International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I2P111

Volume 73 Issue 2, 90-109, February 2025
© 2025 Seventh Sense Research Group®

Review Article

Video Summarization with Neural Networks: A Systematic Comparison of State-of-the-Art Techniques

M. Hamza Eissa[1], Hesham Farouk[2], Kamal Eldahshan[1], Amr Abozeid[1,3]

[1]Department of Mathematics, Faculty of Science, Al-Azhar University, Cario, Egypt.
[2]Department of Computers and Systems Electronics Research Institute. Cario, Egypt.
[3]Department of Computer Science, College of Computer and Information Sciences, Jouf University, Saudi Arabia.

[1]Corresponding Author : mohammed_essa2001@yahoo.com

Received: 31 December 2024 Revised: 27 January 2025 Accepted: 18 February 2025 Published: 28 February 2025

Abstract *- Video summarization represents an essential research field dedicated to developing fast methods that extract valuable content from extensive video collections. An evaluation of video summarization strategies with Neural Networks analyzes methodologies together with architectures and evaluation methods, datasets, and performance assessments. This paper conducts an in-depth analysis of different methodological approaches, including supervised and unsupervised learning, reinforcement learning, hybrid models, and object-centric methods, while assessing their performance traits and existing constraints. The popularity of deep learning techniques such as attention mechanism transformers along with hierarchical reinforcement learning shows continuous growth due to their effectiveness in improving summarization accuracy and efficiency. Summaries achieve higher quality through visual, audio, and textual features, which help produce better outputs for video tutorials combined with tournament highlights and security footage monitoring. The research investigates evaluation frameworks specifically by highlighting weaknesses in existing benchmark metrics and calling for Performance over Random (PoR) as a strong alternative framework. The current model faces ongoing issues with real-time operation, computational speed, and summation generation control based on user needs. The paper explores existing state-of-the-art approaches and proposes research directions focusing on scalability while developing user-customized frameworks and better-assessing metrics.*

Keywords *- Video Summarization, Feature Extraction, Neural Networks, Deep Learning, Reinforcement Learning.*

1. Introduction

Video content expands rapidly across digital platforms, making efficient video management and navigation through vast collections urgently necessary. Search for essential information becomes constrained by human capabilities because online content exceeds billions of hours beyond human comprehension pertaining to time restrictions and information overload. Video summarization tools are critical in resolving video management challenges by creating condensed versions of lengthy videos that maintain all their necessary information. Video condensation remains essential because it provides an excellent user experience by speeding up access times, especially during news broadcasting, entertainment streaming and security surveillance practices [1].

Deep learning emerged as the principal advancement in machine learning techniques that developed video summarization methods during the same period. The combination of basic methods that use feature extraction algorithms and clustering techniques met challenges because of complicated video content when dealing with current video materials. Deep neural networks enabled new methods of producing video summaries that achieved better performance. Research teams achieve video summary creation through short temporal context-preserving sequences by applying convolutional neural networks (CNNs) [1] as spatial filters in combination with recurrent neural networks (RNNs) [2]for temporal navigation.

Video summarization techniques have been improving simultaneously with machine learning technology to reach the breakthrough point of deep learning eventually. The video content processing problems during modern video content analysis were mainly caused by basic methods with feature extraction and clustering algorithms [3]. Implementing deep neural networks created new techniques to produce superior video summaries. The combination of CNNs and RNNs serves as spatial filters and temporal navigators, enabling research teams to generate contextual video summaries with short durations [4]. It is essential to compare these approaches because they reveal the execution challenges between methods while directing upcoming development strategies within this domain [5].

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Video summarization processes require several essential stages to transform full-length videos into their dominant informative sections. The method creates an organized process to manage video data effectively while preserving essential content elements [4]. Uniform sampling represents the initial operation of the workflow system. When inputting video files, the workflow starts by displaying continuous frames in rapid succession to produce animation. Testing begins with separating each frame to create a sequence that simplifies the stored video information analysis. The selected video frames become representative through applications of uniform sampling techniques. By applying this sampling technique, frames are selected evenly throughout the video period [6]. The technique enables a wider range of content to be present in the sample, minimizing the chance of missing important scenes or moments. Such preparatory work provides analytical materials by generating usable frame collections that serve as the basis for following processing techniques. After frame sampling takes place, the following process becomes Redundancy Elimination. Uniform sampling of frames leads to redundant information because the collected sequences contain various repeated and identical pictures that provide minimal value to the summary. The analyzed frames undergo review to determine which ones will get eliminated as redundant. After analysis, the selected frames become a focused subset that represents various segments of the video. The redundancy elimination stage improves subsequent processing efficiency while guaranteeing that the summary consists of distinctive informative material [7].

During feature extraction, the third important step extracts quantitative values that describe the distinct characteristics present in each frame. The following resizing operation applies to non-redundant frames, which are downscaled to a uniform 299 x 299 pixels. Proper data input uniformity demands this scaling operation before the system processes data. The Inception-ResNet-v2 model in CNN processes sets of image frames after their frames experience resizing [8]. The successful processing of visual data through images is attributed to CNNs because they efficiently extract complex features from pictures including edges combined with textures together with patterns [9]. The extracted features from these frames offer an efficient and holistic depiction of visual elements that serve essential classification needs [10].

The workflow finishes by classifying features through the utilization of a Random Forest Classifier. A Random Forest Classifier uses ensemble learning techniques for processing extracted features that originate from the CNN. By combining multiple decision trees, the Random Forest improves accuracy levels and error resistance [11]. Evaluating video frame features leads to selecting the most important frames through keyframe identification. The video summarization process generates keyframes as its end result, which contain the essential information from each video segment. The keyframe algorithm transforms video material into representative segments through this process to create a summary that maintains the original content meaning but decreases its total length [12].

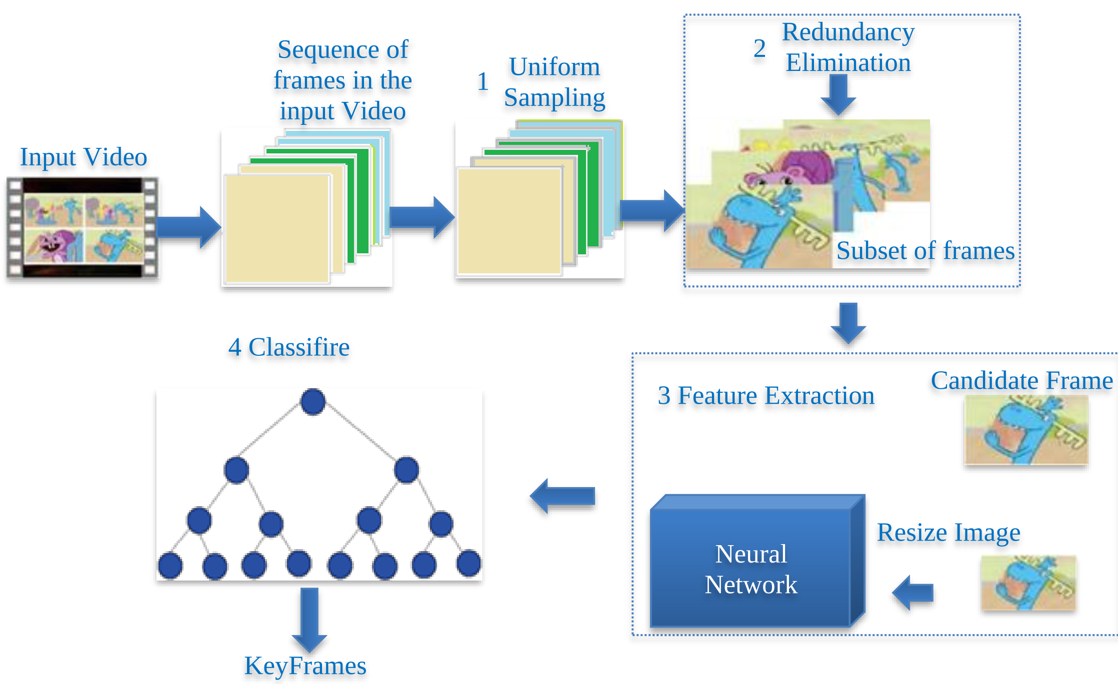

Fig. 1 Video summarization process using Neural Networks

The organized workflow method enables a methodical process for video summary generation. The procedure reaches an efficient and informative state through the combination of techniques including uniform sampling together with redundancy elimination and feature extraction and classification. The strategy successfully handles massive video information across multiple domains, providing a strong approach for video content management and summarization.

The video processing technique begins by dividing the video into frames while removing duplicated sections and extracting features with a CNN before performing feature classification to identify the optimal keyframes for the video summary.

This paper aims to close this gap by providing a full assessment and comparative analysis of existing video summarizing approaches that use deep neural networks. By examining a wide range of approaches, including those that employ CNNs [1], RNNs [2], and their variants, this study aims to provide a clear understanding of the strengths and limitations of each method. The goal is to offer insights that can inform the development of more effective and efficient video summarization tools, ultimately contributing to the broader effort of managing and making sense of the ever-growing volume of video content in the digital age.

The remainder of the paper is organized as follows: Section 2 provides a comprehensive overview of existing video summarization techniques, highlighting their key methodologies, advantages, and limitations. Section 3 outlines the criteria used for comparing different video summarization approaches. Also, the various video summarization techniques are compared based on these criteria. Section 4 discusses comparative findings, analyzing the strengths and weaknesses of each technique and their impact on video summarization research. Finally, Section 5 concludes the paper by summarizing key insights, identifying existing challenges, and providing recommendations for future research directions.

2. Related Work

Video content analysis through summarization has become central to multimedia processing because it enables the effective extraction of significant data from large video datasets. Video summarization research now utilizes deep learning-based models, attention mechanisms, reinforcement learning, and hybrid approaches. This section details an extensive evaluation of primary accomplishments in video summarization while discussing their successful elements together with their associated shortcomings.

2.1. Object-Centric Video Summarization

The main video summarization approach involves identifying objects of interest (OoI) in video images. The authors in [13] presented an OoI-based framework composed of three sequential phases that involve object selection, go into localization, and finally lead to summarization. Users start by picking objects from a pre-established dictionary to eliminate unneeded frames. The YOLOv3 deep learning model identifies and localizes detected objects in video frames during its operation. The selection of keyframes with detected objects leads to the production of a brief summary. Through this approach, video summaries become more effective because personalized user preferences direct the process of maintaining crucial details while reducing unnecessary content. The automatic selection of objects for focus enhances its practical usage for surveillance systems, event detection platforms, and personalized video summary production.

2.2. Attention-Based and Context-Aware Summarization

Video summarization algorithms use attention mechanisms because they allow the effective detection of significant temporal relationships between frames while dynamically assigning frame importance values. This paper in [14] focused on overcoming supervised summarization problems by addressing both short-term contextual attention deficits and distribution inconsistency. An encoder self-attention model combined with encoder-decoder attention allowed the authors to improve both short-term connection detection while maintaining matching scores between actual and predicted significance results. The video summarization system ADSum utilized by researchers led to optimal results on benchmark datasets, including SumMe and TVSum, through their Attentive and Distribution-consistent Video Summarization (ADSum) solution.

In their significant work, the researchers presented SUM-GDA, which stands for Summarized Video Generation, using a Diverse Attention mechanism [15]. Through pairwise temporal relation modeling, the system releases its ability to discover crucial video frames relative to the whole video content, thereby producing more diverse summary outputs. SUM-GDA reduces duplicate content while preserving content variety, which makes it practical for news highlight generation, sports recap production, and academic video-shortening procedures.

2.3. Deep Learning and Reinforcement Learning Approaches

Deep learning has completely transformed video summarization through its ability to find complex patterns inside video content. Video summarization was studied as a sequential decision-making process according to [16] through an introduction of Deep Hierarchical LSTM Networks with Attention (DHAVS). DHAVS applies 3D CNNs to extract spatio-temporal features before using a hierarchical attention-based LSTM module to address long-range dependency constraints. The researchers obtained substantial F-score improvements through hierarchical architectures when testing SumMe and TVSum datasets. The structured process enables effective treatment of extended content blocks, thus making it suitable for documentaries and lectures, along with movie summarization. Through reinforcement learning techniques, developers have improved various methods of video summarization. The authors of [17] developed a hierarchical reinforcement learning method that breaks video summarization into smaller subtasks to enhance convergence and reduce sparse reward difficulties. Managers within the framework establish sub-goals through their network while workers operate through their network to evaluate frame importance using policy gradients. The research of [18] developed a frame selection optimization method through reinforcement learning by combining ResNet and Gated Recurrent Units (GRUs). Adaptation through learning and better summarization results occur when reward-based training methods are implemented. The application of reinforcement learning-based techniques proves advantageous in situations that need to adapt automatically because of their purpose in real-time video summarization and intelligent surveillance.

2.4. Hybrid and Multi-Feature Summarization Approaches

Hybrid extractive techniques use several feature extractions to develop higher-quality summary outputs. The researchers in [19] developed a system for summary production in cricket videos by analyzing audio alongside visual information. Through their framework, this research solution incorporated SGRNN-AM for audio-based event detection and HRF-DBN for scene classification. The method utilized speech-to-text conversion, audio energy levels, and scoreboard visual features to enhance event-based summarization accuracy, allowing it to work in sports highlights and action-filled video summaries. The authors in [20] created a video summarization system using extracted feature variables from coded video bitstreams. The research method started by using stepwise regression to minimize dimensions, and then frame-based temporal sampling was performed through cosine similarity calculations followed by PCA projections. Deep learning LSTMs and CNNs installed within their model enabled better keyframe extraction as they obtained high F-scores across multiple datasets. Combining different methods, known as hybrid approaches, helps integrate three data types, including audio and visual data and text, which applies to medical video summarization and courtroom proceedings.

2.5. Motion and Spatiotemporal Feature-Based Summarization

Standard keyframe-choosing techniques fail to incorporate motion data, which is essential in videos requiring movement. Researchers in [21] created a motion-aware summary system by fusing Capsule Networks with optical flow analysis for extracting space-time features. By employing this technique, high-motion sequences get adequately preserved, so it works best on content like action videos and security and sports material. The authors of [22] developed a 3D spatiotemporal U-Net together with reinforcement learning as a framework for medical video summarization enhancement. Primary diagnosis system developers utilize their approach to recognize spatial-temporal linkages in ultrasound and fetal screening videos because it showcases how spatiotemporal analysis functions in medical diagnostics alongside security analytics and autonomous vehicle guidance.

2.6. Unsupervised and Weakly Supervised Summarization Techniques

Multiple research projects have explored unsupervised learning methods because they lower the need for manually labeled datasets. This research designed an unsupervised framework by uniting traditional vision-based algorithms with deep learning methods for feature extraction [23]. Keyframes were clustered through K-means and Gaussian mixture models by integrating uniform sampling and histograms and SIFT and CNN-based feature extraction methods. The approach proved competent in video summarization tasks that involved dynamic content with multiple viewpoints, which made it suitable for social media applications and personal video organization. Another novel approach [24] introduced Global-and-Local Relative Position Embedding (GL-RPE) to improve unsupervised video summarization. By incorporating relative position embeddings, this method enhances temporal consistency and significantly improves the quality of generated summaries compared to previous unsupervised techniques. Such methods are essential for large-scale applications where manual labeling is infeasible, such as autonomous surveillance and industrial monitoring. The evaluation process takes central importance in the field of video summarization research. Research in [25] demonstrated problems with standard evaluation techniques using F-score and arbitrary dataset splitting. The authors developed Performance over Random (PoR), which serves as an assessment tool to estimate dataset difficulty while improving the reliability of performance evaluation. The researchers demonstrated through their work that standard evaluation methods need to be established because they serve to provide fair evaluations across various summarization approaches. A strong evaluation system guarantees practical usage and wide application compatibility of summarization methods for all video types and usage scenarios.

3. Comparison of Video Summarization Techniques

3.1. Comparison Criteria for Video Summarization Techniques

Video summarization techniques have evolved significantly, reflecting the diverse requirements of various applications and rapid technological advancements. These techniques can be broadly classified based on their methodologies, data processing types, and specific objectives. Below, we outline the key criteria used to compare the video summarization techniques analyzed in this study:

3.1.1. Method

Defines the primary approach or algorithm used, such as recurrent neural networks, convolutional neural networks, or attention-based encoder-decoder models. This criterion helps in understanding the fundamental mechanisms employed to generate video summaries.

3.1.2. Input Video Type

Specifies the category of input videos evaluated using each technique, such as TV series, sports, or instructional videos. This criterion provides insight into the adaptability and generalizability of each technique across different video types.

3.1.3. Dataset

Identifies the dataset used for evaluation, such as SumMe, TVSum, or custom datasets. This criterion is crucial in assessing the representativeness and quality of the data used for benchmarking the techniques.

3.1.4. Evaluation Metric

Details the quantitative measures used to assess the performance of each technique, such as F-measure, F1-score, or precision-recall metrics. Understanding evaluation methods provides insight into the comparative effectiveness of different techniques.

3.1.5. Feature Extraction

Describes the methodology used to extract relevant features from video data, such as deep visual features, optical flow, or ResNet-50 embeddings. This criterion is essential in understanding how video data is processed and analyzed.

3.1.6. Input Representation

Specifies the format of video data input into each technique, such as RGB frames, optical flow sequences, or hybrid representations. This criterion explains the preprocessing and encoding methods utilized in each technique.

3.1.7. Architecture

Identifies the specific neural network architecture implemented, such as LSTMs, transformer models, or encoder-decoder frameworks. Understanding the model architecture provides insights into each technique's computational complexity and design.

3.1.8. Summary Type

Defines the type of summary generated, including static keyframes, dynamic skimming-based summaries, or image-based summaries. This criterion clarifies how summarized content is structured and presented.

3.1.9. Main Contributions

Highlights the strengths and advantages of each technique, such as robustness to varying video lengths, superior temporal modeling, or improved semantic understanding. This criterion helps identify suitable techniques for different applications.

3.1.10. Main Limitations

Outlines the drawbacks or constraints of each technique, such as computational inefficiency, high reliance on labeled data, or challenges in handling diverse video content.

Recognizing these limitations is essential for understanding trade-offs in selecting an appropriate summarization method.

3.2. Comparison of Video Summarization Techniques

The evaluation in Table 1 presents detailed assessments of neural network-based video summarization techniques through multiple evaluative metrics such as effectiveness, efficiency, flexibility, accuracy and adaptability.

This analysis investigates major elements comprising method type, input video attributes, applied datasets, architectural structures and evaluation standards, temporal control mechanisms, extraction methods, and representation techniques.

This criterion evaluates the advantages and drawbacks of proposed methods to understand their complete operational capabilities and practical market potential.

Evaluating both the strengths and weaknesses of approaches is vital to assess their application, particularly video summarization problems and practical implementation scenarios.

4. Discussion

Video summarization is a vital research topic because the massive growth of video content occurs within surveillance and entertainment, along with educational and athletic domains.

Research into efficient video summarization techniques has led scientists to create multiple innovative approaches that apply deep learning together with reinforcement learning and hybrid approaches. The systematic research evaluates multiple video summarization techniques to analyze key characteristics, benefits, and drawbacks that determine their critical role in field advancement.

Five main approaches define the advancement of video summarization techniques, which include Supervised learning, Unsupervised learning, Reinforcement Learning, Hybrid methods and Object-Centric summarization.

The left-hand side of Figure 1 illustrates the number of studies that utilize each methodology.

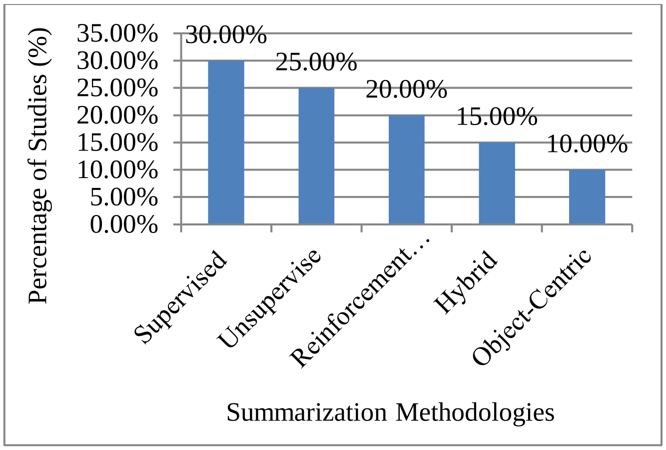

Fig. 2 Evaluation metrics used in video summarization research

The supervised learning method achieves high accuracy but requires extensive labeled datasets for its operation. Unsupervised clustering and autoencoder algorithms reduce dependence on annotations but need different dataset-specific adjustments. The adaptive learning method of reinforcement learning enables models to discover optimal summarization approaches dynamically yet requires large computing power. Hybrid methods combine various feature extraction methods, which results in better accuracy, though they introduce additional complexity to the system. The user-preference-based summarization of object-centric approaches becomes possible through robust object detection models, yet it requires strong detection capabilities.

4.1. Advancements in Supervised and Unsupervised Learning

Supervised learning techniques enhance video summary quality by processing enormous labeled datasets effectively. The combination of Deep Attentive Video Summarization with Distribution Consistency Learning and SUM-GDA brings together deep neural networks alongside attention elements to extract better frames that produce highly pertinent summaries. Labeled data at a large scale creates barriers to video summary applications across different video categories. The Unsupervised Video Summarization Framework Using Keyframe Extraction and Video Skimming with Global-and-Local Relative Position Embedding functions as an unsupervised method, which addresses scalability challenges through unannotated workflow. These methods show great promise in delivering personalized generalizable video summaries even though further tuning processes will be needed to achieve high-quality results.

4.2. Hybrid Approaches and Multi-Feature Integration

Multiple summarization methods obtain greater robustness through combining feature extraction methods. The implementation of SGRNN-AM together with HRF-DBN and the implementation of Static Video Summarization Using Video Coding Features with Frame-Level Temporal Subsampling connected with Deep Learning frameworks delivers superior summarization results by combining video coding features with audiovisual cues. Multiple feature extraction methods enhance single-modality summarization by effectively enriching extracted data features. The high computational requirements and demanding preprocessing steps present obstacles to the real-time usage of these methods.

4.3. Reinforcement Learning and Hierarchical Models

Reinforcement learning introduces fundamental changes to video summarization because models acquire the capability to change their frame selection strategies dynamically. The research of reinforcement learning-based video summarization results with a combination of ResNet and gated recurrent unit and video summarization through reinforcement learning with a 3D spatio-temporal U-Net prove that reinforcement learning frameworks succeed in discovering optimal summarization procedures. Applying hierarchical reinforcement learning in a video summarization model based on deep reinforcement learning with long-term dependency increases model capabilities to handle long-term dependencies. The latest developments in summarization techniques produce smarter conversion methods, although they maintain high processing demands and require refined reward rules to achieve better results.

4.4. Object-Centric and User-Preference-Based Summarization

The introduction of object-centric video summarization added a fresh layer to personal video summarization capabilities. Researchers have proven the value of object detection models along with YOLOv3, particularly through Investigations, including an effective video summarization framework based on the object of interest using deep learning and an optimized deep learning method for video summarization based on the user object of interest. Such methods use predefined objects of interest to create video summaries that improve user engagement. Additional improvements need to be made to object detection accuracy, low-resolution frame processing, and frame occlusion resolution while working with these systems.

4.5. Spatiotemporal and Motion-Aware Summarization

Creating summaries with traditional keyframe-based methods usually results in insufficient coverage of dynamic video content. The summarization approaches A Novel Keyframes Selection Framework for Comprehensive Video Summarization and Exploring Global Diverse Attention via Pairwise Temporal Relation for Video Summarization utilize spatiotemporal features to enhance motion-aware summary generation. Unsupervised Video Summarization Based on Deep Reinforcement Learning with Interpolation maintains temporal consistency in dynamic content through reinforcement learning techniques. The techniques enhance motion-based video summary generation while demanding significant computational resources with optimized processing systems.

4.6. Datasets Utilized in Video Summarization

Various datasets have been employed to evaluate video summarization techniques. The most commonly used datasets include SumMe, TVSum, VSUMM, YouTube datasets, and Custom datasets created for specific applications. The percentage of studies utilizing these datasets is depicted in Figure 2.

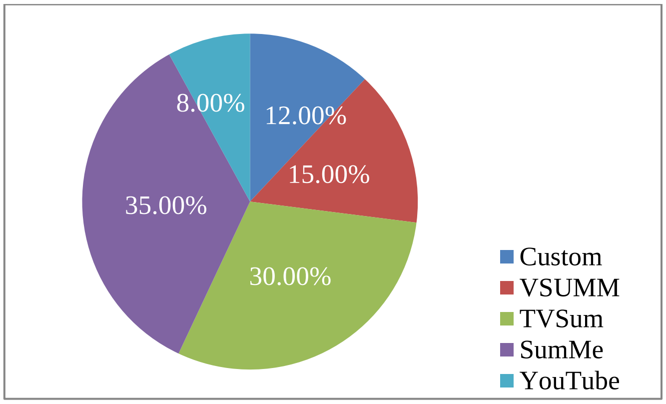

Fig. 3 Percentage of Dataset Usage in Video Summarization Studies

SumMe and TVSum are the most frequently used benchmark datasets, as they provide a diverse range of videos annotated with human-generated summaries. However, the reliance on human annotations introduces subjectivity, making the evaluation process inconsistent. While tailored for specific applications (e.g., sports, medical, and surveillance videos), custom datasets often lack generalizability and cross-dataset validation.

4.7. Evaluation Metrics in Video Summarization

The effectiveness of video summarization models is assessed using various evaluation metrics, including F-score, Precision, Recall, Accuracy, and Kendall's τ. The frequency of these metrics used in recent studies is represented in Figure 3.

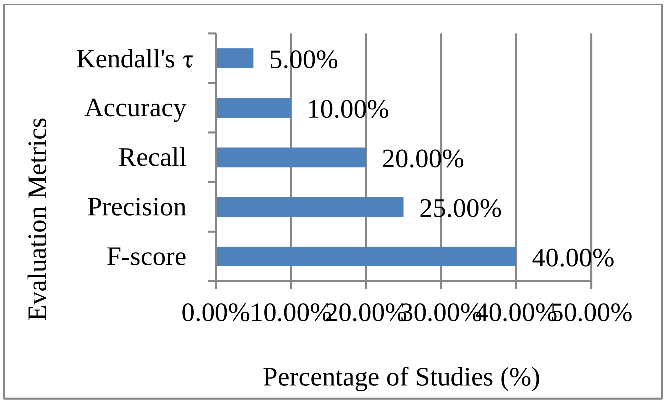

Fig. 4 Evaluation Metrics Used in Video Summarization Research

F-score is the most widely adopted metric, balancing precision and recall to measure summary accuracy. However, some studies have raised concerns over its inability to capture long-term dependencies in summaries. Metrics like Kendall's τ and Spearman's ρ have been proposed to address ranking consistency but are less commonly used due to their complexity in interpretability. The need for more robust and standard evaluation protocols has led to frameworks like Performance over Random (PoR), which accounts for dataset difficulty variations.

4.8. Impact and Future Directions

Research outcomes demonstrate substantial developments in video summarization methods, which include attention-based learning approaches, reinforcement learning systems, and hybrid framework solutions. Supervised video analysis methods deliver exceptional results, yet the need for labeled data creates scalability problems for the approach. Beyond their promise, both unsupervised and weakly supervised models need further development to maximize their summary quality output. The computational challenges are present when implementing reinforcement learning approaches which simultaneously offer better decision-making abilities along with improved adaptability. Object-centric and motion-aware technologies enable improved summary performance alongside better user experiences, although their practical implementation needs improvements for efficiency and generalization capabilities.

Future investigators should direct their efforts towards creating summary systems that maintain both performance and readability alongside fast processing abilities. Future research must develop standardized assessment methods that enable fair evaluation of video summarization approaches throughout different datasets and real-life deployments. The future development of video summarization models depends heavily on deep learning and artificial intelligence evolution, which will produce scalable summary systems for various applications.

5. Conclusion

Video summarization has witnessed significant advancements with the adoption of deep learning, reinforcement learning, and hybrid machine learning models, improving the efficiency and accuracy of automatic summarization techniques. This study has comprehensively analysed various neural network-based video summarization methods, categorizing them based on their architecture, input representation, evaluation metrics, and summarization strategies. The research highlights the evolution from traditional handcrafted feature-based methods to more sophisticated deep learning frameworks, such as attention mechanisms, transformers, and multi-modal learning approaches. While supervised learning methods have demonstrated high accuracy, their reliance on extensive labeled datasets limits scalability. Conversely, unsupervised and weakly supervised approaches mitigate this issue by leveraging clustering and self-supervised learning techniques but often require further fine-tuning for generalizability. Reinforcement learning-based summarization models have shown promising results in adaptive keyframe selection and learning optimal summary generation policies, yet they pose computational challenges due to their complexity. Hybrid models integrating audio, textual, and visual features have further enhanced video summarization quality, particularly in domains such as educational content, sports event analysis, and surveillance footage. Despite these advancements, several challenges persist, including computational efficiency, real-time summarization, scalability, and personalization. Many existing methods require high processing power, making them impractical for real-time and resource-constrained environments. Additionally, a lack of standardized evaluation protocols leads to inconsistent benchmarking across different datasets. Future research should focus on developing lightweight models for real-time applications, improving user-adaptive summarization, and establishing robust evaluation frameworks such as Performance over Random (PoR) to ensure fair comparisons.

References

[1] Lingkun Chen et al., "Convolutional Neural Networks (CNNs)-Based Multi-Category Damage Detection and Recognition of High-Speed Rail (HSR) Reinforced Concrete (RC) Bridges Using Test Images," *Engineering Structures*, vol. 276, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ambre Dupuis, Camélia Dadouchi, and Bruno Agard, "Methodology for Multi-Temporal Prediction of Crop Rotations Using Recurrent Neural Networks," *Smart Agricultural Technology*, vol. 4, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Preeti Meena, Himanshu Kumar, and Sandeep Kumar Yadav, "A Review on Video Summarization Techniques," *Engineering Applications of Artificial Intelligence*, vol. 118, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ambreen Sabha, and Arvind Selwal, "Data-Driven Enabled Approaches for Criteria-Based Video Summarization: A Comprehensive Survey, Taxonomy, and Future Directions," *Multimedia Tools and Applications*, vol. 82, pp. 32635-32709, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Yujie Li et al., "A DCA-Based Sparse Coding for Video Summarization with MCP," *IET Image Processing*, vol. 17, no. 5, pp. 1564-1577, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Anil Singh Parihar, Joyeeta Pal, and Ishita Sharma, "Multiview Video Summarization using Video Partitioning and Clustering," *Journal of Visual Communication and Image Representation*, vol. 74, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ghazaala Yasmin et al., "Key Moment Extraction for Designing an Agglomerative Clustering Algorithm-Based Video Summarization Framework," *Neural Computing and Applications*, vol. 35, no. 7, pp. 4881-4902, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Teja Kattenborn et al., "Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24-49, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Zewen Li et al., "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999-7019, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Nikhil Ketkar, and Jojo Moolayil, *Convolutional Neural Networks*, Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch, pp. 197-242, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Wencheng Zhu et al., "DSNet: A Flexible Detect-To-Summarize Network for Video Summarization," *IEEE Transactions on Image Processing*, vol. 30, pp. 948-962, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Muhammad Rafiq et al., "Scene Classification for Sports Video Summarization Using Transfer Learning," *Sensors*, vol. 20, no. 6, pp. 1-18, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Hafiz Burhan Ul Haq et al., "An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning," *Mathematical Problems in Engineering*, vol. 2022, no. 1, pp. 1-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Zhong Ji et al., "Deep Attentive Video Summarization with Distribution Consistency Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1765-1775, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Evlampios Apostolidis et al., "Performance Over Random: A Robust Evaluation Protocol for Video Summarization Methods," *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1056-1064, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Hansa Shingrakhia, and Hetal Patel, "SGRNN-AM and HRF-DBN: A Hybrid Machine Learning Model for Cricket Video Summarization," *The Visual Computer*, vol. 38, no. 7, pp. 2285-2301, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yunjae Jung et al., "Global-and-Local Relative Position Embedding for Unsupervised Video Summarization," *European Conference on Computer Vision*, Glasgow, UK, pp. 167-183, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Shruti Jadon, and Mahmood Jasim, "Unsupervised Video Summarization Framework using Keyframe Extraction and Video Skimming," *IEEE 5th International Conference on Computing Communication and Automation*, Greater Noida, India, pp. 140-145, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Muhammad Atif Afzal, and Muhammad Sohail Tahir, "Reinforcement Learning based Video Summarization with Combination of ResNet and Gated Recurrent Unit," *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 4, pp. 261-268, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Hafiz Burhan Ul Haq, Watcharapan Suwansantisuk, and Kosin Chamnongthai, “An Optimized Deep Learning Method for Video Summarization Based on the User Object of Interest,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, pp. 244-256, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Ping Li et al., “Exploring Global Diverse Attention via Pairwise Temporal Relation for Video Summarization,” *Pattern Recognition*, vol. 111, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Xu Wang et al., “A Video Summarization Model Based on Deep Reinforcement Learning with Long-Term Dependency,” *Sensors*, vol. 22, no. 19, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Cheng Huang, and Hongmei Wang, “A Novel Key-Frames Selection Framework For Comprehensive Video Summarization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 577-589, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[24] Kazuki Kawamura, and Jun Rekimoto, “FastPerson: Enhancing Video-Based Learning through Video Summarization that Preserves Linguistic and Visual Contexts,” *Proceedings of the Augmented Humans International Conference*, Melbourne, Australia pp. 205-216, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Obada Issa, and Tamer Shanableh, “Static Video Summarization Using Video Coding Features With Frame-Level Temporal Subsampling and Deep Learning,” *Applied Sciences*, vol. 13, no. 10, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Jia-Hong Huang et al., “Causalainer: Causal Explainer for Automatic Video Summarization,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2630-2636, 2023. [Google Scholar] [Publisher Link]

[27] Ui Nyoung Yoon, Myung Duk Hong, and Geun-Sik Jo, “Unsupervised Video Summarization Based on Deep Reinforcement Learning with Interpolation,” *Sensors*, vol. 23, no. 7, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Arati Kushwaha et al., “Human Activity Recognition Based on Video Summarization and Deep Convolutional Neural Network,” *The Computer Journal*, vol. 67, no. 8, pp. 2601-2609, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Vrushali Raut, and Reena Gunjan, “Transfer Learning Based Video Summarization in Wireless Capsule Endoscopy,” *International Journal of Information Technology*, vol. 14, no. 4, pp. 2183-2190, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[30] Tianrui Liu et al., “Video Summarization through Reinforcement Learning with a 3D Spatio-Temporal U-Net,” *IEEE Transactions on Image Processing*, vol. 31, pp. 1573-1586, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Rabbia Mahum et al., “A Robust Framework to Generate Surveillance Video Summaries using Combination of Zernike Moments and r-Transform and Deep Neural Network,” *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 13811-13835, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ye Yuan, and Jiawan Zhang, “Unsupervised Video Summarization via Deep Reinforcement Learning with Shot-Level Semantics,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 445-456, 2022. [CrossRef] [Google Scholar] [Publisher Link]

Appendix

Table 1. Comparing video summarization techniques using neural networks

Paper	Method	Input Video Type	Dataset	Evaluation Metric	Feature Extraction	Input Representation	Architecture	Summary Type	Main contributions	Main limitations
[14].	Customized video summarization framework based on deep learning and object of interest (OoI) detection using YOLOv3.	Various types, including surveillance videos, documentaries, news, etc.	VSUMM TVSum's own dataset	Precision, Recall, F1-score, Accuracy	YOLOv3 for object detection and localization	Frames containing the object of interest	YOLOv3 (You Only Look Once version 3)	Dynamic summaries based on objects of interest	High accuracy in detecting and summarizing videos based on specific objects of interest; Supports multiple static and dynamic objects.	High computational power requirements: Performance can be affected by video quality and object size.
[16].	Sequence-to-sequence learning framework with self-attention and distribution consistency learning.	Various types, including user videos, YouTube videos, etc.	SumMe dataset, TVSum dataset, YouTube dataset	F-measure	GoogleNet features, BiLSTM, LSTM	Frame sequences downsampled to 2 frames per second	BiLSTM encoder, LSTM decoder with self-attention mechanism	Dynamic summaries based on frame-level importance scores	Addressed short-term contextual attention insufficiency and distribution inconsistency issues; Improved performance on benchmark datasets.	Requires large training data; Deficient in modeling very long-term contextual attention.
[18]	Hybrid machine learning framework combining SGRNN-AM for audio-based excitement detection and HRF-DBN for visual scene classification.	Cricket videos	Custom dataset of cricket videos collected from sports broadcasti	Precision, Recall, F1-score, Accuracy, Error rate	Audio energy levels, speech-to-text conversion, scorecard region	Video frames, audio clips, text transcripts from speech-to-text conversion	Stacked Gated Recurrent Neural Network with Attention Module (SGRNN-AM), Hybrid Rotation Forest-Deep Belief Network	Dynamic summaries based on key events (fours, sixes, wickets)	Developed a hybrid approach combining audio and visual features; Improved accuracy in detecting key	High computational requirements; Dependence on accurate speech-to-text conversion and precise feature extraction

			ng channels and YouTube.		analysis, umpire gesture detection		(HRF-DBN)		events; Addressed limitations of using only one feature type.	from scorecards and umpire gestures.
[22]	Deep reinforcement learning framework with ResNet-152 for feature extraction and GRU for sequence modeling.	Various types, including sports, holidays, and events.	SumMe dataset	F-measure	ResNet-152 for deep feature extraction	Video frames converted from videos	ResNet-152 combined with two-layer GRU	Dynamic summaries based on frame-level importance scores	Combined ResNet and GRU for improved video summarization, enhanced performance on benchmark datasets, and efficient feature extraction and sequence modeling.	High computational requirements; Dependence on the quality of feature extraction and sequence modeling.
[20]	Unsupervised learning framework using traditional vision-based algorithms and deep learning for keyframe extraction and video skimming.	Various types, including user-generated videos and dynamic viewpoint videos.	SumMe dataset	F-score	Uniform sampling, image histograms, SIFT, and CNNs trained on ImageNet	Video frames extracted from videos	Traditional vision-based methods combined with deep learning models (CNNs)	Dynamic summaries based on keyframe extraction and video skimming	Proposed an effective unsupervised video summarization framework; Demonstrated superior performance of deep learning-based feature extraction; Addressed the	Depending on the quality of the clustering methods, further optimization is needed for different types of videos.

									need for personalized and generalizable video summaries.	
[26]	Framework using Capsules Net for spatiotemporal information extraction, TED for shot segmentation, and self-attention for keyframe selection.	Various types, including those with significant motion.	VSUMM, TvSum, SumMe, RAI datasets	F-score, subjective and objective performance metrics	Capsules Net for spatiotemporal features, optical flow calculation for motion	Video frames with extracted spatiotemporal features	Capsules Net for feature extraction, self-attention model for keyframe selection	Comprehensive summaries, including static images and motion information	Developed a novel framework integrating spatiotemporal and motion features; Proposed a TED method for shot segmentation; Achieved competitive performance on multiple datasets.	Required optimization for real-time applications and various video types depends on the quality of transition detection and feature extraction.
[24]	Convolutional neural network architecture with a global diverse attention mechanism to capture pairwise temporal relations.	Various types, including user-generated videos, web videos, etc.	SumMe dataset, TVSum dataset, VTW dataset	F-score, Precision, Recall	Pre-trained CNN (GoogLeNet) for frame feature extraction	Video frames with extracted features	Convolutional neural network with a global diverse attention mechanism	Dynamic summaries based on frame-level importance scores	Developed a global diverse attention mechanism, Reduced computational costs, and Achieved state-of-the-art performance on multiple datasets.	Performance varied with different datasets; Required further optimization for specific video types.
[19]	Self-attention mechanism with relative position	Various types, including user-generated videos,	TVSum dataset, SumMe	F-score, Kendall's τ, Spearman's	Pre-trained CNN (GoogLeNet	Video frames sampled at 2fps	Self-attention mechanism with relative position	Dynamic summaries based on frame-level	Captured both local and global temporal	Performance varied with different datasets;

	embedding, combined with global and local input decomposition.	web videos, etc.	dataset	ρ	) for frame feature extraction		embedding	importance scores	dependencies; Improved performance on benchmark datasets with minimal redundancy.	Required further optimization for specific video types.
[17]	A new evaluation protocol called "Performance over Random" (PoR) accounts for the difficulty of data splits.	Various genres, including news, how-to's, documentaries, and user-generated content.	SumMe, TVSum	F-Score, PoR (Performance over Random), PoH (Performance over Human)	Not specified	Not specified	Not specified	Dynamic video summaries (video skims)	Introduced PoR to mitigate the weaknesses of the established evaluation protocol; Conducted extensive experiments to demonstrate PoR's reliability.	Not specified
[17]	Evaluation protocol using performance over random	Various genres, including news, how-to's, documentaries, and user-generated content	SumMe, TVSum	F-Score, Precision, Recall	Not explicitly mentioned	Frame-level importance scores	Not explicitly mentioned	Dynamic video summary (video skim)	Identified shortcomings of the established evaluation protocol; proposed a new evaluation approach considering data split difficulty; demonstrated enhanced representativenes	The study did not explicitly address feature extraction methods or specific neural network architectures used for summarization.

									s and reliability of performance results.	
[27]	Video summarization technique that integrates visual and audio information to create comprehensive summaries.	Lecture videos, educational content.	SumMe, TVSum	Viewing time reduction, comprehension level.	OCR is used for visual text, and speech recognition is used for audio text.	Video frames and audio transcriptions.	Scene change detection, OCR, speech recognition, text summarization, video synthesis.	Dynamic summaries that integrate visual and audio information.	Developed a method to preserve both visual and auditory context in video summaries, Significantly reducing viewing time without compromising comprehension.	Needs improvement in transition quality and audio clarity; User interface could be more intuitive.
[28]	Extraction of feature variables from video bitstreams, stepwise regression for dimensionality reduction, frame-level temporal subsampling using cosine similarity and PCA projections, followed by deep learning architectures.	Various types, including surveillance footage and other digital videos.	TVSum, SumMe, OVP, VSUMM	Precision, Recall, F-score, Computational time	HEVC video coding features	Video frames encoded with HEVC	LSTM networks, 1D-CNNs, Random forests	Static video summaries (keyframes)	Developed new techniques for feature extraction and temporal subsampling; Achieved high accuracy and efficiency; Demonstrated significant improvements over previous methods.	High computational requirements for certain deep learning architectures; Need for further optimization for different video types and real-time applications.
[29]	Causal learning techniques with a	Various types, including visual-	TVSum, QueryVS,	F1-score, state-of-the-	Causal semantics	Video frames and optional text-	Prior and posterior	Dynamic video summaries with	Enhanced model explainability	Potential challenges in

	causal semantics extractor for mutual information distillation.	textual input scenarios.	SumMe	art benchmarks	extraction, spatial-temporal feature extraction	based queries	probabilistic networks, helper distributions	improved explainability	through causal learning; Achieved state-of-the-art performance on benchmarks; Improved handling of multi-modal inputs.	predicting behaviors of data intervention and model outcome due to video noise and blur.
[23]	Deep learning-based video summarization focusing on user-selected objects of interest using YOLOv3.	Surveillance videos are user-generated content.	SumMe dataset, self-created dataset.	Accuracy, summarization rate, F1-score.	YOLOv3 for object detection.	Video frames with detected objects.	YOLOv3 deep learning model.	Object-based video summaries.	Developed a method for user-specific video summarization; Achieved high accuracy and summarization rates; Created a user-friendly graphical interface.	Potential challenges with low-resolution or low-light videos: Requires further testing across diverse video types.
[30]	Deep reinforcement learning with piecewise linear interpolation for keyframe selection.	Various types, including user-generated content and web videos.	SumMe, TVSum	F-score, Precision, Recall	Transformer and CNN-based network for feature extraction.	Video frames are represented as graph-level features.	Transformer encoder network combined with Pointwise Conv 1D network.	Dynamic video summaries with uniformly selected keyframes.	Introduced a method integrating deep reinforcement learning and interpolation; Enhanced performance through graph-level features and a temporal	The method showed lower performance on long videos due to high variance issues.

									consistency reward function.	
[31]	Deep Convolutional Neural Network (DCNN) integrated with video summarization using keyframes.	Various data types are available, including surveillance footage and other video data.	IXMAS, HMDB51, Breakfast, YouTube-8M, Kinetics-600	Precision, Recall, F-Measure, Classification Accuracy, Convergence Rate, Learnable Parameters	Pearson correlation coefficient (PCC) and color moments (CM) for keyframe extraction.	Keyframes extracted from video clips.	Lightweight DCNN with identity skip connections for gradient preservation.	Static video summaries (keyframes)	Developed an efficient DCNN architecture with feature reusability; Integrated keyframe extraction to reduce redundancy; Demonstrated superior performance across multiple datasets.	Computational efficiency depends on keyframe extraction quality; Performance may vary with different video resolutions and complexities.
[31]	Deep convolutional neural network (DCNN) integrated with video summarization using keyframes.	Various data types are available, including surveillance footage and other video data.	IXMAS, HMDB51, Breakfast, YouTube-8M, Kinetics-600	Precision, Recall, F-Measure, Classification Accuracy, Convergence Rate, Learnable Parameters	Pearson correlation coefficient (PCC) and color moments (CM) for keyframe extraction.	Keyframes extracted from video clips.	Lightweight DCNN with identity skip connections for gradient preservation.	Static video summaries (keyframes)	Developed an efficient DCNN architecture with feature reusability; Integrated keyframe extraction to reduce redundancy; Demonstrated superior performance across multiple	Computational efficiency depends on keyframe extraction quality; Performance may vary with different video resolutions and complexities.

									datasets.	
[32]	Transfer learning with Inception V3 and K-means clustering for keyframe extraction.	Wireless capsule endoscopy videos.	KID dataset and original WCE videos from a medical center.	F-measure, Compression Ratio	Inception V3 was used to generate embeddings of video frames.	Video frames extracted from WCE videos.	Inception V3 CNN with K-means clustering.	Static video summaries (keyframes).	Developed a method to reduce redundant frames in WCE videos; Achieved high F-measure and compression ratios; Demonstrated effectiveness in aiding medical diagnosis.	Performance depends on the quality of keyframe extraction; validation is required on a larger and more diverse dataset.
[33]	3D spatio-temporal U-Net combined with reinforcement learning.	General videos and medical videos (fetal ultrasound screenings).	SumME, TVSum, OVP, YouTube, and fetal ultrasound dataset.	F-score, Precision, Recall.	3D spatio-temporal CNN features and Inflated 3D (I3D) features.	The spatio-temporal video features from 3D CNNs.	3D spatio-temporal U-Net (3DST-UNet) with reinforcement learning.	Dynamic video summaries (keyframes and key shots).	Introduced a novel 3DST-UNet-RL framework; demonstrated superior performance on benchmark datasets; applied method to both general and medical video summarization tasks.	The framework required significant computational resources; performance could vary with different video types and dataset complexities.
[23]	The deep learning-based method with User Object of Interest (UOoI)	Surveillance videos	Not explicitly mentioned	Not explicitly mentioned	Convolution al Neural Networks (CNNs)	Video frames	CNN-based feature extraction followed by a deep learning-	Keyframes	Introduced a deep learning framework tailored to user-	It is not explicitly mentioned in the summary provided.

							based selection process		defined objects of interest for efficient video summarization and demonstrated improvements in summarization quality and relevance.	
[34]	Combination of Zernike Moments and R-Transform for feature extraction, KNN clustering, and AlexNet classifier.	Surveillance videos.	KARD, Weizmann, NUCLA, TVSum50, SumMe.	Accuracy, Precision, Recall.	Zernike Moments and R-Transform.	Silhouette images and extracted features.	Feature extraction with Zernike Moments and R-Transform, followed by KNN clustering and AlexNet classifier.	Keyframes highlighting abnormal activities.	Developed a framework combining machine learning and deep learning for efficient video summarization; achieved high accuracy and robustness in detecting abnormal activities.	Performance depends on the quality of silhouette images and feature extraction; validation is required on various datasets.
[25]	Deep reinforcement learning with unsupervised auxiliary summarization loss and a novel reward function.	General videos from various sources.	SumMe, TVSum	F-score, Precision, Recall	Convolutional Neural Network (CNN)	Video frames	Encoder-decoder architecture with CNN for feature extraction and LSTM for temporal dependency capture.	Dynamic video summaries	Introduced an unsupervised auxiliary summarization loss; Improved long-term dependency capture; Demonstrated	Computational efficiency for longer videos; Dependence on the quality of feature extraction methods.

									significant performance improvements on benchmark datasets.	
[35]	Deep reinforcement learning with shot-level semantic reward function.	Various types, including general videos from different sources.	SumMe, TVSum, CoSum, VTW	Precision, Recall, F-Score	Convolutional neural network (CNN) for convolutional feature matrices.	Video frames with shot-level semantics.	Encoder-decoder model with CNN and bidirectional LSTM.	Dynamic video summaries.	Introduced shot-level semantic reward to improve summarization; Achieved state-of-the-art performance on multiple datasets.	The method's performance could vary with different video types; it required robust shot-level semantic extraction for optimal results.