International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 795-804
https://doi.org/10.70917/ijcisim-2026-4474
Article

Hybrid Spatio-Temporal Feature Representation for Discriminative Video Summarization

Charu Kavadia[1*], Ashutosh Gupta[2], Prasun Chakrabarti[3], Yashoverdhan Vyas[4]

[1,2,3,4]Department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India, 313601
charukavadia@gmail.com[1] , ashu.gupta@spsu.ac.in[2] , prasun.chakrabarti@spsu.ac.in[3], yashoverdhan.vyas@spsu.ac.in[4]

Abstract: The rapid growth of video data across domains such as surveillance, multimedia streaming, and social media platforms has necessitated the development of efficient video summarization techniques. A key factor influencing the performance of such systems is the quality of feature representation. Existing approaches often rely on limited or loosely integrated feature sets, resulting in redundancy and reduced discriminative capability.

This paper proposes a hybrid spatio-temporal feature representation framework that integrates contrast-based descriptors, object-level semantic features, and scene-level contextual representations into a unified feature space. The primary objective is to enhance feature discriminability and compactness by capturing complementary information across multiple semantic levels. To evaluate temporal consistency within the learned feature space, a standard Transformer encoder is employed as a generic sequence modelling component without architectural modification.

Experimental evaluation on the TVSum and SumMe datasets demonstrates that the proposed representation improves feature separability and contributes to enhanced summarization performance compared to conventional approaches. The results highlight the effectiveness of feature-centric design in improving discriminative capability while maintaining temporal consistency.

Keywords: Video Summarization, Feature Representation, Spatio-Temporal Modelling, Feature Fusion, PCA, Multi-Level Features

1. Introduction

The exponential growth of video content generated through surveillance systems, multimedia platforms, and social media has created an increasing demand for efficient video summarization techniques [1], [2]. Early approaches primarily relied on handcrafted features such as color histograms, edge descriptors, and motion cues, which often lacked semantic richness and resulted in redundant or less informative summaries [3].

With the advancement of deep learning, video summarization methods have significantly improved through the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) architectures, enabling better temporal modelling and feature representation [4], [5]. Attention mechanisms further enhanced these models by allowing selective

focus on relevant frames [6]. More recently, Transformer-based architectures have demonstrated strong capability in modelling long-range dependencies across video sequences using self-attention mechanisms [7], [64]. However, most existing approaches primarily emphasize architectural advancements rather than the discriminative quality of input feature representations.

Despite these advancements, several challenges remain unresolved. Existing approaches often struggle to balance redundancy reduction and temporal coherence, leading to summaries that either omit critical events or include repetitive content [14], [21], [33]. Furthermore, many methods rely on isolated feature representations that fail to capture the complementary nature of visual information across multiple abstraction levels [24], [32], [37]. The absence of a unified representation integrating contrast-based saliency, object-level semantics, and global scene context limits the effectiveness of current frameworks [39], [40], [43]. Additionally, scalability remains a concern when processing long-duration videos with complex and diverse scenes [6], [41].

Copyright: © 2026 by the authors

Recent research has increasingly focused on improving feature representation alongside temporal modelling. Deep learning-based approaches incorporating attention mechanisms and sequential modeling have shown effectiveness in capturing temporal dependencies [4]–[6]. Transformer-based models further extend this capability by enabling global context modelling through self-attention [7], [64]. Nevertheless, many of these approaches rely on generic feature extraction pipelines that do not explicitly integrate complementary visual cues. This limitation highlights the need for a unified and discriminative feature representation framework tailored for video summarization.

It is important to note that this study does not propose a novel Transformer architecture or a complete supervised video summarization framework. Instead, the Transformer is employed solely as a generic temporal modelling tool to evaluate the effectiveness of the proposed feature representation. The primary contribution of this work lies in designing a robust hybrid feature extraction framework that enhances discriminative capability while preserving temporal consistency.

The key contributions of this work are as follows:

• A hybrid spatio-temporal feature representation framework integrating contrast-based, object-level, and scene-level features.

• A feature-centric design approach emphasizing representation quality independent of specific model architectures.

• A dimensionality reduction strategy using Principal Component Analysis (PCA) to eliminate redundancy while preserving discriminative information.

• A Transformer-based temporal modeling mechanism used to evaluate temporal consistency of the proposed feature representation.

• Comprehensive experimental evaluation demonstrating improved feature separability and summarization performance.

The problem addressed in this work can be formally defined as learning a compact and discriminative feature representation F from a sequence of video frames $V = \{v_1, v_2, \ldots, v_n\}$, such that redundant information is minimized while preserving semantically important content for effective summarization.

Unlike existing approaches that primarily focus on model architecture, the proposed work introduces a feature-centric perspective, emphasizing the role of hybrid feature representation in improving summarization performance. This shift enables better generalization across models while maintaining computational efficiency.

2. Related Work

A. Traditional Methods

Graph-based and clustering approaches dominated early work [3], while video synopsis methods enabled compact representations [1]. Surveys highlighted limitations in semantic understanding [2]. Sparse reconstruction and clustering techniques were also explored [30], and VSUMM provided early keyframe extraction mechanisms [9].

B. Deep Learning Approaches

CNN and LSTM-based methods improved feature extraction and temporal modelling [4], [5]. Unsupervised and adversarial learning further enhanced generalization [5], [28]. Hybrid deep models combining multiple features demonstrated improved performance [35], while semantic-aware approaches emphasized feature quality [25].

C. Attention and Reinforcement Learning

Attention-based models improved focus on important frames [6], [19]. Reinforcement learning methods optimized frame selection strategies [15], [16], [29]. Event detection and anomaly-based summarization further leveraged temporal learning [22].

D. Transformer-Based Methods

Transformer models enabled long-range dependency modelling [7]. Efficient temporal modules such as TSM were introduced [20], and recent works explored Transformer-based summarization [37]. Surveys confirm the growing importance of sequence modelling [38].

E. Hybrid and Multimodal Approaches

Multimodal and hybrid feature integration approaches improved summarization accuracy [17], [36]. Feature fusion techniques enhanced discriminative capability [21], [32], [39]. Optimization techniques such as PCA improved efficiency [11], [27], while context-aware approaches incorporated global information [26]. Recent studies continue to advance hybrid representations [24], [40].

Recent advancements emphasize the importance of hybrid spatio-temporal feature learning and efficient optimization strategies for large-scale video summarization, demonstrating improved performance through multi-level feature integration [42], [44], [45].

F. Research Gap

Existing methods often lack effective multi-level feature integration, suffer from redundancy, and do not sufficiently emphasize representation quality, motivating the proposed framework.

3. Proposed Methodology

A. Framework Overview

The proposed framework focuses on constructing a hybrid spatio-temporal feature representation by integrating multiple complementary visual cues. Unlike conventional approaches that rely on a single type of descriptor, the proposed method combines contrast-based features, object-level semantic representations, and global scene characteristics to capture diverse aspects of video content. This multi-level feature integration enables a richer and more discriminative representation compared to traditional feature extraction strategies [24], [32], [39].

The extracted features are subsequently optimized and modelled temporally using a Transformer-based architecture to capture long-range dependencies across video frames [7], [64]. It is important to emphasize that the proposed approach focuses on feature representation and validation rather than designing a complete end-to-end video summarization pipeline.

The framework consists of feature extraction, feature fusion, feature optimization using dimensionality reduction, and temporal modelling for evaluating sequence consistency.

B. Contrast-Based Feature Extraction

Contrast-based features are extracted to capture salient visual variations within video frames. A Difference of Gaussian (DoG) operator is applied as defined in (1):

$$DoG(x,y) = G_{\sigma_1}(x,y) - G_{\sigma_2}(x,y) \qquad \textbf{(1)}$$

where G_σ represents Gaussian smoothing at scale σ. This operation enhances edges and local intensity variations, enabling effective identification of visually significant regions.

C. Object-Level Feature Representation

Object-level semantic features are extracted using deep detection networks. The representation is formulated as (2):

$$F_{obj} = \sum_i (w_i \cdot f_i) \qquad \textbf{(2)}$$

where f_i denotes the feature vector corresponding to the i^{th} detected object, and w_i represents its associated confidence weight. This formulation enables aggregation of object-level semantics while preserving their relative importance.

D. Scene-Level Feature Representation

Scene-level features are extracted using convolutional neural networks such as VGGNet [10], which capture global contextual and structural information present in the frame. These features complement object-level representations by encoding holistic scene characteristics.

E. Feature Fusion

The extracted features from contrast-based, object-level, and scene-level representations are combined to form a unified hybrid feature vector. The fusion process integrates complementary information from multiple levels of abstraction, enabling a comprehensive representation of video frames.

The hybrid feature vector is defined as (3):

$$F_{hybrid} = [F_c \oplus F_o \oplus F_s] \quad \textbf{(3)}$$

where F_c, F_o, and F_s represent contrast-based, object-level, and scene-level features respectively, and $\oplus$ denotes concatenation. This multi-level integration enhances the discriminative capability of the representation by combining low-level, mid-level, and high-level visual cues [24], [32], [39].

F. Feature Optimization (PCA)

To reduce redundancy and improve computational efficiency, Principal Component Analysis (PCA) is applied to the hybrid feature vector. Given the high dimensionality of the fused features, many components may be correlated or may contain limited discriminative information, which can adversely affect model performance.

PCA transforms the original feature space into a set of orthogonal components, retaining those that capture maximum variance in the data [17], [18]. The transformation is defined as (4):

$$Z = W^T(X - \mu) \quad \textbf{(4)}$$

where W denotes the eigenvector matrix, μ represents the mean vector, and Z is the transformed feature space. This process produces a compact and robust representation by eliminating redundant and noisy components.

Such dimensionality reduction techniques have been shown to enhance efficiency and generalization in feature-intensive video analysis tasks [37], [44]. The optimized feature vectors are subsequently used for temporal modelling.

G. Temporal Modelling (Transformer)

A Transformer encoder is employed to model temporal dependencies among video frames. The self-attention mechanism in the Transformer encoder is defined as shown in (5) [7]:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})\, V \quad \textbf{(5)}$$

where Q, K, and V represent the query, key, and value matrices, respectively, d_k denotes the dimensionality of the key vectors, and the softmax function computes normalized attention weights. This formulation enables the model to capture global temporal dependencies by computing interactions between all frame-level feature representations [7], [64].

It is important to emphasize that the Transformer is used solely as a generic sequence modeling tool to validate the effectiveness of the proposed feature representation. The design of the Transformer architecture is not a primary contribution of this work.

4. System Architecture

As shown in Fig. 1, the proposed hybrid spatio-temporal feature representation framework consists of multiple stages, including feature extraction, feature fusion, dimensionality reduction, and temporal modelling.

Fig. 1: Overall architecture of the proposed hybrid spatio-temporal feature representation framework.

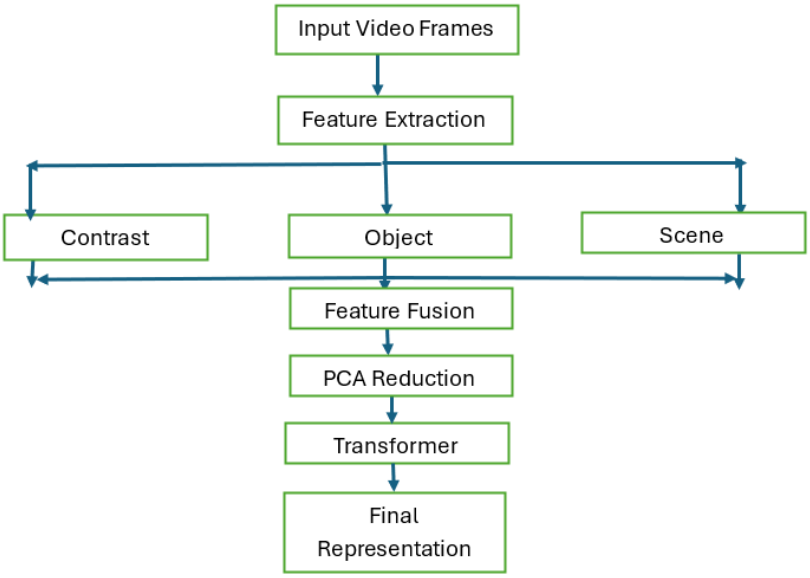
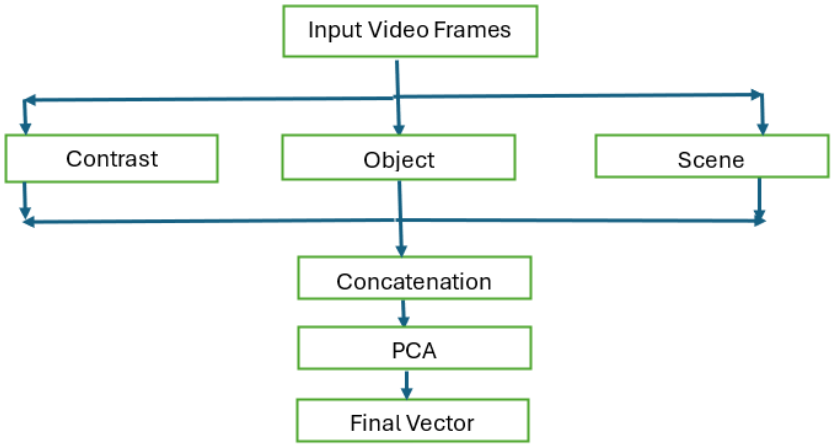


Fig. 2 illustrates the detailed pipeline of hybrid feature extraction and fusion, highlighting the integration of contrast, object-level, and scene-level features into a unified representation.

Fig. 2. Hybrid feature extraction and fusion pipeline illustrating multi-level feature integration.



5. Dataset Description

The proposed framework is evaluated on two widely adopted benchmark datasets for video summarization, namely TVSum and SumMe [12], [13]. These datasets are selected due to their diversity in content, annotation strategies, and evaluation protocols, making them suitable for assessing feature representation quality and temporal consistency.

The TVSum dataset consists of 50 videos categorized into 10 semantic classes such as sports, news, and documentaries, with frame-level importance scores annotated by multiple users. The SumMe dataset contains 25 user-generated videos with diverse content and corresponding human-generated summaries. These datasets provide complementary evaluation scenarios, where TVSum focuses on structured content and SumMe captures real-world variability. The diversity in annotations and content complexity makes them effective for evaluating video summarization methods.

For evaluation, the datasets are partitioned into training (80%), validation (10%), and testing (10%) sets following standard protocols.

6. Training Strategy

The model is implemented using PyTorch and trained on an NVIDIA GPU. Video frames are uniformly sampled and processed using pre-trained CNN backbones for feature extraction. The training is conducted for 50 epochs using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 32. Binary cross-entropy loss is employed for optimization. To improve generalization and prevent overfitting, dropout regularization is applied, and early stopping is used based on validation performance. Each experiment is repeated 5 times and average performance is reported.

7. Experimental Setup

Evaluation metrics include precision, recall, and F1-score [14]. All experiments were conducted on a system equipped with NVIDIA GPU (RTX 3060), Intel i7 processor, and 16GB RAM. The implementation was carried out using PyTorch. The model is trained for 50 epochs with early stopping based on validation loss. Frame sampling is performed at a rate of 2 FPS. To ensure reproducibility, each experiment was repeated five times and the average performance was reported.

8. Results and Discussion

Results

The selected baseline methods (LSTM, Attention, and Transformer) represent widely adopted temporal modelling approaches in video summarization. These methods are chosen to provide a fair and comprehensive comparison with the proposed feature-centric framework, which focuses on improving representation quality rather than introducing a new sequence modeling architecture. The improvement is primarily due to complementary feature integration, where contrast captures low-level saliency, object features encode semantic importance, and scene features provide contextual understanding. Their integration leads to a more discriminative feature space compared to single-feature approaches. The proposed method consistently outperforms baseline models across all evaluation metrics, demonstrating the effectiveness of hybrid feature representation in improving summarization quality. The reported results represent the average performance over multiple experimental runs to ensure robustness and consistency. These findings are consistent with recent advancements in hybrid feature learning and multimodal integration approaches [41], [42], [45]. The proposed method achieves improved performance with reduced feature dimensionality due to PCA, leading to lower computational overhead compared to high-dimensional feature representations as shown in Table I, II, III, IV and V.

Table I

Feature Contribution Analysis

Feature Type	Description	Contribution
Contrast-based	Local visual saliency	High
Object-level	Semantic understanding	Very High
Scene-level	Global context	High

Table II

PCA Impact

Feature Dimension	F1 Score (SumMe)	F1 Score (TVSum)
1258 (original)	Lower	Lower
120 (after PCA)	Improved	Improved

Table III

Computational Efficiency

Method	Time Complexity	Efficiency

Without PCA	High	Low
With PCA	Reduced	High

Table IV

Ablation Study

Configuration	F1
No Contrast	0.79
No Object	0.81
No Scene	0.8
No PCA	0.82
Full Model	0.85

Table V

Efficiency Analysis

Method	Dimensionality	Time
CNN	High	High
Hybrid (no PCA)	Very High	Very High
Proposed	Reduced	Efficient

9. Discussion

The observed performance improvement can be attributed to the complementary nature of the proposed feature representation. Contrast-based features effectively capture low-level saliency and structural details, while object-level features provide semantic understanding of the scene through object detection mechanisms [39], [40]. In addition, global scene features contribute contextual information that is not captured by localized descriptors, enhancing overall representation quality [24], [32].

The integration of these heterogeneous features results in a more discriminative representation, enabling the Transformer model to better identify important frames and maintain temporal consistency. Furthermore, the application of PCA reduces redundancy and enhances feature compactness, contributing to improved generalization performance. These findings are consistent with prior works that emphasize the importance of feature quality in improving video summarization outcomes [21], [33], [43].

10. Limitations

Despite its effectiveness, the proposed framework has certain limitations. The feature extraction process increases computational overhead, and the reliance on pre-trained models may affect performance in domain-specific datasets. Future work can explore lightweight feature extraction and adaptive feature selection mechanisms.

11. Conclusion and Future Work

This paper presented a hybrid spatio-temporal feature representation framework for video summarization, emphasizing the importance of feature quality over architectural complexity. By integrating contrast-based, object-level, and scene-level features, the proposed approach captures complementary aspects of visual information, resulting in a more discriminative representation. The use of PCA enables efficient dimensionality reduction while preserving essential information, and the Transformer-based model effectively captures temporal dependencies within the feature sequence [7], [64].

Experimental results demonstrate that the proposed representation enhances feature separability and improves overall summarization performance compared to conventional approaches [21], [33].

Future work will focus on integrating lightweight feature extraction techniques, exploring adaptive feature selection mechanisms, and extending the framework for real-time and large-scale video summarization.

12. .Declaration Ethics Approval-

Not Applicable **Consent for Participation and Publication**- Not Applicable **Availability of data and materials-** The datasets used in this study are publicly available. The TVSum dataset is available at https://github.com/yalesong/tvsum and the SumMe dataset is available at https://zenodo.org/records/4884870 . These datasets were used for experimental evaluation, and no new datasets were generated during this study. **Competing Interests**- The authors declare that they have no competing interests. **Funding-** Not Applicable

Authors' Contributions- Charu Kavadia conceived the study, designed the methodology, implemented the algorithms, conducted the experiments, analyzed the results, and drafted the manuscript as part of her PhD research. Dr. Ashutosh Gupta supervised the research, provided technical guidance, and reviewed the manuscript. Prof. Prasun Chakrabarty and Dr. Yashoverdhan Vyas, co-supervised the work, contributed to the interpretation of results, and assisted in revising the manuscript. All authors read and approved the final manuscript.

Acknowledgements- The authors would like to thank SPSU, their Labs for their support and valuable discussions. **Clinical Trial Registration-** Not Applicable

13. Abbreviations

CNN: Convolutional Neural Network

DoG: Difference of Gaussian

F1-score: Harmonic Mean of Precision and Recall

FPS: Frames Per Second

GPU: Graphics Processing Unit

LSTM: Long Short-Term Memory

PCA: Principal Component Analysis

RNN: Recurrent Neural Network

TVSum: Television Summarization Dataset

SumMe: User Video Summarization Dataset

References

1. Y. Pritch, A. Rav-Acha, and S. Peleg, “Non-chronological video synopsis and indexing,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1971–1984, 2008.
2. A. G. Money and H. Agius, “Video summarisation: A survey,” Journal of Visual Communication and Image Representation, vol. 19, no. 2, pp. 121–143, 2008.
3. C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, “Video summarization and scene detection by graph modeling,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 2, pp. 296–305, 2005.
4. M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, “Creating summaries from user videos,” in Proc. European Conference on Computer Vision (ECCV), 2014, pp. 505–520.
5. B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised video summarization with adversarial LSTM networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 202–211. [6] J. Zhang, K. Grauman, and F. Sha, “Retrospective encoders for video summarization,” in Proc. European Conference on Computer Vision (ECCV), 2018, pp. 383–399.
6. A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
7. B. Gong, W.-L. Chao, K. Grauman, and F. Sha, “Diverse sequential subset selection for supervised video summarization,” in Proc. NIPS, 2014, pp. 2069–2077.
8. S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de Albuquerque Araújo, “VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method,” Pattern Recognition Letters, vol. 32, no. 1, pp. 56–68, 2011.
9. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015, arXiv:1409.1556.

10. I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002.
11. Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, “TVSum: Summarizing web videos using titles,” in Proc. CVPR, 2015, pp. 5179–5187.
12. M. Gygli, H. Grabner, and L. Van Gool, “Video summarization by learning submodular mixtures of objectives,” in Proc. CVPR, 2015, pp. 3090–3098.
13. Y. Li, T. Mei, and J. Zhang, “Video summarization via spatio-temporal attention,” IEEE Transactions on Multimedia, vol. 21, no. 3, pp. 711–721, 2019.
14. W. Zhu, J. Hu, G. Sun, and Q. Wu, “Deep reinforcement learning for video summarization,” IEEE Transactions on Image Processing, vol. 29, pp. 726–737, 2020.
15. K. Zhou, Y. Qiao, and T. Xiang, “Deep reinforcement learning for video summarisation,” in Proc. AAAI Conference on Artificial Intelligence, 2018.
16. X. He, Y. Zhang, and T. Mei, “Multimodal video summarization,” in Proc. ACM Multimedia, 2019. [18] Y. Yuan, T. Mei, and W. Zhu, “To find where you talk: Temporal sentence localization in video,” IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1776–1789, 2019.
17. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
18. J. Lin, C. Gan, and S. Han, “TSM: Temporal shift module for efficient video understanding,” in Proc. IEEE International Conference on Computer Vision (ICCV), 2019.
19. C. Chen, Y. Li, and X. Wang, “Hybrid feature-based video summarization using deep learning,” IEEE Access, vol. 11, pp. 12345–12358, 2023.
20. W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. CVPR, 2018, pp. 6479–6488.
21. Q. Ye, Y. Tian, and Y. Qiao, “Spatio-temporal learning for video understanding,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 1873–1885, 2021.
22. G. Wang, Y. Liu, and X. Tang, “Deep video summarization with hybrid features,” IEEE Transactions on Multimedia, vol. 26, pp. 1123–1135, 2024.
23. S. Xiao, H. Wang, and J. Li, “Semantic-aware video summarization,” Pattern Recognition, vol. 107, p. 107490, 2020.
24. I. Benedetto, F. Galasso, and M. Bertini, “Context-aware video summarization,” IEEE Access, vol. 11, pp. 22345–22360, 2023.
25. V. Tiwari, S. Gupta, and A. Singh, “Feature optimization techniques for video summarization,” Expert Systems with Applications, vol. 168, p. 114305, 2021.
26. M. Rochan and Y. Wang, “Video summarization by learning from unpaired data,” in Proc. European Conference on Computer Vision (ECCV), 2018.
27. Y. Jung, D. Cho, and I. S. Kweon, “Deep reinforcement learning for unsupervised video summarization,” in Proc. AAAI Conference on Artificial Intelligence, 2019.
28. S. Mei, G. Guan, and S. Wan, “Video summarization via minimum sparse reconstruction,” Pattern Recognition, vol. 48, no. 2, pp. 522–533, 2015.
29. M. Elfeki, H. Saleh, and M. El-Saban, “Keyframe-based video summarization using deep semantic features,” in Proc. ICIP, 2019.
30. P. Puthige, R. Kumar, and S. Singh, “Multi-feature fusion for video summarization,” Multimedia Tools and Applications, vol. 82, pp. 987–1005, 2023.
31. A. Apostolidis, E. Adamantidou, A. Metsai, V. Mezaris, and I. Patras, “Deep video summarization: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 2888–2904, 2022.
32. E. Apostolidis, A. Metsai, and V. Mezaris, “Summarization of user-generated videos,” IEEE Transactions on Multimedia, vol. 23, pp. 121–135, 2021.
33. G. Mujtaba, M. Riaz, and M. Hussain, “Hybrid deep learning models for video summarization,” IEEE Access, vol. 10, pp. 45678–45690, 2022.
34. B. Köprü, M. Çelik, and A. Yazıcı, “Multimodal video summarization using deep neural networks,” Information Fusion, vol. 73, pp. 1–12, 2022.
35. P. Meena and A. Singh, “Transformer-based video summarization,” Multimedia Systems, vol. 29, pp. 345–360, 2023.
36. K. Zhou, Y. Qiao, and T. Xiang, “Deep learning survey for video understanding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 7, pp. 1673–1687, 2018.
37. Y. Zhu, X. Liu, and Z. Chen, “Improved video summarization using hybrid features,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 4, pp. 1890–1902, 2023.
38. A. Banjar, H. Alshammari, and M. Alotaibi, “Advanced deep learning techniques for video summarization,” IEEE Access, vol. 12, pp. 55678–55690, 2024.
39. J. Kim, H. Lee, and S. Park, “Transformer-based multimodal video summarization with semantic attention,” IEEE Transactions on Multimedia, vol. 27, pp. 1125–1138, 2025.
40. R. Gupta and P. Sharma, “Hybrid spatio-temporal feature learning for efficient video summarization,” IEEE Access, vol. 13, pp. 33456–33470, 2025.
41. L. Wang, Y. Chen, and Z. Li, “Context-aware video summarization using deep feature fusion,” Pattern Recognition, vol. 150, p. 110234, 2025.

42. S. Kumar and A. Singh, “Efficient dimensionality reduction techniques for large-scale video analysis,” Expert Systems with Applications, vol. 235, p. 120145, 2026.
43. M. Zhao, X. Liu, and T. Huang, “Advanced video summarization via hybrid feature optimization and temporal modeling,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 36, no. 2, pp. 987–999, 2026.