Video Summarization Techniques: A Comprehensive Review

Toqa Alaa[1], Ahmad Mongy[1], Assem Bakr[1], Mariam Diab[1] and Walid Gomaa[1,2]
[1]*Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology, Alexandria, Egypt*
[2]*Faculty of Engineering, Alexandria University, Alexandria, Egypt*
{*toqa.alaa, ahmad.aboelnaga, asem.abdelhamid, mariam.diab, walid.gomaa*}*@ejust.edu.eg*

Keywords: Keyframe Selection, Event-Based Summarization, Supervised Methods, Unsupervised Methods, Attention Mechanism, Multi-Modal Learning, Generative Adversarial Networks.

Abstract: The rapid expansion of video content across various industries, including social media, education, entertainment, and surveillance, has made video summarization an essential field of study. This survey aims to explore the latest techniques and approaches developed for video summarization, with a focus on identifying their strengths and drawbacks to guide future improvements. Key strategies such as reinforcement learning, attention mechanisms, generative adversarial networks, and multi-modal learning are examined in detail, along with their real-world applications and challenges. The paper also covers the datasets commonly used to benchmark these techniques, providing a comprehensive understanding of the current state and future directions of video summarization research.

1 INTRODUCTION

The rapid advancement of technology has integrated camcorders into many devices, leading to an explosion of video content as people capture and share their daily lives on social media (Pritch et al., 2008). Traditional video representation methods, such as viewing consecutive frames, struggle to meet the demands of modern multimedia services like content-based search, retrieval, and navigation. To address these challenges, automatic video content summarization and indexing techniques have been developed, enabling more efficient access and categorization of video content (Pritch et al., 2008). The growing interest in video summarization is evident from the increasing number of research papers published annually, as shown in Figure 1.

Video summarization methods can be broadly categorized into static, event-based, and personalized approaches. Static summarization selects keyframes representing significant scenes or events, while event-based methods focus on summarizing specific actions, such as in sports videos (Banjar et al., 2024). Personalized summarization tailors content based on user preferences, generating topic-related summaries to meet individual needs (Zhu et al., 2023).

Recent advancements in deep learning have enhanced video summarization through attention mechanisms and reinforcement learning, enabling models to generate more concise and meaningful summaries (Zhou et al., 2018). Multi-modal learning further enriches summaries by integrating audio, text, and visual features (Zhu et al., 2023).

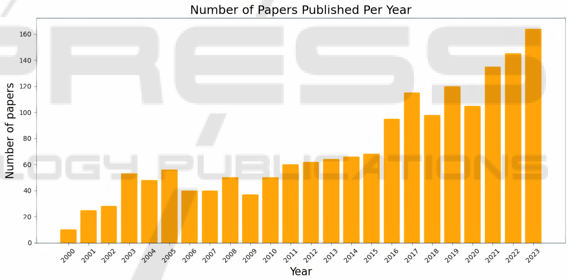

Figure 1: The number of papers per year containing in their title the phrase ”video summarization” according to Google Scholar.

Research has also explored single-view and multi-view summarization strategies, with single-view focusing on a single perspective and multi-view integrating information from multiple viewpoints (Parihar et al., 2021). Additionally, the inclusion of interactive features, such as user feedback mechanisms, has made video summarization tools more versatile and user-friendly (Wu et al., 2022). Given the rapid advancements in video summarization techniques and their diverse applications, this paper seeks to answer the following research questions:

Alaa, T., Mongy, A., Bakr, A., Diab, M. and Gomaa, W.
Video Summarization Techniques: A Comprehensive Review.
DOI: 10.5220/0012936400003822

In *Proceedings of the 21st International Conference on Informatics in Control, Automation and Robotics (ICINCO 2024) - Volume 2*, pages 141-148
ISBN: 978-989-758-717-7; ISSN: 2184-2809

1. What are the most recent and effective techniques for video summarization, and how do they perform across different tasks?
2. What are the common limitations and challenges of these techniques in real-world scenarios?
3. How can future research address these limitations to improve the efficiency and accuracy of video summarization?

This paper is organized as follows: Section 1 provides an overview of video summarization; Section 2 examines various techniques; Section 3 explores relevant datasets; Section 4 discusses applications; and Section 5 concludes the paper with future directions.

2 VIDEO SUMMARIZATION TECHNIQUES

This section discusses the various techniques used in video summarization. These techniques can be grouped according to the learning method, the type of the extracted features, the type of the input video, or the type of the output summary.

2.1 Learning Paradigm-Based Methods

Video summarization techniques can be categorized into supervised, unsupervised, weakly supervised, and reinforcement methods.

2.1.1 Supervised Methods

Supervised video summarization involves using annotated or labeled data during the training phase to learn how to generate summarizes. These methods involve datasets where each video is paired with a corresponding summary that highlights the most important events in the video. Supervised methods in video summarization typically use Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) Networks, or Transformer architecture (He et al., 2023; Zhu et al., 2023; Lin et al., 2022). These methods use large labeled datasets to learn patterns but require extensive manual labeling and may not generalize well to new or varied content.

2.1.2 Unsupervised Methods

Unsupervised video summarization has gained significant attention, particularly through clustering methods and generative adversarial networks (GANs). The approach in (Mahmoud et al., 2013) relies on clustering color and texture features extracted from the video frames using a modified Density-based clustering non-parametric algorithm (DBSCAN) (Ester et al., 1996) to summarize the video content. The original video undergoes pre-sampling. Next, color features are extracted using a color histogram in the HSV color space, and texture features are extracted using a two-dimensional Haar wavelet transform in the HSV color space. Video frames are then clustered with the modified DBSCAN algorithm, and key frames are selected. Finally, these key frames are arranged in their original order to aid visual understanding. The authors in (Lee et al., 2012) used clustering of frame color histograms to segment temporal events. Temporal video segmentation is used in (Potapov et al., 2014) for detecting shot or scene boundaries. This approach takes into account the differences between all pairs of frames.

GAN-based approaches involve adversarial training, where a generator attempts to create summaries that fool a discriminator into thinking they are user-generated. This method is exemplified by the Fully Convolutional Sequence Network (FCSN) (Rochan and Wang, 2019), which selects key frames, while a summary discriminator differentiates between real and artificial summaries. Figure 2 illustrates the interaction between the generator and discriminator in this process. Despite their effectiveness, many GAN-based methods that incorporate LSTM units face challenges with low variation in frame-level importance scores, limiting their impact. To address this, Attentive Conditional GANs (AC-GANs) (He et al., 2019) use a BiLSTM to predict importance scores, while (Lin et al., 2022) introduced a deep semantic and attentive network (DSAVS) to minimize the distance between video and text representations, incorporating self-attention for capturing long-range dependencies. Additionally, the architecture proposed

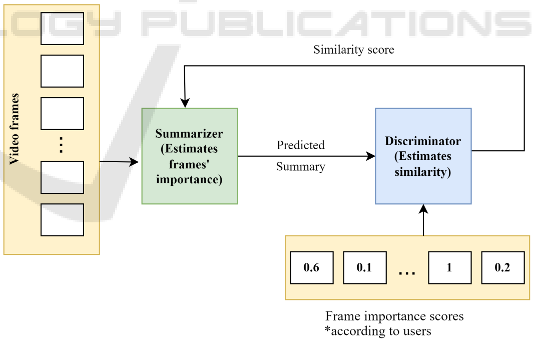

Figure 2: Generator and discriminator in video summarization process.

in (Apostolidis et al., 2020) embeds an Actor-Critic model within a GAN, framing the selection of key fragments as a sequence generation task that is incrementally refined through rewards from the discriminator.

2.1.3 Weakly Supervised Methods

Video Summarization has a very high cost for data-labeling tasks, so it is desirable for the machine learning techniques to work with weak supervision. There are three types of weak supervision. The first is incomplete supervision, only a subset of training data is given with labels while the other data remain unlabeled. The second type is inexact supervision, i.e., only coarse-grained labels are given, such as in (Ye et al., 2021) in which the model can learn to detect highlights by mining video characteristics with video level annotations (topic tags) only. The third type is inaccurate supervision, i.e., the given labels are not always ground-truth. Such a situation occurs, e.g., when the image annotator is careless or weary, or some images are difficult to categorize. A contextual temporal video encoder and a segment scoring transformer are used in (Narasimhan et al., 2022) to rank segments by their significance. This approach avoids the need for manual annotations and enhances scalability for large datasets.

2.1.4 Reinforcement Learning Methods

The process of video summarization is inherently sequential, as the choice of one segment can influence the importance and selection of subsequent segments. RL is well-suited for sequential decision-making tasks, enabling the model to optimize the selection process over time to achieve the best summary.

An end-to-end, reinforcement learning-based framework was proposed in (Zhou et al., 2018). It designs a novel reward function that jointly accounts for diversity and representativeness of generated summaries and does not rely on labels or user interactions at all. However, this approach suffered from some limitations. The sparse reward problem leads to hard convergence as an agent receives the reward only after the whole summary is generated in conventional reinforcement learning methods. The authors in (Wang et al., 2024) tried to address this problem by proposing a Progressive Reinforcement Learning network for Video Summarization (PRLVS) in an unsupervised learning fashion. In this method, the summarization task is decomposed hierarchically. The model trains an agent to modify the summary progressively. The agent chooses to replace one frame with some frame in the neighborhood and receives a reward for the whole summary at each step.

Unsupervised video summarization with reinforcement learning and a 3D spatio-temporal U-Net is implemented by (Liu et al., 2022) to efficiently encode spatio-temporal information of the input videos for downstream reinforcement learning.

2.2 Input-Based Methods

Video summarization techniques differ depending on whether the input is single-view or multi-view, as well as the modality of the input video.

2.2.1 SingleView Methods

Single View video summarization has been extensively researched, leading to various effective approaches.

Techniques such as equal partition-based clustering (Kumar et al., 2016) and key-frame selection using frame difference (Tirupathamma, 2017) are notable for their simplicity and efficiency in generating summaries. These methods typically focus on event detection and redundancy reduction by selecting key frames that capture significant changes in the video. Methods like SUM-GDA (Li et al., 2021) process a single view of visual data, applying attention mechanisms to focus on important visual elements in the video. The authors in (Lin et al., 2022) introduced a new framework for video summarization called Deep Hierarchical LSTM Networks with Attention (DHAVS). This methodology extracts spatio-temporal features using 3D ResNeXt-101 and employs a deep hierarchical LSTM network with an attention mechanism to capture long-range dependencies and focus on important keyframes. Additionally, it proposes a cost-sensitive loss function to improve the selection of critical frames under class imbalance. DHAVS was tested on the SumMe and TVSum datasets, and the results show that DHAVS outperforms state-of-the-art methods, demonstrating significant advancements in video summarization. This technique operates on a single video view but incorporates both spatial and temporal aspects. Single-view video summarization methods are simpler and more efficient, focusing on key events and significant changes from one perspective. However, they may miss broader context and important details that could be captured from multiple views.

2.2.2 MultiView Methods

MultiView video summarization excels in capturing diverse perspectives by avoiding the redundancy and repetition inherent in single-view approaches (Elfeki et al., 2022). Methods like seqDPP (Gong et al., 2014) and Multi-DPP (Elfeki et al., 2022) ensure that summaries are both comprehensive and contextually rich by leveraging the temporal structures of multiple video streams. Multi-view video summarization captures diverse perspectives and reduces redundancy by integrating multiple video streams, resulting in comprehensive and contextually rich summaries. However, it comes with increased computational complexity and challenges in maintaining temporal coherence across streams.

2.2.3 Modality-Based Methods

Single modality approaches often struggle with complex scenes, leading to inaccurate predictions. Multimodal methods address this by integrating information from various modalities—visual, audio, and textual—thereby enhancing the quality and accuracy of video summaries. Visual modalities help segment videos into meaningful parts by detecting scenes, objects, and activities. Audio modalities use automatic speech recognition (ASR) to extract crucial dialogues and narrative elements, while textual modalities add context through subtitles and overlays. Recent advancements in multi-modal learning demonstrate that audio and visual modalities can share a consistency space, offering complementary perspectives on activities (Zhao et al., 2021). Techniques such as the AudioVisual Recurrent Network (AVRN) (Ye et al., 2021) improve summarization by combining these signals, showing significant performance gains on datasets like SumMe and TVsum. Furthermore, methods like those proposed in (Palaskar et al., 2019) leverage both visual and textual data to create summaries that are semantically rich and contextually relevant. Multi-modal Transformers have also been introduced to adaptively fuse these features, compensating for the limitations of single-modality approaches (Zhu et al., 2023; Narasimhan et al., 2021; He et al., 2023). These models, which include feature extraction, learning, and frame selection modules, have proven effective in generating high-quality video summaries by capturing cross-modal correlations and improving decision-making processes.

2.3 Output-Based Methods

This section discusses the video summarization methods according to the type of the output summary.

2.3.1 Generic Video Summarization

Generic video summarization creates a synopsis by selecting key parts of the video, often through storyboard summarization or video skimming. This type of summary is useful when users need a quick overview without prior knowledge of the content. While simple summaries can be generated by uniformly or randomly sampling frames, more advanced methods utilize video processing and computer vision techniques to select frames that best represent the video's semantics.

The GVSUM technique proposed in (Basavarajaiah and Sharma, 2021) offers a versatile approach applicable to various types of videos, including surveillance and sports. GVSUM extracts I-frames through partial decoding and groups them based on visual features using unsupervised clustering. Frames are included in the summary whenever there is a change in cluster number, indicating a shift in the visual scene.

2.3.2 Personalized Video Summarization

Personalized video summarization is a developing technology that creates customized video summaries based on individual viewer preferences. The work in (Zhu et al., 2023) emphasizes generating multiple topic-related summaries, addressing subjective viewer needs. This involves a multi-label classification task where each video frame is assessed by multiple binary classifiers to determine its relevance to various topics. Figure 3 illustrates the key developments in personalized video summarization, including the gap in research following the first publication.

This approach is particularly useful for creating personalized sports highlights, enabling viewers to focus on content relevant to their interests. The framework proposed in (Tejero-de Pablos et al., 2018) uses neural networks to personalize sports video summaries by analyzing player actions. Additionally, the method in (Banjar et al., 2024) tailors summaries to user-defined lengths by detecting excitement scores in the audio stream.

2.3.3 Query-Based Video Summarization

Query-based video summarization generates summaries tailored to specific user queries or topics, offering a more personalized approach than traditional methods that rely on predefined rules. The task is formulated in (Xiao et al., 2020) as computing similarity between video shots and the query, utilizing a convolutional network with local self-attention and query-aware global attention mechanisms to learn visual information.

The "IntentVizor" framework introduced in (Wu et al., 2022) advances interactive video summarization by integrating textual and visual queries. It features an intent module to extract user intent and

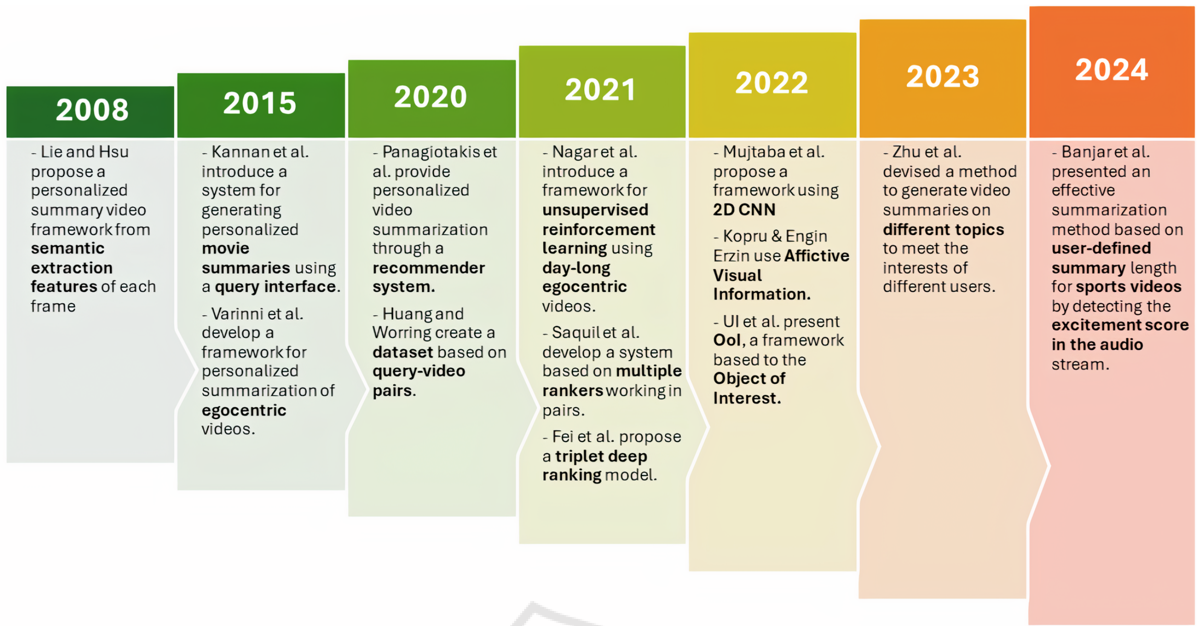
2008
- Lie and Hsu propose a personalized summary video framework from semantic extraction features of each frame
2015
- Kannan et al. introduce a system for generating personalized movie summaries using a query interface.
- Varinni et al. develop a framework for personalized summarization of egocentric videos.
2020
- Panagiotakis et al. provide personalized video summarization through a recommender system.
- Huang and Worring create a dataset based on query-video pairs.
2021
- Nagar et al. introduce a framework for unsupervised reinforcement learning using day-long egocentric videos.
- Saquil et al. develop a system based on multiple rankers working in pairs.
- Fei et al. propose a triplet deep ranking model.
2022
- Mujtaba et al. propose a framework using 2D CNN
- Kopru & Engin Erzin use Affictive Visual Information.
- Ul et al. present OoI, a framework based to the Object of Interest.
2023
- Zhu et al. devised a method to generate video summaries on different topics to meet the interests of different users.
2024
- Banjar et al. presented an effective summarization method based on user-defined summary length for sports videos by detecting the excitement score in the audio stream.

Figure 3: Milestones in personalized video summarization.

a summary module for summarization, both supported by the Granularity-Scalable Ego-Graph Convolutional Network (GSE-GCN). This framework allows for multi-modal queries, enabling users to adjust summarization interactively based on an adjustable distribution of learned basis intents.

3 DATASETS

In video summarization research, the choice of datasets is crucial for training and evaluating the performance of various techniques. Table 1 outlines the key datasets commonly used in the video summarization techniques, highlighting their characteristics, contents, and relevance to different summarization methods.

4 APPLICATIONS

Video summarization offers transformative benefits across various industries, enhancing the management and utilization of video content. In streaming platforms, video summarization is crucial for creating trailers and highlight reels, providing viewers with quick overviews of content and boosting engagement. As short-form video platforms gain popularity, the need for automated highlight identification from untrimmed videos has become even more pronounced (Ye et al., 2021).

Recommendation systems benefit from video summarization by refining content suggestions to match individual users' preferences. Topic-aware summarization, as proposed by (Zhu et al., 2023), further enhances this by generating summaries on various topics, catering to diverse user interests and improving the overall viewing experience.

In terms of efficient indexing and retrieval, summarized videos streamline the process by condensing content, making search and retrieval more efficient. This approach facilitates better keyword extraction (Apostolidis et al., 2021), reduces storage requirements, and accelerates processing times (Tiwari and Bhatnagar, 2021).

For anomaly detection in surveillance, video summarization supports the rapid deployment of models by leveraging pre-trained deep models and denoising autoencoders (DAEs)(Sultani et al., 2018). Frameworks like IntentVizor(Wu et al., 2022) enhance decision-making by providing valuable insights into unusual activities.

In education, video summarization helps students quickly grasp key concepts by distilling lengthy lectures and tutorials into concise summaries, thus improving the learning experience (Benedetto et al., 2023).

Social media platforms utilize video summarization to create engaging promotional content and product demonstrations. This method captures viewer attention effectively within a brief time frame (Sabha and Selwal, 2023).

Table 1: Datasets for video summarization.

Dataset	Description	Size	Duration (min)	Modality
TVSum (Song et al., 2015)	Title-based containing news, how-to, documentary, vlog, egocentric genres and 1,000 annotations of shot-level importance scores obtained via crowdsourcing.	50	2-10	Uni
SumMe (Gygli et al., 2014)	Each video is annotated with at least 15 human summaries.	25	1-6	Uni
BLiSS (He et al., 2023)	Contains livestream videos and transcripts with multimodal summaries (674 videos, 12,629 text)	13,303	5	Multi
TopicSum (Zhu et al., 2023)	Contains frame-level importance scores, and topic labels annotations.	136	5	Multi
Multi-Ego (Elfeki et al., 2022)	Contains videos recorded simultaneously by three cameras, covering a wide variety of real-life scenarios.	41	3-7	Uni
How2 (Palaskar et al., 2019)	Short instructional videos from different domains, each video is accompanied by a transcript.	80,000	1-2	Multi
EDUVSUM (Ghauri et al., 2020)	Educational videos with subtitles from three popular e-learning platforms: Edx,YouTube, and TIB AV-Portal.	98	10-60	Multi
COIN (Tang et al., 2019)	Contains 11,000 instructional videos covering 180 tasks. Used for training by creating pseudo summaries.	8,521	3	Uni
FineGym (Shao et al., 2020)	A hierarchical video dataset for fine-grained action understanding.	156	10	Multi
CrossTask (Xu et al., 2019)	Contains 4,700 instructional videos covering 83 tasks; used for generating pseudo summaries and training.	3,675	3	Uni
WikiHow Summaries (Hassan et al., 2018)	High-quality test set created by scraping WikiHow articles that include video demonstrations and visual depictions of steps.	2,106	2-7	Multi
HowTo100M (Miech et al., 2019)	A large-scale dataset with more than 100 million video clips from YouTube, annotated with textual descriptions.	136M	6-7	Multi
VTW (Zeng et al., 2016)	Large dataset containing annotated videos from YouTube with highlighted key shots	2,529	2-5	Uni

5 CONCLUSIONS

Video summarization is a crucial area of research, contributing to numerous valuable applications, such as video retrieval, personalized video recommendations, and surveillance systems. This comprehensive review has explored the most recent and effective techniques for video summarization, highlighting advancements in both extractive and abstractive methodologies. Learning-based approaches, including reinforcement learning, attention mechanisms, and multi-modal learning, have significantly improved the ability to generate concise and meaningful video summaries, enhancing their performance across various tasks.

Despite the progress made, several challenges remain. Current techniques often struggle with limitations in real-world scenarios, particularly in achieving real-time summarization and context-awareness. Additionally, there is a lack of robust multi-modal datasets that incorporate diverse features such as text, audio, and visual data, which are critical for further improving the adaptability of summarization models across different applications, from personalized video recommendations to security monitoring.

Looking forward, future research must focus on addressing these limitations by developing techniques that are more efficient and accurate, particularly in real-time and context-sensitive environments. Furthermore, the creation of more comprehensive, multimodal datasets will be essential to unlocking the full potential of video summarization technologies.

In summary, while significant progress has been made, this review has identified key gaps in current methodologies. Overcoming these challenges will be vital to fully harness the potential of video summarization as video content continues to proliferate across various domains.

REFERENCES

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., and Patras, I. (2020). Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3278–3292.

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., and Patras, I. (2021). Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

Banjar, A., Dawood, H., Javed, A., and Zeb, B. (2024). Sports video summarization using acoustic symmetric ternary codes and svm. *Applied Acoustics*, 216:109795.

Basavarajaiah, M. and Sharma, P. (2021). Gvsum: generic video summarization using deep visual features. *Multimedia Tools and Applications*, 80(9):14459–14476.

Benedetto, I., La Quatra, M., Cagliero, L., Canale, L., and Farinetti, L. (2023). Abstractive video lecture summarization: applications and future prospects. *Education and Information Technologies*, 29(3):2951–2971.

Elfeki, M., Wang, L., and Borji, A. (2022). Multi-stream dynamic video summarization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 339–349.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*.

Ghauri, J. A., Hakimov, S., and Ewerth, R. (2020). Classification of important segments in educational videos using multimodal features. *arXiv preprint arXiv:2010.13626*.

Gong, B., Chao, W.-L., Grauman, K., and Sha, F. (2014). Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27.

Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014). Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 505–520. Springer.

Hassan, S., Saleh, M., Kubba, M., et al. (2018). Wikihow: A large scale text summarization dataset. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3243–3252.

He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., and Wang, Z. (2023). Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878.

He, X., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Robertson, N., and Guan, H. (2019). Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the 27th ACM International Conference on multimedia*, pages 2296–2304.

Kumar, K., Shrimankar, D. D., and Singh, N. (2016). Equal partition based clustering approach for event summarization in videos. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 119–126.

Lee, Y. J., Ghosh, J., and Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353.

Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., and Shao, L. (2021). Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*, 111:107677.

Lin, J., Zhong, S.-h., and Fares, A. (2022). Deep hierarchical lstm networks with attention for video summarization. *Computers & Electrical Engineering*, 97:107618.

Liu, T., Meng, Q., Huang, J.-J., Vlontzos, A., Rueckert, D., and Kainz, B. (2022). Video summarization through reinforcement learning with a 3d spatio-temporal u-net. *IEEE Transactions on Image Processing*, 31.

Mahmoud, K. M., Ismail, M. A., and Ghanem, N. M. (2013). *VSCAN: An Enhanced Video Summarization Using Density-Based Spatial Clustering*, page 733–742. Springer Berlin Heidelberg.

Miech, A., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million video clips. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 2630–2640.

Narasimhan, M., Nagrani, A., Sun, C., Rubinstein, M., Darrell, T., Rohrbach, A., and Schmid, C. (2022). Tl; dw? summarizing instructional videos with task relevance and cross-modal saliency. In *European Conference on Computer Vision*, pages 540–557. Springer.

Narasimhan, M., Rohrbach, A., and Darrell, T. (2021). Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000.

Palaskar, S., Libovickỳ, J., Gella, S., and Metze, F. (2019). Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901*.

Parihar, A. S., Pal, J., and Sharma, I. (2021). Multiview video summarization using video partitioning and clustering. *Journal of Visual Communication and Image Representation*, 74:102991.

Potapov, D., Douze, M., Harchaoui, Z., and Schmid, C. (2014). *Category-Specific Video Summarization*, page 540–555. Springer International Publishing.

Pritch, Y., Rav-Acha, A., and Peleg, S. (2008). Nonchronological video synopsis and indexing. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1971–1984.

Rochan, M. and Wang, Y. (2019). Video summarization by learning from unpaired data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7902–7911.

Sabha, A. and Selwal, A. (2023). Data-driven enabled approaches for criteria-based video summarization: a comprehensive survey, taxonomy, and future directions. *Multimedia Tools and Applications*, 82(21):32635–32709.

Shao, D., Zhao, Y., Dai, B., and Lin, D. (2020). Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625.

Song, Y., Vallmitjana, J., Stent, A., and Jaimes, A. (2015). Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187.

Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tang, Z., Xiong, Y., Xu, Y., Wang, W., Hua, X.-S., and Zhang, J. (2019). Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216.

Tejero-de Pablos, A., Nakashima, Y., Sato, T., Yokoya, N., Linna, M., and Rahtu, E. (2018). Summarization of user-generated sports video by using deep action recognition features. *IEEE Transactions on Multimedia*, 20(8):2000–2011.

Tirupathamma, S. (2017). Key frame based video summarization using frame difference. *International Journal of Innovative Computer Science & Engineering*, 4(3).

Tiwari, V. and Bhatnagar, C. (2021). A survey of recent work on video summarization: approaches and techniques. *Multimedia Tools and Applications*, 80(18):27187–27221.

Wang, G., Wu, X., and Yan, J. (2024). Progressive reinforcement learning for video summarization. *Information Sciences*, 655:119888.

Wu, G., Lin, J., and Silva, C. T. (2022). Intentvizor: Towards generic query guided interactive video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10503–10512.

Xiao, S., Zhao, Z., Zhang, Z., Yan, X., and Yang, M. (2020). Convolutional hierarchical attention network for query-focused video summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07).

Xu, Z., Buch, S., Ramanan, D., and Niebles, J. C. (2019). Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545.

Ye, Q., Shen, X., Gao, Y., Wang, Z., Bi, Q., Li, P., and Yang, G. (2021). Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7950–7959.

Zeng, K.-H., Chen, T.-H., Niebles, J. C., and Sun, M. (2016). Title generation for user generated videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 609–625. Springer.

Zhao, B., Gong, M., and Li, X. (2021). Audiovisual video summarization. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):5181–5188.

Zhou, K., Qiao, Y., and Xiang, T. (2018). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhu, Y., Zhao, W., Hua, R., and Wu, X. (2023). Topic-aware video summarization using multimodal transformer. *Pattern Recognition*, 140:109578.