Extending CNNs to 3D convolutions or combining them with other models, such as RNNs or attention mechanisms, can also capture temporal dynamics, improving the model's ability to understand complex motion patterns and video semantic content [27]. Recent advancements have introduced vision transformers (ViTs), which leverage transformer architecture to process video data by treating video frames as sequences of image patches. ViTs have shown promising results in capturing spatial and temporal features, providing a powerful alternative to traditional CNN-based approaches [28]. The second approach is hybrid models, as shown in Figure 3, which integrate CNNs with various advanced techniques such as convolutional recurrent neural networks (CRNNs), two-stream networks, 3D CNNs, attention mechanisms, combinations of CNNs with transformer-based architectures, multi-modal fusion, and ensemble models, as summarized in this paper.

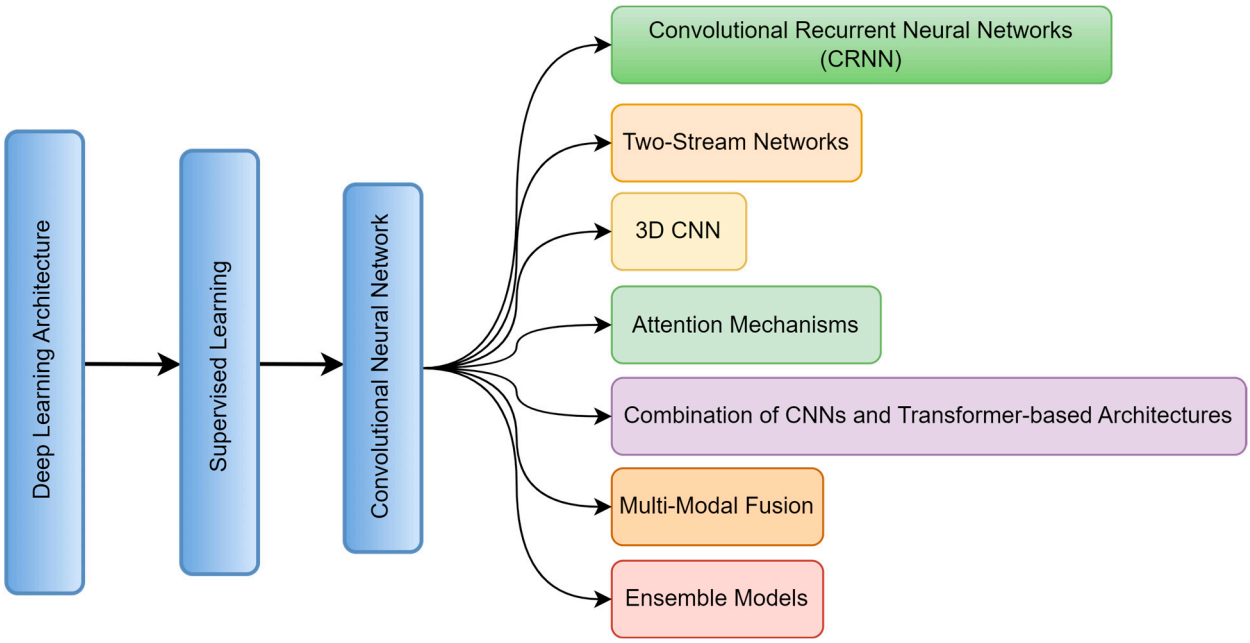

Figure 3. The integration of the CNN model with other models.

3.2. Parallel Processing in Video Classification

Video classification is computationally intensive due to the large volume of data and the complexity of analyzing both spatial and temporal dimensions. Parallel processing addresses these challenges by breaking down the video classification task into smaller, independent units that can be processed concurrently using multiple threads or computed units [29]. This approach significantly improves efficiency, reduces processing times, and enables effective parallelism in video data's spatial and temporal domains.

In the context of 3D-CNN architectures, parallel processing enhances performance by decomposing the classification task into multiple sub-tasks. This decomposition is especially beneficial for hybrid models, which combine various neural network architectures such as CRNNs, two-stream networks, attention mechanisms, and combinations of CNNs

and transformer-based architectures. These hybrid models also incorporate multi-modal fusion and ensemble methods to improve classification accuracy and robustness.

Each sub-task in these hybrid models is processed independently using 3D-CNNs. For instance, convolutional networks can handle spatial feature extraction, while recurrent networks manage temporal dependencies [30]. Attention mechanisms can further refine the focus on relevant video segments, and transformer-based architectures can capture long-term dependencies across video frames. Table 5 below summarizes several studies that utilize hybrid models for video classification, highlighting their architectures, features, and operational mechanisms.

Table 5. Studies on video classification using hybrid models.

Approach	Architecture	Features	Operational Mechanism
Li [31]	3D-CNN	Multi-class, temporally downsampled, increment of the new class.	Utilizing each 3D-CNN as a binary classifier for a distinct video class streamlines training and decreases computational overhead.
Jing et al. [32]	Semi-supervised	Supervisory signals extracted from unlabeled data, 2D images for semi-supervised learning of 3D video clips.	Three loss functions are employed to optimize the 3D network: video cross-entropy loss on labeled data, pseudo-cross-entropy loss on unlabeled data's pseudo-labels, and soft cross-entropy loss on both labeled and unlabeled data to capture appearance features.
Wu et al. [33]	Multi-Stream, ConvNets	Multi-stream, multi-class.	Effectively recognizes video semantics with precise and discriminative appearance characteristics; motion stream traina ConvNet model operates on stacked optical flows.
Yue-Hei Ng et al. [34]	Convolutional temporal	CNN feature computation, feature aggregation.	Pooling feature methods that were max-pooling local information through time and LSTM, whose hidden state evolves with each sequential frame.
Wu et al. [35]	Short-term motion	Short-term spatial–motion patterns, long-term temporal clues.	Extracts spatial and motion features with two CNNs trained on static frames and stacked optical flow.
Tavakolian et al. [36]	Heterogeneous Deep Discriminative Model (HDDM)	Unsupervised pre-training, redundancy-adjacent frames, spatio-temporal variation patterns.	HDDM weights are initialized by an unsupervised layer-wise pre-training stage using Gaussian Restricted Boltzmann Machines (GRBM)
Liu et al. [37]	Simple Recurrent Units method (SRU)	Feature extraction, feature fusion, and similarity measurement.	SRU network can obtain the overall characterization of video features to a certain extent through average pooling.
Varadarajan et al. [38]	Max Calibration mixtuRe of Experts (MiCRObE)	Hand-crafted.	MiCRObE can be used as a frame-level classification that does not require human-selected and frame-level ground truth.

Table 5. *Cont.*

Approach	Architecture	Features	Operational Mechanism
Mihanpour et al. [39]	Deep bidirectional LSTM (DB-LSTM)	Frame extraction, forward and backward passes of DB-LSTM.	The DB-LSTM recurrent network is used in forward and backward transitions, and the final classification is performed.
Jiang et al. [40]	Hybrid deep learning	Multi-model clues, static spatial motion patterns.	Integrating a comprehensive set of multi-modal clues for video categorization by employing three independent CNN models: one operating on static frames, another on stacked optical flow images, and the third on audio spectrograms to compute spatial, motion, and audio features.

Figure 4 illustrates the parallel processing framework with hybrid models for video classification, which is particularly effective for large-scale video datasets. The input video is divided into frames and processed in parallel. After frame processing, hybrid models train the data, resulting in class scores or predictions for the objects within each frame.

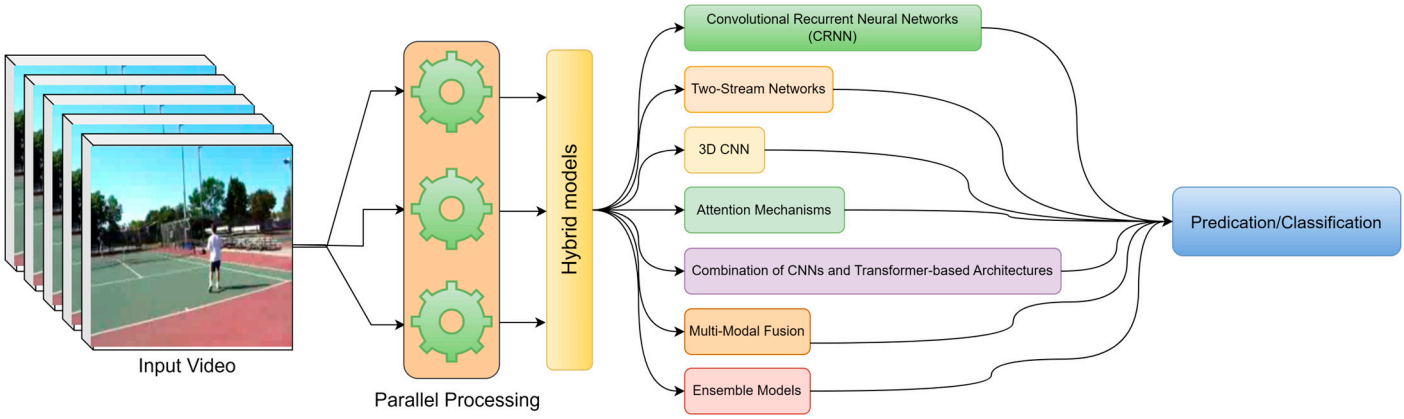

Figure 4. Overview of the video classification process, including parallel processing with hybrid models.

3.3. The Methods Used in Video Classification

This section describes the difference between the video classification methods used to label a short video clip and the human activity performed in the video clip. We describe the difference between single frame, late fusion, early fusion, and slow fusion, as shown in Figure 5. Single-frame video classification involves breaking down a video into individual frames and treating each frame as an independent image for classification. Features are extracted from each frame using techniques such as CNNs, and then a classifier predicts the category or label of each frame. While this approach simplifies processing and is easy to implement, it overlooks crucial temporal information for understanding video context [26]. It is adequate for tasks with less critical temporal details and limited computational resources, but it needs help with dynamic videos requiring a deeper understanding of temporal relationships.

Late fusion for video classification is a powerful technique that offers distinct advantages. Unlike other methods, it does not fuse information at the feature level. Instead, it aggregates predictions from individual classifiers, and the models are trained on different input modalities, such as visual, audio, or textual data [41]. These individual predictions are combined using techniques like averaging, voting, or weighted averaging to produce a final classification result. Late fusion is advantageous because it allows each modality to be processed independently, enabling the use of specialized models for each modality. This

approach is efficient when different modalities contain complementary information about the video content.

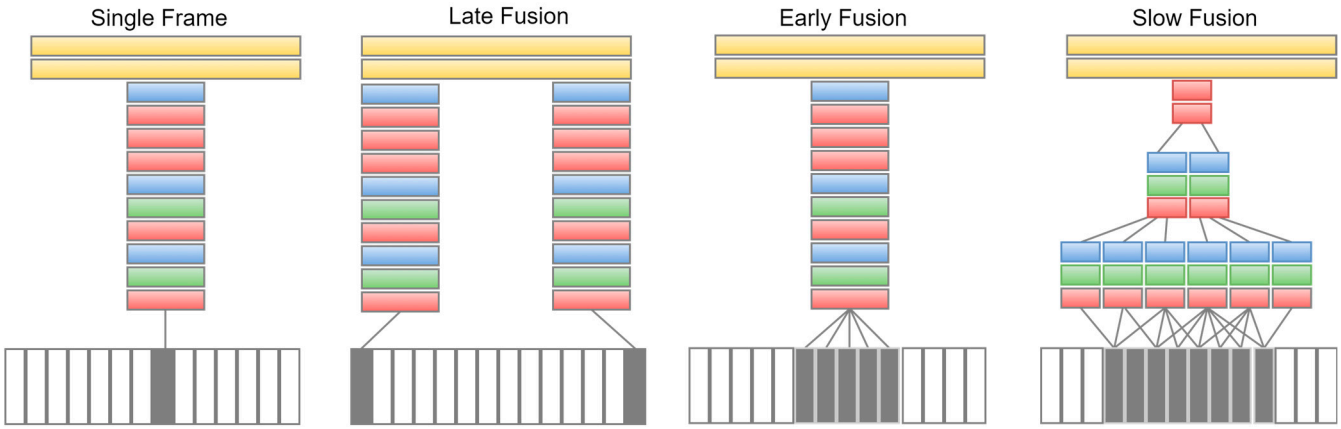

Figure 5. Different strategies for video classification tasks: single-frame, late-fusion, early-fusion, and slow-fusion neural networks.

Early fusion for video classification is a unique approach that involves combining information from different modalities and sources early in the classification process. Instead of processing each modality independently and combining their predictions later, early fusion merges the features extracted from different modalities into a single representation before passing it through a classifier [35]. This combined representation captures visual, audio, and possibly textual information in a unified feature space. Early fusion is helpful because it allows for the joint learning of features across modalities, potentially enhancing a model's ability to capture complex relationships in the data.

Slow fusion for video classification involves gradually integrating information from different sources and modalities to improve classification accuracy. Instead of combining information at a single stage, slow fusion aggregates features and predictions from different modalities across multiple time steps or frames. This approach allows the model to capture the video sequence's temporal dependencies and contextual information. Slow fusion typically involves techniques such as RNN and LSTM networks to integrate information over time [42]. Slow fusion is helpful because of the capability to capture the video's spatial and temporal characteristics, which enables more robust and nuanced classification.

3.4. Hybrid Models for Video Classification

In this section, we summarize the hybrid models for video classification that leverage various neural network architectures, and the techniques proposed to improve the accuracy of video classification systems, as shown in Figure 6. These hybrid models combine different approaches to effectively capture spatial and temporal features in videos, enhancing the model's ability to understand complex motion patterns and semantic content. In video classification, the spatial and temporal streams are two different approaches to capturing information from video data. The spatial stream focuses on individual frames, extracting features such as objects and scenes using CNNs. This stream captures the static-appearance information from the video.

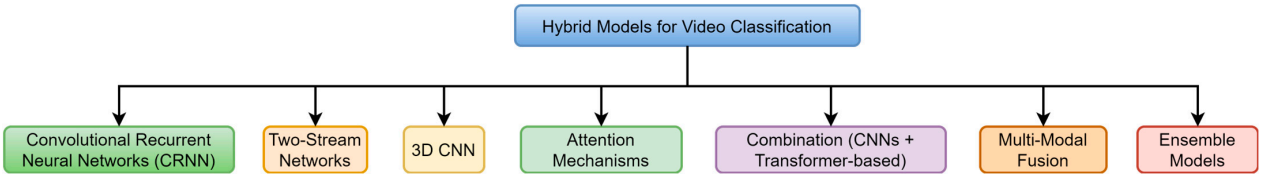

Figure 6. The hybrid learning model approach is used in video classification.

On the other hand, the temporal stream analyzes the motion between frames, capturing dynamic information such as movement and actions. This technique uses methods like optical flow or RNNs to understand the temporal evolution of the video. By combining these two streams, models can achieve a more comprehensive understanding of the appearance and motion within the video, leading to improved classification performance.

The choice of model depends on several factors, including the complexity of the classification task, the availability of computational resources, and the characteristics of the input data [35]. By carefully selecting and combining different neural network architectures and techniques, practitioners can tailor hybrid models to suit specific video classification challenges, leading to more accurate and reliable results. Integrating transformer models with CNNs has become a powerful approach to enhance further spatial and temporal feature modeling in video classification tasks. CNNs are adept at extracting detailed spatial features from individual frames, making them highly effective for recognizing objects and identifying features within a static context. However, video classification requires understanding the content of individual frames and capturing the temporal dynamics across sequences of frames.

Transformers excel in modeling long-range dependencies and capturing temporal relationships due to their attention mechanisms, which allow them to focus on relevant parts of the input sequence over time. Combining these strengths, hybrid models leverage CNNs to process the spatial information within each frame and transformers to understand the temporal evolution across frames. This fusion facilitates comprehensive spatio-temporal feature extraction, enabling more accurate and robust video classification. This technique, which includes fine-tuning pre-trained models, is commonly employed to integrate these architectures, leveraging transfer learning to enhance performance and reduce training time. Recent research has shown that this hybrid approach significantly improves the accuracy of video classification models by effectively addressing the complexities of spatial and temporal feature extraction. Details of the differences between the strengths and weaknesses of these models are shown in Table 6.

Abdullah et al. [43] proposed a hybrid deep learning model tailored for video classification tasks, specifically a CNN-RNN system. The proposed approach involves utilizing a single-layer LSTM network to learn temporal features from stacked spatial features extracted by a CNN for each video instance. Notably, instead of employing general-purpose object classifiers, such as pre-trained CNNs trained on ImageNet, they opt to use a CNN model specialized in extracting emotion features. The hybrid model design aims to leverage the strengths of both CNNs and RNNs, enhancing the system's ability to classify videos accurately by modeling spatial and temporal information and capturing nuanced emotional and motion-related features in video facial expression recognition tasks.

Feichtenhofer et al. [44] introduced CNNs for human action recognition in videos, presenting various methods to effectively integrate appearance and motion information. Despite its advantages, the two-stream architecture has encountered two primary limitations. Firstly, pixel-wise correspondences between spatial and temporal features need learning. Secondly, the architecture faces constraints at a temporal scale, given that spatial CNNs operate solely on individual frames, while temporal CNNs operate on temporal sequences independently. These limitations degrade the model's ability to capture fine-grained spatio-temporal relationships, potentially impacting its performance in accurately recognizing human actions in videos.

Fan et al. [45] present a video-based emotion-recognition system centered around a hybrid network as a core module. This hybrid network combines RNN and 3D convolutional networks (C3D) in a late-fusion manner. The RNN and C3D are tasked with encoding appearance and motion information differently within the system. Specifically, the RNN receives appearance features extracted by a CNN from individual video frames as input, and subsequently encodes the motion information. On the other hand, C3D simultaneously models both the appearance and motion aspects of the video. This hybrid approach aims to leverage the complementary strengths of RNNs and C3Ds to effectively capture appearance

and motion features, enhancing the system's ability to recognize emotions accurately from video data.

Li et al. [46] proposed a hybrid attention model (HAM) within a deep CNN framework for image classification tasks. The model utilizes an intermediate feature map as its input by integrating a HAM. The HAM operates in two main steps: (1) generating a single-channel attention map and a refined feature map through the channel submodel; (2) then, based on this attention map, the spatial submodel divides the refined feature map into two groups along the channel axis. This division process facilitates the creation of a pair of spatial attention descriptors, providing enhanced focus on specific regions within the image. The HAM aims to improve the discriminative power of the model by dynamically attending to relevant spatial features, thereby enhancing its performance in image classification tasks.

Mekruksavanich et al. [47] introduced a novel approach to human activity recognition using multiple CNNs. The proposed method creates an effective architecture by incorporating a hybrid CNN with a channel attention mechanism. This mechanism enhances the network's ability to capture intricate spatio-temporal characteristics hierarchically, enabling it to discern between various human movements in daily activities. By leveraging the channel attention mechanism, the model dynamically focuses on relevant features, improving its capacity to recognize and differentiate between different human activities more accurately.

Ullah et al. [48] present an innovative end-to-end hybrid framework for anomaly detection, integrating a CNN with a vision transformer-based architecture. This framework analyzes surveillance videos and identifies anomalous events using spatial and temporal information. The model processes the spatial and temporal features extracted from the surveillance video in the first step. Then, in the second step, these features are passed through the vision transformer-based model to capture the long-term temporal relationships among various complex surveillance events. The features obtained from the backbone model are subsequently inputted into a sequential learning model, where temporal self-attention mechanisms are employed to generate an attention map. This attention map highlights relevant spatio-temporal regions within the video, aiding in accurately detecting anomalous activities. Xu et al. [49] proposed a multi-modal emotional classification framework to capture user emotions in social networks. This framework features a 3D convolutional–LSTM (3DCLSTM) model, a hybrid model for classifying visual emotions, and a CNN-RNN hybrid model for classifying text-based emotions. The hybrid multi-modal fusion approach leverages the benefits of both feature fusion and decision-layer fusion strategies, effectively overcoming their limitations. This approach uses C3D to learn the spatio-temporal characteristics inputted into the convolutional LSTM model during training and classification. This combination enables the model to accurately capture and classify complex emotional cues from visual and textual data.

Jagannathan et al. [50] proposed automatic vehicle classification, which plays a vital role in intelligent transportation and visual traffic surveillance systems. Adaptive histogram equalization and a Gaussian mixture model are implemented to enhance the quality of the collected vehicle images and detect vehicles from denoised images. The extracted features are then used as the input for an ensemble deep learning technique for vehicle classification.

Table 6. Summary of strengths and weaknesses of hybrid models for video classification.

Hybrid Model	Strengths	Weaknesses
CNN-RNN [43]	Combines spatial features with temporal dynamics; specialized CNN for emotional features.	Complexity in training; high computational resources.
Two-stream architecture [44]	Effectively integrates appearance and motion; separate processing for spatial and temporal data.	Pixel-wise correspondences between spatial and temporal features need learning; constraints on temporal scale.

Table 6. *Cont.*

Hybrid Model	Strengths	Weaknesses
Hybrid network (RNN + 3D CNN) [45]	Late-fusion of RNN and C3D; encodes appearance and motion separately.	Increased model complexity; potential for overfitting with limited data.
HAM [46]	Dynamically attends to relevant spatial features and improves discriminative power.	Requires significant tuning and potentially high computational cost.
Hybrid CNN with channel attention mechanism [47]	Captures intricate spatio-temporal characteristics; dynamic focus on relevant features.	High complexity; requires substantial computational resources.
CNN + vision transformer [48]	Long-term temporal relationship modeling; effective for anomaly detection in surveillance.	High computational and memory requirements; complex architecture.
3DCLSTM + CNN-RNN [48]	Effective multi-modal fusion; captures spatio-temporal and textual emotional cues.	Very high complexity; demanding computational resources.
Ensemble models [50]	Enhanced image quality; accurate vehicle classification.	Requires extensive pre-processing; high computational cost.

3.5. Challenges and Future Directions of Video Classification

In this section, we cover the challenges faced in video classification tasks and summarize future directions to improve performance, including advancements in hardware, improved algorithms, and the use of transfer learning models.

3.5.1. Challenges in Video Classification

This section presents several benchmark datasets commonly used for video classification tasks, providing a comprehensive overview of their characteristics and challenges. We present various aspects of these datasets, including data complexity, computational resource requirements, and real-time processing capabilities. The complexity of video data, the substantial computational power needed to process large-scale datasets, and the necessity for efficient algorithms to handle real-time processing are significant challenges in the video classification domain. Table 7 summarizes some notable datasets, detailing the dataset name, the total number of videos, the resolutions of the videos, the number of classes present, and the year of publication. This summary helps illustrate the diverse range of datasets available, each with unique features and applications. These datasets serve as critical benchmarks for developing and evaluating video classification models, providing a standard for comparing performance across different approaches.

Table 7. Dataset benchmarks for video classification.

Datasets	# of Videos	Resolutions	# of Classes	Year
KTH	2.391	160 × 120	6	2004
Weizmann	81	180 × 144	9	2005
Kodak	1.358	768 × 512	25	2007
Hollywood	430	400 × 300, 300 × 200	8	2008
YouTube Celebrities Face	1.910	-	47	2008
Hollywood2	1.787	400 × 300, 300 × 200	12	2009
UCF11	1.600	720 × 480	1600	2009
UCF sports	150	720 × 480	10	2009
MCG-WEBV	234.414	-	15	2009
Olympic Sports	800	90 × 120	16	2010

Table 7. *Cont.*

Datasets	# of Videos	Resolutions	# of Classes	Year
HMDB51	6.766	320 × 240	51	2011
CCV	9.317	-	20	2011
JHMDB	960	-	21	2011
UCF-101	133.20	320 × 240	101	2012
THUMOS 2014	183.94	-	101	2014
MED-2014 (Dev. set)	31.000	-	20	2014
Sports-1M	113.3158	320–240	487	2014
MPII Human Pose	25 K	-	410	2014
ActivityNet	279.01	1280 × 720	203	2015
EventNet	953.21	-	500	2015
FCVID	912.23	-	239	2015
Kinetics	650.000	-	400, 600, 700	2017
Something-something V1	110.000	100 × (~)	174	2017
YouTube-8M	6.1 M	-	3862	2018
Moments in Time	802.264	340 × 256	339	2018
EPIC-KITCHENS	396 K	-	149	2018
Charades-Ego	685.36	-	157	2019
AVA-Kinetics	230 k	-	80	2020
Something-something V2	220.847	-	174	2021

3.5.2. Future Directions of Video Classification

In video-processing tasks, hardware components are critical for enhancing time performance during training on large-scale datasets. For example, Kyrkou et al. [51] proposed an SVM cascade architecture implemented on a Spartan-6 field-programmable gate array (FPGA) platform. They evaluated it for object detection on 800 × 600 (super video graphics array) resolution images. The proposed system, enhanced by a neural network capable of processing cascade information, achieves a real-time processing rate of 40 frames per second in the face detection benchmark application. Karpathy et al. [26] proposed large-scale video classification with CNNs. Since CNNs typically take weeks to train on large-scale datasets, even on the fastest available GPUs, this approach to speeding up the networks reduces the number of layers and neurons in each layer and the resolution of images during the training phase. This approach achieved good performance in terms of training time. Pérez et al. [52] present a scalable, low-power, low-resource-utilization accelerator architecture for inference on the MobileNet V2 CNN. The architecture uses a different system with an embedded processor as the central controller, external memory to store network data, and dedicated hardware implemented on reconfigurable logic with scalable processing elements.

Developers increasingly use FPGAs in image-processing applications, especially for real-time tasks like vehicle number plate recognition, video classification, and medical imaging (e.g., X-rays and CT scans). An FPGA embedded with a camera can perform image processing rapidly, as the images are streamed rather than processed as a sequence of individual frames. FPGAs allow for the custom hardware design of algorithms using hardware description languages optimized for performance and resource efficiency [53]. This process entails the utilization of algorithms by deploying a collection of application-specific intellectual property cores sourced from a library. High-level synthesis is employed to transform a C-based representation of the algorithm into synthesizable hardware. Subsequently, the system maps the algorithm onto a parallel array of programmable soft-core processors.

Using the FPGA for video classification begins with capturing video frames using sensors. An acquisition device initially processes these frames and then transfers them to memory, where resources can be shared with the CPU/GPU. Finally, the FPGA processes the data, leveraging its memory for efficient video classification. The image processing method using the FPGA library is shown in Figure 7.

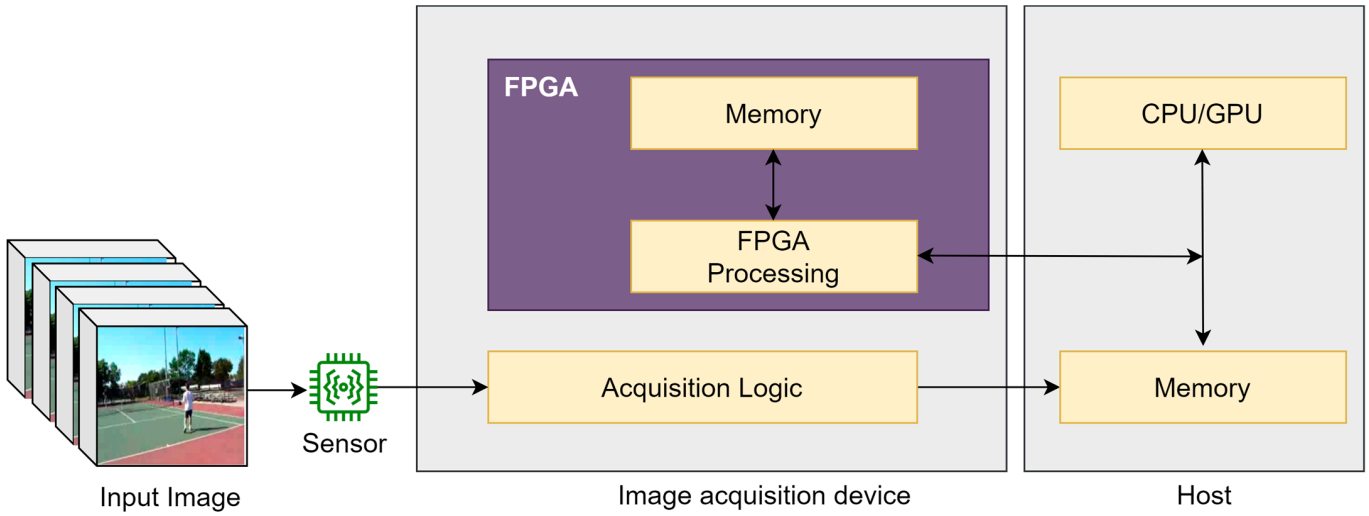

Figure 7. Image processing using FPGA.

Given these capabilities, how can FPGAs be optimized further for video classification using deep learning techniques? What are this approach's potential advantages and limitations, and how can it be leveraged to enhance the accuracy and efficiency of real-time video-processing applications? All these questions are open research topics for researchers and industries to consider in future work to improve the parallel processing of large datasets via video classification tasks and real-time video classification applications.

3.6. Overview of Deep Learning Frameworks and Hybrid Models for Video Classification

This section describes two different approaches for video classification: single and hybrid deep learning. Single deep learning entails the utilization of a singular neural network architecture for classification tasks. This method utilizes the potency of deep neural networks to learn features and patterns directly from raw video data autonomously. The primary advantages include simplicity in design and implementation and the capacity to train end-to-end models without extensive manual feature engineering.

In contrast, hybrid deep learning combines multiple neural network architectures and integrates traditional machine learning techniques with deep learning models to enhance performance. This approach capitalizes on the strengths of various models, such as combining CNNs for spatial feature extraction with RNNs for temporal sequence modeling. Hybrid models can further incorporate pre-processing steps, feature extraction techniques, and ensemble methods to augment accuracy and robustness. The principal advantages of hybrid deep learning encompass heightened flexibility, superior performance on intricate tasks, and the potential to reduce labeled data requirements through techniques like transfer learning.

Table 8 summarizes the difference between single deep learning and hybrid deep learning models. These frameworks adeptly manage the complexities of vast volumes of high-dimensional video data. Their flexibility facilitates the creation of custom neural network architectures customized for video classification tasks, empowering researchers and developers to explore various model configurations to achieve optimal results. Within deep learning frameworks, three primary processes contribute to evaluating video classification tasks: (1) data augmentation, (2) the utilization of pre-trained models, and (3) the exploration of diverse model architectures. Each process plays a pivotal role in enhancing performance across video classification tasks.

Table 8. Comparison between single and hybrid deep learning.

Attributes	Single Deep Learning	Hybrid Deep Learning
Applications	Limited application	Diverse application
Classifier diversity	Limited to Softmax	Softmax or ML-based
Feature extraction	Limited scope	Larger scope
Hardware resource	Uses little resources	Uses more resources
Performance evaluation	Low performance	Superior performance
Program complexity	Low complexity	High complexity
Transfer learning	Limited options for transfer learning	More options for transfer learning

3.6.1. Data Augmentation

The initial stage of data augmentation is pivotal for video classification tasks, as it marks the first step in building a robust video classification system. This stage is a foundational component for enhancing accuracy as the process progresses. Mao et al. [54] propose various techniques to improve data augmentation. They apply a random rotation of up to 70 degrees to the video frames and a random shift of 224 × 224 pixels. Similarly, Takahashi et al. [55] apply random image cropping and patching for deep CNNs to improve the accuracy of model training. Kim et al. [56] propose using brightness adjustment techniques to enhance training, and involve representative colors using a power function to improve the training process.

Furthermore, Taylor et al. [57] introduce methods to strengthen deep learning through generic data augmentation by applying exposure techniques during image classification training to achieve high performance. On the contrary, Sayed et al. [58] propose an improved method for handling motion blur in online object detection that yields promising results for object detection tasks. Additionally, Kim et al. [59] introduced a data augmentation technique utilizing the adaptive inverse peak signal-to-noise ratio they devised while examining the impact of color attributes within the training images.

Diba et al. [60] proposed a novel noise filtration algorithm for denoising neuromorphic camera data using a graph neural network (GNN)-driven transformer approach. The key innovation in their work is the introduction of the GNN–transformer algorithm, which classifies active event pixels in raw streams into real log-intensity variations or noise. They utilized a message-passing framework called EventConv within the GNN to capture spatiotemporal correlations among events while maintaining their asynchronous nature. Additionally, they introduced the known-object ground-truth labeling (KoGTL) approach to generate labeled datasets under various lighting conditions, including moonlight, to train and test their algorithm. The proposed algorithm outperformed state-of-the-art methods by at least 8.8% in terms of filtration accuracy when tested on unseen datasets. To augment the data for training and testing, they conducted experiments in various lighting conditions, including very good lighting, office lighting, low light, and moonlight, recording scenes with static and moving cameras in different directions to capture dynamic event and noise-generation scenarios.

3.6.2. Pre-Training on Hybrid Models

The visual applications of deep learning include object detection, visual object recognition, image segmentation, and more. The pre-training step involves training a deep neural network on a vast amount of labeled video data, typically employing techniques like supervised learning. This process enables a model to ascertain general features and patterns inherent in various types of videos, such as motion, shapes, and textures. Pre-trained models undergo fine-tuning for specific video classification tasks. Fine-tuning entails adjusting the pre-trained model's parameters to better align with the characteristics of the target dataset or system. This step often requires less labeled data than training a model from scratch, rendering it more feasible for video classification tasks with limited annotated samples.

Pre-trained models for video classification commonly rely on CNNs. Among the prevalent methods are those based on CNNs, including two primary approaches: (1) RNNs with LSTM networks, and (2) optical flow, as shown in Figure 8. These models incorporate architectures such as 3D CNNs to capture spatio-temporal features from video sequences directly. Ramesh et al. [61] proposed utilizing pre-trained CNNs, such as AlexNet, GoogLeNet, and MobileNet, to enhance video classification accuracy and improve sports training and performance. To construct a custom video-style classifier, Aryal et al. [62] presented several pre-trained models, such as VGG16, InceptionV3, and ResNet50. They utilized Keras, a powerful deep learning library that implements these deep models, to build the classifiers.

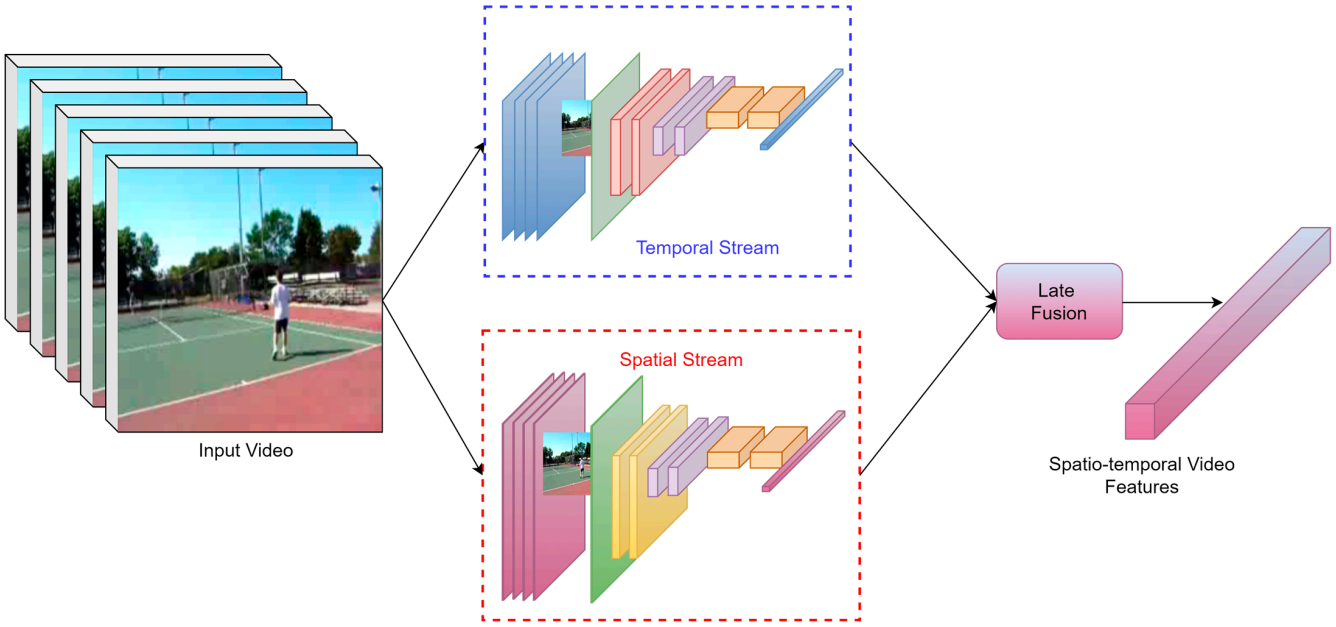

Figure 8. Two-stream architecture with optical flow for video classification.

Multi-stream models are built on the idea of separating the temporal stream and the spatial stream. The temporal stream in action recognition provides significant advantages by capturing motion information that static images cannot transfer. This stream analyzes the sequence and flow of frames over time, allowing the system to understand the dynamics and transitions of actions. By focusing on temporal changes, the stream can detect subtle movements and patterns, improving the accuracy of recognizing complex activities, such as distinguishing between similar actions that differ in speed or direction. The temporal stream enhances the robustness of action recognition systems in varying conditions, as it can better account for changes in perspective, lighting, and occlusions that might affect single-frame analysis.

The spatial stream in action recognition offers significant advantages by capturing detailed spatial information from individual frames, which is crucial for understanding the context and environment of an action. This stream focuses on the appearance and layout of objects and scenes, allowing the system to identify essential features such as shapes, textures, and edges. By analyzing these spatial characteristics, the spatial stream enhances the system's ability to recognize and differentiate between various objects and their configurations within a frame. It is particularly beneficial for identifying the context in which an action occurs, such as distinguishing between different sports or activities based on surroundings.

Additionally, another approach to the primary model involves combining CNNs with transformer architectures for video classification on pre-trained models, representing a novel approach that leverages the respective advantages of the architectures [63]. CNNs are adept at extracting spatial features from individual video frames, making them ideal

for object recognition and feature extraction tasks. Meanwhile, transformers are excellent at capturing long-range dependencies and temporal relationships within sequential data, leveraging attention mechanisms to focus on the relevant parts of the input sequence. By integrating CNNs for spatial feature extraction and transformers for temporal modeling, the combined model can effectively analyze the visual content of each frame and the temporal dynamics across frames.

This comprehensive understanding of spatial and temporal information enables more accurate and robust video classification. Whether through joint training or fine-tuning pre-trained models, this approach leverages transfer learning to improve performance while reducing training time. Selva et al. [28] introduced a transformer architecture modification to efficiently handle video data, reduce redundancy, reintroduce functional inductive biases, and capture long-term temporal dynamics for video classification tasks. As depicted in Figure 9, the sequence length M depends on the data sample, N is the desired number of sequences, and FC denotes the number of channels.

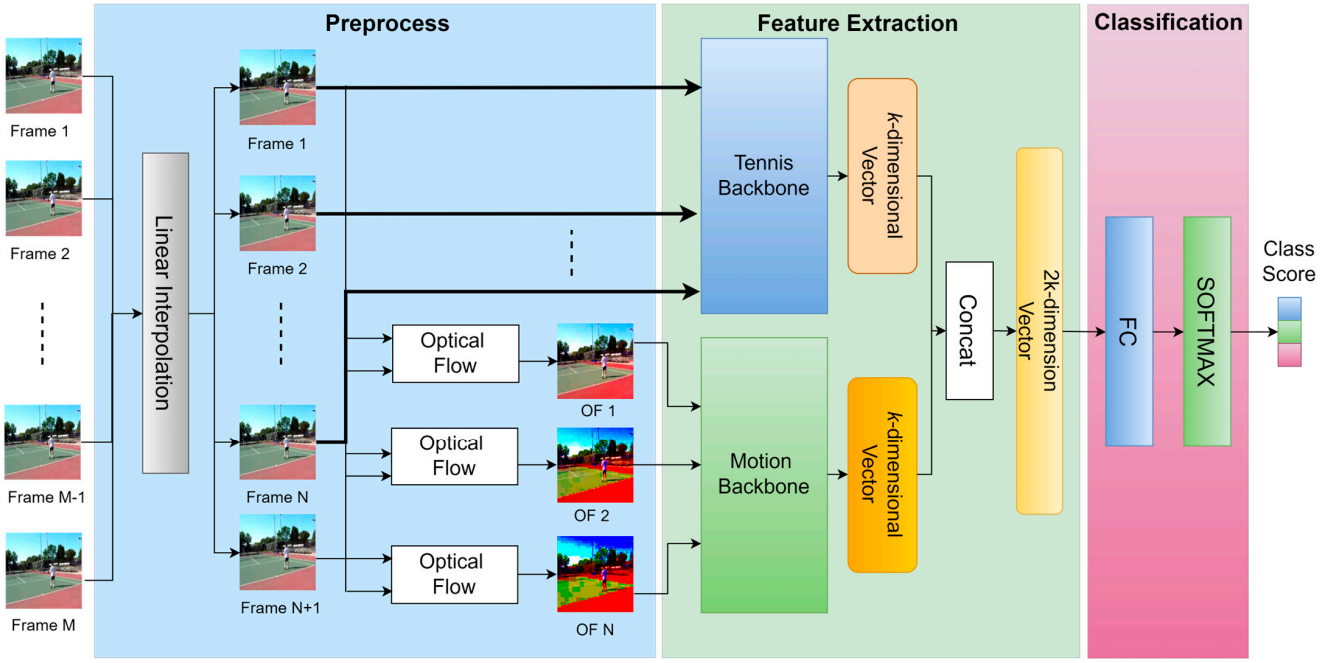

Figure 9. The architecture of the transformer model for extracting frames from a video.

3.6.3. Hybrid Approach to Training Models on Video Classification

The evolution of models for video classification using the hybrid approach showcases remarkable progression. Table 9 covers various aspects of these hybrid models, including the authors' proposed approach, the features, the models employed, the datasets used for the experiments, the problems with the approach, the results/findings, and the year of publication. These studies focus on hybrid architectures, particularly those combining CNN with CRNNs, two-stream networks, attention mechanisms, combinations of CNNs and transformer-based architectures, multi-modal fusion, and ensemble models. Classification accuracy is a standard evaluation metric that frequently uses the UCF-101 dataset for experimentation. Multi-classification problems are addressed in almost all cases. The superior performance of 3D convolutions for spatio-temporal feature learning is noted. Various fusion techniques such as average fusion, kernel average fusion, weight fusion, logistic regression fusion, and multiple kernel learning fusion are generally less effective than the multi-stream, multi-class fusion technique using hybrid architectures in video classification tasks.

Table 9. Video classification training with the hybrid approach.

Approach	Features	Model Architectures	Datasets Used	Problem	Results/Findings	Year
Wu et al. [35]	Spatial, short-term motion.	CNN + LSTM	UCF-101, CCV	Content semantics	UCF-101: 91.3%, CCV: 83.5%.	2015
Jiang et al. [40]	Corresponding features, motion features, and multi-modal features.	CNN + LSTM	UCF-101, CCV	Multi-modal clues	UCF-101: 93.1%, CCV: 84.5%.	2018
De Souza et al. [64]	Spatio-temporal features.	FV-SVM	UCF-101, HMDB-51	Content of video	UCF-101: 90.6%, HMDB-51: 67.8%.	2018
Jaouedi et al. [65]	Spatial features: reduce the size of the data processed; various object features.	GMM + KF + GRNN	UCF Sport, UCF-101, KTH	Facilitate clues	KTH: 96.30%	2020
Zuo et al. [9]	Fuzzy local spatio-temporal descriptors	HFLSTD + iFDT	UCF-50, UCF-101	Uncertainty in pixel voting due to varying numbers of bins	UCF-50: 95.4%, UCF-101: 97.3%	2020
Wu et al. [33]	Multi-modal	CNN + LSTM	UCF-101, CCV	Multi-stream	UCF-101: 92.2%, CCV: 84.9%	2016
Kumaran et al. [66]	Latent features, spatio-temporal features.	CNN-VAE	T15, QMUL, 4WAY	Classify the times series	T15: 99.0%, QMUL: 97.3%, 4WAY: 99.5	2018
Ijjina et al. [67]	Action bank features.	Hybrid deep neural network + CNN	UCF50	-	UCF50: 99.68%	2015
De Souza et al. [68]	Hand-crafted, spatio-temporal features.	iDT + STA + DAFS + DN	UCF-101, HMDB-51, Hollywood2, High-Five, Olympics	Large video data	UCF-101: 90.6%, HMDB-51: 67.8%, Hollywood2: 69.1%, High-Five: 71.0%, Olympics: 92.8%.	2016
Lei et al. [69]	High-level features, robust action features.	CNN-HMM	Weizmann, KTH	Complex temporal dynamics	Weizmann: 89.2%, KTH: 93.97%	2016
Dash et al. [70]	Sophisticated hand-crafted motion features.	SIFT-CNN	UCF, KTM	Action recognition	UCF: 89.5%, KTM: 90%.	2021

4. Evaluation Metrics and Comparison of Existing State-of-the-Art Video Classification Tasks

This section covers several benchmark datasets utilized for video classification tasks. The datasets for action recognition have evolved to become highly complex and realistic. The earliest datasets, such as KTH and Weizman, featured a fixed number of actions and minimal action categories. A list of datasets used for action recognition is given in Table 7. The details provided for these datasets include the dataset name, the total number of videos in each dataset, the resolution of the videos, the number of classes present in the dataset, and the year of publication of the dataset.

4.1. Performance Metrics for Evaluation in Video Classification

Throughout this section, we describe how the evaluation of video classification models is mostly achieved using different performance measures. The most common measures to evaluate models include accuracy, precision (precision measures include conducting positive meaning determination), recall (precision tests for the productive detection of the classifier's positive result), F1 score, micro-F1 score, and K-fold cross-validation. Several recent studies have utilized these measures, as shown in Table 10.

Table 10. List of performance metrics used in the video classification task.

Evaluation Metric	Year of Publication	Reference
Accuracy	2018	Xie et al. [71]
Precision	2014	Karpathy et al. [26]
Recall	2016	Abu-El-Haija et al. [4]
F1 score	2021	de Oliveira Lima et al. [42]
Micro-F1 score	2021	Moskalenko et al. [72]
K-Fold cross-validation	2021	Naik et al. [73]
Top-k	2015	Varadarajan et al. [38]

4.2. Comparison of Datasets for Video Classification

Various datasets have been curated for training and evaluating video classification models. Each dataset has unique characteristics, presents distinct challenges, and is suitable for specific tasks. In Table 11, we comprehensively compare these datasets, which serve as benchmarks for evaluating model performance, and give a summary with invaluable resources to train and test the different algorithms across various tasks and scenarios that match the dataset characteristics.

Table 11. Comparison of datasets for video classification.

Ref.	Dataset	Characteristics	Challenges	Suitability for Tasks
[74]	UCF-101	Consists of 101 action categories.	Limited diversity in activities and scenarios.	Basic action recognition.
[75]	HMDB-51	Contains videos from a diverse set of activities.	The limited number of samples per class.	Basic action recognition.
[76]	Kinetics-400	Large-scale dataset with 400 action classes.	Requires significant computation resources.	Complex action recognition.
[77]	ActivityNet	Contains untrimmed videos annotated with activities.	Temporal localization and annotation.	Activity detection and temporal action localization.
[78]	AVA	Focuses on human–object interactions in video.	Requires fine-grained action annotations.	Human–object interaction recognition.
[79]	Something-something v. 2	Addresses fine-grained action recognition with interventions involving everyday objects.	Limited in vocabulary and scale.	Fine-grained action recognition.

4.3. Comparison of some Existing Approaches on the UCF-101 Dataset

The UCF-101 dataset is a widely utilized benchmark in computer vision and video classification and was published by researchers from the University of Central Florida in 2012 [80]. It comprises 13,320 videos across 101 action categories, ranging from everyday human activities to sports and leisure activities. The videos are typically short, lasting a few seconds, and were captured from YouTube videos. The dataset offers various actions, including running, jumping, playing musical instruments, and more. UCF-101 is a valuable resource for training and evaluating video classification algorithms, aiding researchers in developing robust models for action recognition and related tasks. Table 12 displays existing papers and their comparative results using the UCF-101 dataset.

Table 12. Comparison of video classification methods using the UCF-101 dataset.

Ref.	Method	Accuracy (%)	Year of Publication
Varadarajan et al. [38]	OmniVec	99.6	2023
Wu et al. [81]	BIKE	98.9	2022
Li et al. [82]	ZeroI2V ViT-L/14	98.6	2023
Wu et al. [83]	Text4Vis	98.2	2022
Carreira et al. [84]	Two-Stram I3D	98.0	2017
Huang et al. [85]	BQN	97.6	2021
Zhang et al. [86]	CCS + TSN	97.4	2019
Tran et al. [87]	R [2 + 1]D-TwoStream	97.3	2017
Hong et al. [88]	Multi-stream I3D	97.2	2019
Zhao et al. [89]	AMD	97.1	2023
Sharir et al. [90]	STAM-32	97.0	2021
Zhu et al. [91]	FASTER32	96.9	2019
Qiu et al. [92]	LGD-3D Flow	96.8	2019
Zhang et al. [93]	VidTr-L	96.7	2021
Shou et al. [94]	I3D RGB + DMC-Net	96.5	2019
Chen et al. [95]	A2-Net	96.4	2018
Sun et al. [96]	Optical-Flow-Guided Feature	96.0	2017
Crasto et al. [97]	MARS + RGB + Flow	95.8	2019
Liu et al. [98]	Prob-Distill	95.7	2019
Carreira et al. [84]	RGB-I3D	95.6	2017
Tran et al. [87]	R[2 + 1]D-Flow	95.5	2017
Fan et al. [99]	TVNet + IDT	95.4	2018
Huang et al. [100]	TesNet	95.2	2020
Carreira et al. [84]	RGB-I3D	95.1	2017
Tran et al. [87]	R[2 + 1]D-TwoStream	95.0	2017
Christoph et al. [101]	ST-ResNet + IDT	94.6	2016
Liu et al. [102]	R-STAN-101	94.5	2019
Wang et al. [103]	ARTNet with TSN	94.3	2018
Wang et al. [104]	Temporal Segment Networks	94.2	2016
Ma et al. [105]	TS-LSTM	94.1	2017
Ranasinghe et al. [106]	SVT + ViT-B	93.7	2021
Tran et al. [87]	R[2 + 1]D-RGB	93.6	2017
Carreira et al. [84]	Two-stream I3D	93.4	2017
Tran et al. [87]	R[2 + 1]D-Flow	93.3	2017
Tan et al. [107]	VIMPAC	92.7	2021
Feichtenhofer et al. [44]	S:VGG-16, T:VGG-16	92.5	2016
Shou et al. [94]	DMC-Net	92.3	2019
Zhao et al. [108]	Two-in-one two stream	92.0	2019
Varol et al. [109]	LTC	91.7	2016
Wang et al. [110]	TDD + IDT	91.5	2015
Wang et al. [111]	Very deep two-stream ConvNet	91.4	2015
Shalmani et al. [112]	3D ResNeXt-101 + Confidence Distillation	91.2	2021
Peng et al. [113]	MR Two-Sream R-CNN	91.1	2016
Ranasinghe et al. [106]	SVT	90.8	2021
Bilen et al. [114]	Dynamic Image Networks + IDT	89.1	2016
Yue-Hei Ng et al. [34]	Two-steam + LSTM	88.6	2015
Simonyan et al. [115]	Two-Stream	88.0	2014
Zhang et al. [116]	MV-CNN	86.4	2016
Nguyen et al. [117]	Dynamics 2 for DenseNet-201 Transformer	86.1	2023
Tran et al. [118]	Res3D	85.8	2017
Ng Hei-Yue et at. [119]	ActioFlowNet	83.9	2016
Tran et at. [120]	C3D	82.3	2014
Parmar et at. [121]	HalluciNet	79.8	2019
Pan et al. [122]	R[2 + 1]D	78.7	2021
Pan et al. [122]	3D-ResNet18	74.1	2021
Karpathy et al. [26]	Slow Fusion + Fine-Tune Top 3 layers	65.4	2014
Mazari et al. [123]	MLGCN	63.2	2019
Zhu et al. [124]	CD-UAR	42.5	2018
Choutas et al. [125]	I3D + PoTion	29.3	2018

The progression of research utilizing the UCF-101 dataset, as depicted in Figure 10, offers a compelling narrative of advancement in video classification. This research, a collaborative effort of many in the field, has led to significant breakthroughs. Choutas et al. [125] propose using an interactive three-dimensional (I3D) method combined with Pose moTion (PoTion), achieving an accuracy of 29.3%. Combining I3D with PoTion involves integrating the PoTion representation, which encodes the movement of pose critical points over a video clip, with the I3D architecture for action recognition. These heatmaps are colored based on the frame's time and encode the probability of each pixel containing a specific joint. The colorized heatmaps are summed over all frames to create the PoTion representation. This representation is then used to train a shallow CNN with six convolutional layers and one fully connected layer for action classification. The PoTion representation is combined with the I3D approach, which uses spatio-temporal convolutions and pooling operators from an image classification network. When combined with I3D on RGB and optical flow streams, low accuracy is observed in some classes, such as tying a bow tie and making sushi. This is due to factors like the poor visibility of human actors, especially in first-person videos, the partial visibility of joints, and videos focusing more on objects than human actors. In these cases, the PoTion representation may not capture relevant information effectively, leading to lower accuracy.

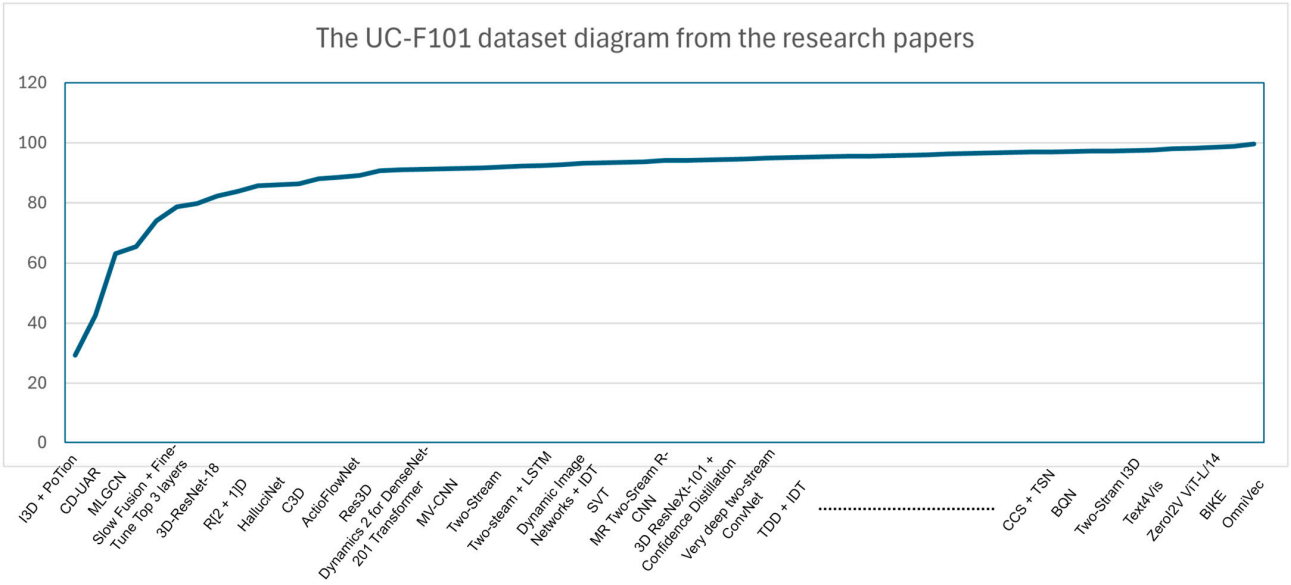

Figure 10. Diagram of the research era of video classification using the UCF-101 dataset.

This marked a starting point for subsequent studies to build upon. Varadarajan et al. [38] introduced the omnidirectional vector (OmniVec) model, achieving a remarkable 99.6% accuracy on the UCF-101 dataset. The OmniVec model is a unified learning framework designed to learn embeddings from multiple modalities and tasks using a shared backbone network. Its three main components are modality-specific encoders, a shared backbone network, and task-specific heads. OmniVec trains sequentially on different tasks and modalities, grouping tasks by the extent of information they exploit across modalities. For example, semantic segmentation embeds more local information than classification. Training data involves mixing samples from each modality for a given task, with modality encoders replaced while task heads and the backbone remain the same. OmniVec handles RGB images and videos, using a standard backbone for processing and facilitating knowledge sharing. It infuses cross-domain information, aligning embeddings from different modalities in a shared space. An iterative training mechanism mixes modalities and groups tasks for better learning. OmniVec achieves high generalization on unseen datasets and tasks, performing state-of-the-art work on various images and videos. It excels in cross-modal knowledge transfer, outperforming other methods in tasks like point cloud

classification, semantic segmentation, and text summarization. OmniVec achieves 99.6% accuracy on the UCF-101 dataset by learning robust representations from multiple modalities and tasks, leveraging a shared backbone network for meaningful representation extraction and generalization. This dramatic improvement underscores the dataset's pivotal role as a cornerstone resource for both researchers and industries engaged in exploring video classification tasks.

5. Discussion

The advancements in deep learning for video classification have been substantial, providing significant improvements in accuracy and efficiency. This section delves into the comparative performance of various deep learning approaches that emphasize the distinction between single and hybrid models.

5.1. Single and Hybrid Models

Single models, primarily leveraging CNNs and RNNs, have been foundational in video classification. CNNs are adept at extracting spatial features from video frames, while RNNs excel at capturing temporal dependencies. However, their isolated use often limits their ability to exploit the intricate spatio-temporal dynamics inherent in videos entirely.

Hybrid models, which combine CNNs with RNNs or transformers models, have emerged as a superior alternative. These models benefit from the strengths of CNNs' spatial feature extraction capabilities and RNNs' or transformers' temporal modeling. For instance, integrating CNNs with transformers has shown remarkable success in enhancing spatial and temporal feature modeling [30], as evidenced by improved accuracy metrics across various benchmarks, including the UCF-101 dataset [80].

5.2. Data Augmentation and Pre-Training

Data augmentation and pre-training on large-scale datasets have also played a pivotal role in advancing video classification. Techniques such as random rotation, shift, and other augmentation methods have improved model robustness and performance [54–59]. Moreover, pre-trained models, which leverage transfer learning, have significantly improved classification tasks by utilizing learned features from extensive datasets to enhance performance on target tasks [70–74].

5.3. Challenges in Video Classification

Despite these advances, several challenges are present in the domain of video classification:

- **Handling large-scale datasets:** Efficiently processing and training on large-scale video datasets remains a significant challenge. The computational cost and time required are substantial, necessitating the development of more efficient algorithms and hardware optimizations.
- **Generalization capabilities:** It is crucial to ensure that models generalize well across diverse video datasets and real-world scenarios. Current models often struggle with overfitting to specific datasets, limiting their applicability.
- **Temporal consistency and long-term dependencies:** Capturing long-term temporal dependencies and maintaining temporal consistency across frames is another critical area where existing models can improve. While hybrid models have made strides in this direction, there is still room for enhancing effectiveness.
- **Reporting and comparing performance metrics:** The inconsistent reporting of performance metrics, such as the absence of standard deviation values, poses a significant challenge. Standard deviations are essential for assessing the variability and reliability of the reported accuracies. Future research should ensure the inclusion of these statistics to facilitate a more robust comparison of algorithm performance. This will enhance the reliability and relevance of the reported results and aid in developing more robust models.

5.4. Future Directions

Future research in video classification should focus on addressing these challenges through several promising avenues:

- **Efficient model architectures:** Developing more efficient model architectures that can handle large-scale datasets without compromising performance is essential. This includes exploring novel neural network designs and leveraging advancements in hardware acceleration.
- **Advanced data augmentation techniques:** Incorporating more sophisticated data augmentation techniques to simulate real-world variations will improve model robustness and generalization.
- **Integration of multimodal data:** Utilizing multimodal data, such as combining video with audio and text, can provide a more comprehensive understanding of the content, leading to better classification performance.
- **Improved temporal modeling:** Enhancing temporal modeling capabilities will be crucial, particularly for long-term dependencies. This will involve developing new types of recurrent units or attention mechanisms explicitly tailored to video data.

6. Conclusions

This paper has comprehensively reviewed deep learning methodologies applied to video classification, offering detailed summaries of significant studies and highlighting key findings. The exploration includes research leveraging the UCF-101 dataset, and clarifies the methodologies and techniques proposed to enhance accuracy. In examining the landscape of video classification tasks, the review covered techniques applied in the image-processing stage, the integration of CNN models with other models for video classification, and open research questions regarding parallel processing techniques for handling large-scale datasets. Additionally, we explained the differences between single-frame, late-fusion, early-fusion, and slow-fusion approaches. We also discussed the challenges and future directions of video classification, evaluation metrics, comparisons from existing papers, and proposed approaches for conducting experiments with large datasets. Our review highlighted the deep learning approaches that outperform other state-of-the-art video classification methods. The integration of CNNs with other models, such as RNNs and transformers, has significantly improved the capture of spatial and temporal features. However, significant gaps still need to be addressed, particularly in handling large-scale datasets efficiently and improving the generalization capabilities of models.

Author Contributions: Conceptualization, project administration, writing—review and editing, M.H. Conceptualization, editing the manuscript, A.L. Format analysis, writing—original draft preparation, writing—review and editing, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2022R1I1A3069371), was funded by BK21 FOUR (Fostering Outstanding Universities for Research) (No.:5199990914048), and was supported by the Soonchunhyang University Research Fund.

Data Availability Statement: Data are available on request due to restrictions, e.g., privacy or ethical.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Global Media Insight Home Page. Available online: https://www.globalmediainsight.com/blog/youtube-users-statistics/ (accessed on 7 June 2024).
2. Youku Home Page. Available online: https://www.youku.com/ (accessed on 7 June 2024).
3. TikTok Home Page. Available online: https://www.tiktok.com/ (accessed on 7 June 2024).
4. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, A.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv* **2016**, arXiv:1609.08675.

5. Fujimoto, Y.; Bashar, K. Automatic classification of multi-attributes from person images using GPT-4 Vision. In Proceedings of the 6th International Conference on Image, Video and Signal Processing, New York, NY, USA, 14–16 March 2024; pp. 207–212.
6. Anushya, A. Video Tagging Using Deep Learning: A Survey. *Int. J. Comput. Sci. Mob. Comput.* **2020**, *9*, 49–55.
7. Rani, P.; Kaur, J.; Kaswan, S. Automatic video classification: A review. *EAI Endorsed Trans. Creat. Technol.* **2020**, *7*, 163996. [CrossRef]
8. Li, Y.; Wang, C.; Liu, J. A Systematic Review of Literature on User Behavior in Video Game Live Streaming. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3328. [CrossRef]
9. Zuo, Z.; Yang, L.; Liu, Y.; Chao, F.; Song, R.; Qu, Y. Histogram of fuzzy local spatio-temporal descriptors for video action recognition. *IEEE Trans. Ind. Inform.* **2019**, *16*, 4059–4067. [CrossRef]
10. Islam, M.S.; Sultana, S.; Kumar Roy, U.; Al Mahmud, J. A review on video classification with methods, findings, performance, challenges, limitations and future work. *J. Ilm. Tek. Elektro Komput. Dan Inform.* **2020**, *6*, 47–57. [CrossRef]
11. Ullah, H.A.; Letchmunan, S.; Zia, M.S.; Butt, U.M.; Hassan, F.H. Analysis of Deep Neural Networks for Human Activity Recognition in Videos—A Systematic Literature Review. *IEEE Access* **2021**, *9*, 126366–126387. [CrossRef]
12. ur Rehman, A.; Belhaouari, S.B.; Kabir, M.A.; Khan, A. On the Use of Deep Learning for Video Classification. *Appl. Sci.* **2023**, *13*, 2007. [CrossRef]
13. Zhang, J.; Yu, X.; Lei, X.; Wu, C. A novel deep LeNet-5 convolutional neural network model for image recognition. *Comput. Sci. Inf. Syst.* **2022**, *19*, 1463–1480. [CrossRef]
14. Fu'Adah, Y.N.; Wijayanto, I.; Pratiwi, N.K.C.; Taliningsih, F.F.; Rizal, S.; Pramudito, M.A. Automated classification of Alzheimer's disease based on MRI image processing using convolutional neural network (CNN) with AlexNet architecture. *J. Phys. Conf. Ser.* **2021**, *1844*, 012020. [CrossRef]
15. Tammina, S. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *Int. J. Sci. Res. Publ. (IJSRP)* **2019**, *9*, 143–150. [CrossRef]
16. Butt, U.M.; Letchmunan, S.; Hassan, F.H.; Zia, S.; Baqir, A. Detecting video surveillance using VGG19 convolutional neural networks. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 1–9. [CrossRef]
17. Kieffer, B.; Babaie, M.; Kalra, S.; Tizhoosh, H.R. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In Proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), Montreal, QC, Canada, 28 November 2017; pp. 1–6.
18. Singla, A.; Yuan, L.; Ebrahimi, T. Food/non-food image classification and food categorization using pre-trained googlenet model. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Amsterdam, The Netherlands, 16 October 2016; pp. 3–11.
19. Kuttiyappan, D. Improving the Cyber Security over Banking Sector by Detecting the Malicious Attacks Using the Wrapper Stepwise Resnet Classifier. *KSII Trans. Internet Inf. Syst.* **2023**, *17*, 1657–1673.
20. Hidayatuloh, A.; Nursalman, M.; Nugraha, E. Identification of tomato plant diseases by Leaf image using squeezenet model. In Proceedings of the International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, 22 October 2018; pp. 199–204.
21. Li, H. Image semantic segmentation method based on GAN network and ENet model. *J. Eng.* **2021**, *10*, 594–604. [CrossRef]
22. Chen, Z.; Yang, J.; Chen, L.; Jiao, H. Garbage classification system based on improved ShuffleNet v2. *Resour. Conserv. Recycl.* **2022**, *178*, 106090. [CrossRef]
23. Zhang, K.; Guo, Y.; Wang, X.; Yuan, J.; Ding, Q. Multiple feature reweight densenet for image classification. *IEEE Access* **2019**, *7*, 9872–9880. [CrossRef]
24. Zhao, L.; He, Z.; Cao, W.; Zhao, D. Real-time moving object segmentation and classification from HEVC compressed surveillance video. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *28*, 1346–1357. [CrossRef]
25. Sivasankaravel, V.S. Cost Effective Image Classification Using Distributions of Multiple Features. *KSII Trans. Internet Inf. Syst.* **2022**, *16*, 2154–2168.
26. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F.-F. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24 June 2014; pp. 1725–1732.
27. Huang, D.; Zhang, L. Parallel Dense Merging Network with Dilated Convolutions for Semantic Segmentation of Sports Movement Scene. *KSII Trans. Internet Inf. Syst.* **2022**, *16*, 1–14.
28. Selva, J.; Johansen, A.S.; Escalera, S.; Nasrollahi, K.; Moeslund, T.B.; Clapés, A. Video Transformers: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12922–12943. [CrossRef]
29. Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; Luo, P. End-to-end dense video captioning with parallel decoding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6847–6857.
30. Gong, H.; Li, Q.; Li, C.; Dai, H.; He, Z.; Wang, W.; Li, H.; Han, F.; Tuniyazi, A.; Mu, T. Multi-scale Information Fusion for Hyperspectral Image Classification Based on Hybrid 2D-3D CNN. *Remote Sens.* **2021**, *13*, 2268. [CrossRef]
31. Li, J. Parallel two-class 3D-CNN classifiers for video classification. In Proceedings of the 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Xiamen, China, 6–9 November 2017; pp. 7–11.
32. Jing, L.; Parag, T.; Wu, Z.; Tian, Y.; Wang, H. Videossl: Semi-supervised learning for video classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, virtual event, 5–9 January 2021; pp. 1110–1119.

33. Wu, Z.; Jiang, Y.G.; Wang, X.; Ye, H.; Xue, X. Multi-stream multi-class fusion of deep networks for video classification. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 1 October 2016; pp. 791–800.
34. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 12 June 2015; pp. 4694–4702.
35. Wu, Z.; Wang, X.; Jiang, Y.G.; Ye, H.; Xue, X. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In Proceedings of the 23rd ACM international Conference on Multimedia, Brisbane, Australia, 13 October 2015; pp. 461–470.
36. Tavakolian, M.; Hadid, A. Deep discriminative model for video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 382–398.
37. Liu, M. Video Classification Technology Based on Deep Learning. In Proceedings of the 2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS), Xi'an, China, 14 August 2020; pp. 154–157.
38. Varadarajan, B.; Toderici, G.; Vijayanarasimhan, S.; Natsev, A. Efficient large scale video classification. *arXiv* **2015**, arXiv:1505.06250.
39. Mihanpour, A.; Rashti, M.J.; Alavi, S.E. Human action recognition in video using DB-LSTM and ResNet. In Proceedings of the 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, 22–23 April 2020; pp. 133–138.
40. Jiang, Y.G.; Wu, Z.; Tang, J.; Li, Z.; Xue, X.; Chang, S.F. Modeling multi-modal clues in a hybrid deep learning framework for video classification. *IEEE Trans. Multimed.* **2018**, *20*, 3137–3147. [CrossRef]
41. Long, X.; Gan, C.; Melo, G.; Liu, X.; Li, Y.; Li, F.; Wen, S. Multi-modal keyless attention fusion for video classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 1–8.
42. de Oliveira Lima, J.P.; Figueiredo, C.M.S. A temporal fusion approach for video classification with convolutional and LSTM neural networks applied to violence detection. *Intel. Artif.* **2021**, *24*, 40–50. [CrossRef]
43. Abdullah, M.; Ahmad, M.; Han, D. Facial expression recognition in videos: An CNN-LSTM based model for video classification. In Proceedings of the 2020 International Conference on Electronics, Information, and Communication, Barcelona, Spain, 19–22 January 2020; pp. 1–3.
44. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
45. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multi-modal Interaction, Tokyo, Japan, 31 October 2016; pp. 445–450.
46. Li, G.; Fang, Q.; Zha, L.; Gao, X.; Zheng, N. HAM: Hybrid attention module in deep convolutional neural networks for image classification. *Pattern Recognit.* **2022**, *129*, 108785. [CrossRef]
47. Mekruksavanich, S.; Jitpattanakul, A. Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition. *Sci. Rep.* **2023**, *13*, 12067. [CrossRef] [PubMed]
48. Ullah, W.; Hussain, T.; Ullah, F.U.M.; Lee, M.Y.; Baik, S.W. TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106173. [CrossRef]
49. Xu, G.; Li, W.; Liu, J. A social emotion classification approach using multi-model fusion. *Future Gener. Comput. Syst.* **2020**, *102*, 347–356. [CrossRef]
50. Jagannathan, P.; Rajkumar, S.; Frnda, J.; Divakarachari, P.B.; Subramani, P. Moving vehicle detection and classification using gaussian mixture model and ensemble deep learning technique. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 5590894. [CrossRef]
51. Kyrkou, C.; Bouganis, C.S.; Theocharides, T.; Polycarpou, M.M. Embedded hardware-efficient real-time classification with cascade support vector machines. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 99–112. [CrossRef]
52. Pérez, I.; Figueroa, M. A Heterogeneous Hardware Accelerator for Image Classification in Embedded Systems. *Sensors* **2021**, *21*, 2637. [CrossRef] [PubMed]
53. Ruiz-Rosero, J.; Ramirez-Gonzalez, G.; Khanna, R. Field Programmable Gate Array Applications—A Scientometric Review. *Computation* **2019**, *7*, 63. [CrossRef]
54. Mao, M.; Va, H.; Hong, M. Video Classification of Cloth Simulations: Deep Learning and Position-Based Dynamics for Stiffness Prediction. *Sensors* **2024**, *24*, 549. [CrossRef] [PubMed]
55. Takahashi, R.; Matsubara, T.; Uehara, K. Data Augmentation Using Random Image Cropping and Patching for Deep CNNs. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2917–2931. [CrossRef]
56. Kim, E.K.; Lee, H.; Kim, J.Y.; Kim, S. Data Augmentation Method by Applying Color Perturbation of Inverse PSNR and Geometric Transformations for Object Recognition Based on Deep Learning. *Appl. Sci.* **2020**, *10*, 3755. [CrossRef]
57. Taylor, L.; Nitschke, G. Improving Deep Learning with Generic Data Augmentation. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Bengaluru, India, 18–21 November 2018; pp. 1542–1547.
58. Sayed, M.; Brostow, G. Improved Handling of Motion Blur in Online Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1706–1716.
59. Kim, E.; Kim, J.; Lee, H.; Kim, S. Adaptive Data Augmentation to Achieve Noise Robustness and Overcome Data Deficiency for Deep Learning. *Appl. Sci.* **2021**, *11*, 5586. [CrossRef]

60. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv* **2017**, arXiv:1711.08200.
61. Ramesh, M.; Mahesh, K. A Performance Analysis of Pre-trained Neural Network and Design of CNN for Sports Video Classification. In Proceedings of the International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 28–30 June 2020; pp. 213–216.
62. Aryal, S.; Porawagama, A.S.; Hasith, M.G.S.; Thoradeniya, S.C.; Kodagoda, N.; Suriyawansa, K. Using Pre-trained Models As Feature Extractor To Classify Video Styles Used In MOOC Videos. In Proceedings of the IEEE International Conference on Information and Automation for Sustainability (ICIAfS), Colombo, Sri Lanka, 21–22 December 2018; pp. 1–5.
63. Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Jiang, Y.G.; Zhou, L.; Yuan, L. Bevt: Bert pre-training of video transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 14–18 June 2022; pp. 14733–14743.
64. De Souza, C.R.; Gaidon, A.; Vig, E.; Lopez, A.M. System and Method for Video Classification Using a Hybrid Unsupervised and Supervised Multi-Layer Architecture. U.S. Patent 9,946,933, 17 April 2018. pp. 1–20.
65. Jaouedi, N.; Boujnah, N.; Bouhlel, M.S. A new hybrid deep learning model for human action recognition. *J. King Saud Univ.-Comput. Inf. Sci.* **2020**, *32*, 447–453. [CrossRef]
66. Kumaran, S.K.; Dogra, D.P.; Roy, P.P.; Mitra, A. Video trajectory classification and anomaly detection using hybrid CNN-VAE. *arXiv* **2018**, arXiv:1812.07203.
67. Ijjina, E.P.; Mohan, C.K. Hybrid deep neural network model for human action recognition. *Appl. Soft Comput.* **2016**, *46*, 936–952. [CrossRef]
68. De Souza, C.R.; Gaidon, A.; Vig, E.; López, A.M. Sympathy for the details: Dense trajectories and hybrid classification architectures for action recognition. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 697–716.
69. Lei, J.; Li, G.; Zhang, J.; Guo, Q.; Tu, D. Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model. *IET Comput. Vis.* **2016**, *10*, 537–544. [CrossRef]
70. Dash, S.C.B.; Mishra, S.R.; Srujan Raju, K.; Narasimha Prasad, L.V. Human action recognition using a hybrid deep learning heuristic. *Soft Comput.* **2021**, *25*, 13079–13092. [CrossRef]
71. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 305–321.
72. Moskalenko, V.V.; Zaretsky, M.O.; Moskalenko, A.S.; Panych, A.O.; Lysyuk, V.V. A model and training method for context classification in cctv sewer inspection video frames. *Radio Electron. Comput. Sci. Control.* **2021**, *3*, 97–108. [CrossRef]
73. Naik, K.J.; Soni, A. Video Classification Using 3D Convolutional Neural Network. In *Advancements in Security and Privacy Initiatives for Multimedia Images*; IGI Global: Hershey, PA, USA, 2021; pp. 1–18.
74. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
75. Solmaz, B.; Assari, S.M.; Shah, M. Classifying web videos using a global video descriptor. *Mach. Vis. Appl.* **2013**, *24*, 1473–1485. [CrossRef]
76. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
77. Xu, H.; Das, A.; Saenko, K. Two-stream region convolutional 3D network for temporal activity detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2319–2332. [CrossRef] [PubMed]
78. AVA Home Page. Available online: https://research.google.com/ava/ (accessed on 7 June 2024).
79. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5842–5850.
80. Srivastava, S.; Sharma, G. Omnivec: Learning robust representations with cross modal sharing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Walkoloa, HI, USA, 4–8 January 2024; pp. 1236–1248.
81. Wu, W.; Wang, X.; Luo, H.; Wang, J.; Yang, Y.; Ouyang, W. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6620–6630.
82. Li, X.; Wang, L. ZeroI2V: Zero-Cost Adaptation of Pre-trained Transformers from Image to Video. *arXiv* **2023**, arXiv:2310.01324.
83. Wu, W.; Sun, Z.; Ouyang, W. Revisiting classifier: Transferring vision-language models for video recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 June 2023; pp. 2847–2855.
84. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulum, HI, USA, 21–26 June 2017; pp. 6299–6308.
85. Huang, G.; Bors, A.G. Busy-quiet video disentangling for video classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1341–1350.
86. Zhang, J.; Shen, F.; Xu, X.; Shen, H.T. Cooperative cross-stream network for discriminative action representation. *arXiv* **2019**, arXiv:1908.10136.

87. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6450–6459.
88. Hong, J.; Cho, B.; Hong, Y.W.; Byun, H. Contextual action cues from camera sensor for multi-stream action recognition. *Sensors* **2019**, *19*, 1382. [CrossRef]
89. Zhao, Z.; Huang, B.; Xing, S.; Wu, G.; Qiao, Y.; Wang, L. Asymmetric Masked Distillation for Pre-Training Small Foundation Models. *arXiv* **2023**, arXiv:2311.03149.
90. Sharir, G.; Noy, A.; Zelnik-Manor, L. An image is worth 16x16 words, what is a video worth? *arXiv* **2021**, arXiv:2103.13915.
91. Zhu, L.; Tran, D.; Sevilla-Lara, L.; Yang, Y.; Feiszli, M.; Wang, H. FASTER Recurrent Networks for Efficient Video Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13098–13105.
92. Qiu, Z.; Yao, T.; Ngo, C.W.; Tian, X.; Mei, T. Learning spatio-temporal representation with local and global diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12056–12065.
93. Zhang, Y.; Li, X.; Liu, C.; Shuai, B.; Zhu, Y.; Brattoli, B.; Chen, H.; Marsic, I.; Tighe, J. VidTr: Video Transformer without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canda, 11–17 October 2021; pp. 13577–13587.
94. Shou, Z.; Lin, X.; Kalantidis, Y.; Sevilla-Lara, L.; Rohrbach, M.; Chang, S.F.; Yan, Z. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1268–1277.
95. Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J. A^2-nets: Double attention networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–10.
96. Sun, S.; Kuang, Z.; Sheng, L.; Ouyang, W.; Zhang, W. Optical flow guided feature: A fast and robust motion representation for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1390–1399.
97. Crasto, N.; Weinzaepfel, P.; Alahari, K.; Schmid, C. Motion-augmented rgb stream for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7882–7891.
98. Liu, M.; Chen, X.; Zhang, Y.; Li, Y.; Rehg, J.M. Attention distillation for learning video representations. *arXiv* **2019**, arXiv:1904.03249.
99. Fan, L.; Huang, W.; Gan, C.; Ermon, S.; Gong, B.; Huang, J. End-to-end learning of motion representation for video understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6016–6025.
100. Huang, G.; Bors, A.G. Learning spatio-temporal representations with temporal squeeze pooling. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2103–2107.
101. Christoph, R.; Pinz, F.A. Spatiotemporal residual networks for video action recognition. *Adv. Neural Inf. Process. Syst.* **2016**, *2*, 3468–3476.
102. Liu, Q.; Che, X.; Bie, M. R-STAN: Residual spatial-temporal attention network for action recognition. *IEEE Access* **2019**, *7*, 82246–82255. [CrossRef]
103. Wang, L.; Li, W.; Li, W.; Van Gool, L. Appearance-And-Relation Networks for Video Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18 June 2018; pp. 1430–1439.
104. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.
105. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Process. Image Commun.* **2019**, *71*, 76–87. [CrossRef]
106. Ranasinghe, K.; Naseer, M.; Khan, S.; Khan, F.S.; Ryoo, M.S. Self-supervised video transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–20 June 2022; pp. 2874–2884.
107. Tan, H.; Lei, J.; Wolf, T.; Bansal, M. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv* **2021**, arXiv:2106.11250.
108. Zhao, J.; Snoek, C.G. Dance with flow: Two-in-one stream action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9935–9944.
109. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef]
110. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314.
111. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y. Towards good practices for very deep two-stream convnets. *arXiv* **2015**, arXiv:1507.02159.
112. Shalmani, S.M.; Chiang, F.; Zheng, R. Efficient action recognition using confidence distillation. In Proceedings of the 26th International Conference on Pattern Recognition, Montral, QC, Canada, 21–25 August 2022; pp. 3362–3369.

113. Peng, X.; Schmid, C. Multi-region two-stream R-CNN for action detection. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 744–759.
114. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; Gould, S. Dynamic image networks for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3034–3042.
115. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
116. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with enhanced motion vector CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2718–2726.
117. Nguyen, H.P.; Ribeiro, B. Video action recognition collaborative learning with dynamics via PSO-ConvNet Transformer. *Sci. Rep.* **2023**, *13*, 14624. [CrossRef] [PubMed]
118. Tran, D.; Ray, J.; Shou, Z.; Chang, S.F.; Paluri, M. Convnet architecture search for spatiotemporal feature learning. *arXiv* **2017**, arXiv:1708.05038.
119. Ng, J.Y.H.; Choi, J.; Neumann, J.; Davis, L.S. Actionflownet: Learning motion representation for action recognition. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1616–1624.
120. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chils, 7–13 December 2015; pp. 4489–4497.
121. Parmar, P.; Morris, B. HalluciNet-ing spatiotemporal representations using a 2D-CNN. *Signals* **2021**, *2*, 604–618. [CrossRef]
122. Pan, T.; Song, Y.; Yang, T.; Jiang, W.; Liu, W. Videomoco: Contrastive video representation learning with temporally adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11205–11214.
123. Mazari, A.; Sahbi, H. MLGCN: Multi-Laplacian graph convolutional networks for human action recognition. In Proceedings of the British Machine Vision Conference, Cardiff, UK, 9–12 September 2019; pp. 1–27.
124. Zhu, Y.; Long, Y.; Guan, Y.; Newsam, S.; Shao, L. Towards universal representation for unseen action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9436–9445.
125. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. Potion: Pose motion representation for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7024–7033.