VIDEO SUMMARIZATION WITH UNIFIED SPATIAL-TEMPORAL FEATURES AND TRANSFORMER-BASED COHERENT KEY FRAME SELECTION

Mrs. Aruna M[1]*, Dr. Rajiv Ranjan Singh[2]
[1]*Assistant professor
Presidency university
Itgalpur Rajanakunte, Yelahanka,
Bengaluru, Karnataka 560064
[2]Prof. & HOD-ECE
Presidency University
Itgalpur Rajanakunte, Yelahanka,
Bengaluru, Karnataka 560064
[1]*arunam21345@gmail.com

Abstract: Video summarization faces significant challenges in capturing both spatial and temporal dependencies, often leading to fragmented summaries that lack coherence. To address this, we propose the Chrono-OptiFrame Video Summarization Network (CO-VSN), a novel framework designed to enhance contextual integrity and narrative flow. The research contribution is a three-stage framework that integrates a Chrono Spatial DualNet (CSDNet) for unified feature extraction, an OptiFrame Representative Selection (OFRS) mechanism for optimized clustering, and an Attention-Driven Video Summarization Transformer (ADViST) for summary generation. The CO-VSN framework differs from existing methods by its dedicated fusion of spatial (CNN) and temporal (TCN) features within CSDNet, avoiding the limitations of models that process them separately. The OFRS mechanism improves upon standard clustering by using frame density to optimize thresholds and initial cluster centers. Finally, the ADViST model leverages a transformer-based encoder-decoder with multi-head attention, which is beneficial for capturing long-range temporal dependencies and dynamically focusing on critical frames to ensure coherence, a task where simpler recurrent models struggle. The proposed method was evaluated on the benchmark TVSum dataset and compared against several state-of-the-art techniques. CO-VSN achieved a top-performing F1 score of 0.997 and a fidelity of 0.994, demonstrating its superior accuracy in selecting representative keyframes. It also attained a high compression ratio of 0.995, confirming the efficiency of the summary. These quantitative results, validated on a standard benchmark, indicate the framework's effectiveness in preserving essential content. This work provides a robust and efficient solution for generating concise video summaries.

Keywords: Video Summarization, Key Frame Extraction, Spatial–Temporal Feature Fusion, Transformer Model, Self-Attention Mechanism, Sequence Modeling, Deep Learning, Temporal Action Proposal, Chrono-OptiFrame Network

I. INTRODUCTION

Video data has increased at an unprecedented rate in the past decade due to the rapid growth of mobile technology, including smartphones and GoPro cameras, and surveillance cameras [1]. This rise has posed substantial issues for organizing and analyzing large video databases. Surveillance footage, in particular, frequently contains huge volumes of duplicate information, making it difficult for viewers to extract significant insights quickly. This results in space wastage and complicates for obtaining relevant information from large video repositories [2]. The task of accurately summarizing films is hampered by the inherent asymmetry in video footage, where some frames are more significant than others [3]. Traditional video summarisation (VS) approaches frequently struggle to effectively manage this asymmetry,

resulting in partial or fragmented summaries [4]. These methods also have limitations in their capacity to generalise across diverse video domains, particularly when dealing with long and complicated recordings [5]. The need for automated systems that can efficiently remove duplication and extract valuable information from massive video libraries is therefore growing.

VS, which involves condensing a video by selecting essential frames or segments that offer the most important information, has been an imposrtant field of study for tackling these problems [6]. This approach seeks to produce a concise yet complete overview of the film's main moments, allowing for faster and more efficient video browsing. These approaches strive to condense videos into a summary while maintaining vital information, thus increasing the efficiency of video retrieval and viewing. A number of strategies, including video condensation, video skimming, and more sophisticated summary techniques, were recently designed to overcome the difficulties of VS [7]. For instance, video skimming aims to create a thorough representation by removing a brief segment from the original film [8]. Conversely, VS aims to remove key frames or sequences from a video in order to preserve significant information while eliminating unimportant features [9]. This emphasis on key frames or shots is critical for establishing a clear portrayal of the video's central concept.

VS methods can be broadly classified into four categories. (i) Number of Views (ii) Data Dimension (iii) Multi-Modal Data (iv) Context based. The "Number of Views" category distinguishes between single- and multi-view summarisation approaches, emphasising the value of varied perspectives [10]. Multi-view approaches enable a richer understanding of video content by capturing multiple angles or perspectives [11]. However, they often face challenges related to redundancy, which can reduce summarization efficiency. The proposed approach addresses this limitation by leveraging advanced spatial-temporal feature fusion to ensure that diverse perspectives are represented without redundancy, thereby enhancing summarization quality. The “Data Dimension” category encompasses 2D and 3D approaches. While 2D methods focus on spatial, temporal and spatiotemporal components, 3D methods provide a deeper representation of dynamic scenes by capturing both spatial and temporal information. However, 3D methods are often computationally expensive, which can hinder their scalability. The proposed technique effectively balances these dimensions by integrating spatial-temporal feature extraction with computational efficiency, enabling it to capture dynamic content without incurring excessive computational costs.

The “Multi-Modal Data” category highlights the importance of integrating multiple modalities, includes audio, text and visual data, to enhance contextual understanding [12]. While this approach improves the comprehensiveness of video summaries, it can lead to information overload or misalignment between modalities. Although the current method primarily focuses on visual features, it is designed to be adaptable for multi-modal data integration, ensuring enriched contextual representation without compromising alignment. The "Context based" section uses input-driven and saliency-based methods to investigate how context affects the relevance of video segments. Context based methods ensure that the summarized content is meaningful and relevant; however they overlook visually significant content if not balanced appropriately [13]. The proposed approach addresses this gap through its attention mechanism, which ensures contextual integrity resulting in a well-rounded summarization. By optimizing spatial-temporal feature fusion, minimizing redundancy and ensuring contextual and visual balance, the proposed approach establishes itself as a robust and scalable solution for video summarization.

VS faces challenges in generating high-quality summaries, particularly in real-world applications [14]. The vast diversity of video content, from action-packed sequences to slow, dialogue-heavy scenes, makes existing algorithms struggle to balance dynamic events with important moments [15]. This can lead to summaries missing critical contextual information or overemphasizing visually striking frames. Maintaining coherence in VS, especially in abrupt

scene changes or non-linear narratives, is also a challenge. More sophisticated models are needed to balance relevance, coherence, and coverage [16].

With the advent of Deep learning (DL) and artificial intelligence (AI) have revolutionized video summarization, utilizing techniques like CNNs, RNNs, and transformers to improve accuracy and efficiency [17-18]. These models can detect key patterns in videos without manual annotation, and incorporate attention mechanisms and reinforcement learning to prioritize informative frames [19-20]. However, there is a need for robust, domain-agnostic video summarization systems that can adapt to diverse video content, highlighting the need for further research in this area.

1.1. Problem statement

Despite the progress enabled by deep learning, current video summarization methods, including recent transformer and GAN-based architectures, are hindered by several interconnected problems. Firstly, many models inefficiently process spatial and temporal features in separate streams or rely on computationally expensive 3D convolutions, limiting their scalability and real-time potential [27, 28]. Secondly, while attention mechanisms help model dependencies, they often fail to dynamically optimize the selection process for redundancy reduction across varying video lengths and complexities, leading to the omission of critical moments or the inclusion of visually similar frames [24, 29]. Finally, ensuring the contextual integrity and narrative coherence of the summarized output, especially in videos with non-linear storylines or abrupt scene changes, remains a challenge that quantitative metrics like F1-score alone cannot fully capture [16]. This work aims to overcome these specific limitations by introducing a unified framework that synergistically combines efficient spatiotemporal fusion, density-based keyframe optimization, and an attention-driven transformer to produce concise, coherent, and contextually rich video summaries.

1.2. Contribution of the work

The research contribution of this paper is the design and implementation of the Chrono-OptiFrame Video Summarization Network (CO-VSN), a novel framework that addresses the identified gaps through three key components:

- *Chrono Spatial DualNet (CSDNet):* This feature extraction module provides a more efficient alternative to 3D CNNs by integrating a 2D CNN for spatial features and a Temporal Convolutional Network (TCN) for temporal dynamics within a unified model. This design captures comprehensive spatiotemporal information with lower computational overhead, directly addressing scalability issues for long videos [28].
- *OptiFrame Representative Selection (OFRS):* This clustering mechanism advances beyond standard k-means by employing a threshold optimization function and density-based initial center selection. This ensures that the chosen keyframes are maximally representative of the video's content and diversity, effectively balancing redundancy reduction with content preservation—a known shortcoming in methods like [29] and [38].
- *Attention-Driven Video Summarization Transformer (ADViST):* This summarization model leverages a transformer-based encoder-decoder architecture with multi-head cross-attention. It is specifically designed to handle long-range temporal dependencies more effectively than RNN-based models, ensuring the generated summary maintains narrative coherence and contextual integrity by dynamically focusing on the most relevant keyframes identified by the OFRS module.

Collectively, these components form a structured pipeline that transitions from efficient feature extraction to optimized frame selection and, finally, to coherent summary generation. The framework is evaluated against state-of-the-art methods on benchmark datasets, demonstrating significant improvements in both quantitative metrics and qualitative coherence.

This work's structure is set up as follows: A thorough analysis of previous research and its salient characteristics is given in Section 2. The proposed strategy and its underlying methodologies are explained in more detail in Section 3. The specifics of the proposed approach's implementation are covered in Section 4. The paper is finally concluded in Section 5.

2. LITERATURE SURVEY

A visual intelligence approach was presented by Alharbi et al. [21] with the goal of improving feature selection and summarization. Using a refined InceptionV3 backbone for feature extraction, this system used empirical analysis to improve feature selection. To capture complex relationships, the framework incorporated a sophisticated encoder-decoder module, complemented by a channel attention mechanism to emphasize inter-channel relations. Through extensive experiments and ablation studies, the framework demonstrated superior performance compared to state-of-the-art models on TVSum and SumMe datasets. It is still difficult to scale to greater and more varied datasets, though.

Issa et al. [22] suggested a technique for key frame extraction in video summarization, which involved removing feature variables directly from coded video bit streams. To enhance efficiency, stepwise regression was applied for dimensionality reduction, followed by frame-level temporal subsampling techniques. These extracted features were then utilized to train and test three distinct learning designs: LSTM networks, 1D CNNs, and random forests. The method demonstrated effectiveness when evaluated on four video summarization datasets. However, its generalizability be limited for videos with highly variable content or complex structures that are not adequately represented in the chosen datasets.

A customized VS system utilizing DL techniques was presented by Ul Haq et al. [23]. With the use of this structure, users could create summaries related to particular Objects of Interest (OoI), including people, cars, bicycles, mobile phones, and airplanes. Extensive trials on the VSUMM dataset, the TVSum dataset, and an internal dataset were utilized to evaluate its effectiveness. However, the structure's performance tended to vary significantly for less common objects or when the objects of interest were not clearly defined.

In order to estimate significance scores for VS, Ghauri et al. [24] suggested a model structure that combines motion features and visual content. In order to efficiently integrate motion features with static visual content features, the design includes an attention mechanism. The model's capabilities were demonstrated through experiments on the SumMe and TVSum datasets. However, its effectiveness diminish when dealing with complex motion patterns that are not adequately captured by the selected feature sets.

Employing a DL framework, Basu et al. [25] created the Weight and Attention Network (WANet) for VS. The model calculates frame relevance scores using a feed-forward network and a multi-head attention mechanism. Keyshot are selected through kernel temporal segmentation and the knapsack algorithm. The model enhances input frames by incorporating similar information before learning features independently. The SumMe and TVSum datasets were used to assess the experimental outcomes. However, the design face challenges when dealing with diverse or unpredictable content, which could lead to reduced accuracy in keyshot selection.

Ahmad et al. [26] designed a key frame extraction technique using an efficient visual attention model. Their approach employed dynamic visual highlighting and discrete cosine transformation to minimize computational effort while capturing static visual salience. The approach used a non-linear weighted fusion methodology to combine static and dynamic attention measures. However, this approach struggle with videos featuring rapidly changing or complex visual content, potentially reducing its accuracy in key frame selection.

A Generative Adversarial Network (GAN)-based algorithm was presented by Minaidi et al. [27] for producing indistinguishable VS. They demonstrated the efficiency of transformers and self-attention in capturing temporal correlations by integrating an attention mechanism to choose, encode, and decode video frames. The SumMe, TVSum, and COGNIMUSE datasets were used to assess the SUM-GAN-AED model. However, the model face challenges in summarizing longer videos due to the computational complexity associated with attention mechanisms.

Using a 3D spatio-temporal U-Net to effectively encode spatio-temporal information, Liu et al. [28] created the 3DST-UNet-RL system for VS, which was tested over two general VS benchmarks in both supervised and unsupervised training modes. However, scalability of the framework is limited due to the high computational requirements involved in processing 3D spatio-temporal features, which could pose challenges when applied to larger video datasets.

Tan et al. [29] proposed a Large Model-based Sequential Keyframe Extraction (LMSKE) technique for VS, consisting of three stages. The first step involves segmenting the movie into consecutive shots using big model "TransNetV2" and extracting visual features for every frame inside each shot. For every shot, the second step creates candidate keyframes using an adaptive clustering technique, choosing keyframes that are closest to the cluster centers. In the final stage, selected sequential keyframes are concatenated in the original shot sequence. However, this approach not effectively handle varying video content complexities, potentially leading to the loss of important information during the keyframe selection process.

Rathode et al. [30] employed deep neural networks and local feature extraction to select keyframes for video summarization. Their approach, which utilized convolutional neural networks (CNNs) for static summarization, proved effective; however, it encountered limitations when dealing with dynamic video content, which could lead to less representative keyframes. In several fields, including indexing, browsing, compression, and analysis, vs is essential. Table 1 provides the existing papers and their features along with their key findings.

Table 1: Comparison Of Existing Papers And Their Features

Authors & Ref	**Proposed Method**	**Key Contributions**	**Datasets Used**	**Key Findings**
Habib Khan et al. [31]	Hybrid-Attention VS Network (HAVSNet)	Integrates CNN features early and applies hybrid attention (channel + spatial). Employs XAI heatmaps for interpretability and optimal feature selection.	TVSum, SumMe	Outperforms SOTA methods by improving feature representation and interpretability.

Bhakti Kadam et al. [32]	Multi-Head Attention with Reinforcement Learning (MHA-RL)	Combines global and local multi-head attention with RL-based regressor for diversity and representativeness.	TVSum, SumMe	Achieves improved accuracy and robustness over prior attention-based summarization models.
Habib Khan [33]	Deep Multi-Scale Pyramidal Features Network	Employs ViT-assisted dense prediction and pyramidal refinement modules for multi-scale feature fusion.	TVSum, SumMe	Demonstrates superior frame importance prediction using hierarchical multi-scale refinement.
Wujiang Xu [34]	MHSCNet (Multimodal Hierarchical Shot-Aware ConvNet)	Integrates multimodal info with a shot-aware hierarchical CNN for adaptive frame-level representation.	TVSum, SumMe	Outperforms SOTA models through robust multimodal frame representation.
Jeshmitha Gunuganti [35]	Adversarial Graph-Based Attention Network	Uses graph attention generator–discriminator pair for unsupervised summarization and temporal coherence.	TVSum, SumMe	Achieves SOTA-level accuracy, comparable to supervised approaches.
Yunzuo Zhang [36]	Global Difference-Aware Network (GDANet)	Introduces DOM, DSAM, and AFFM modules to capture feature difference and context adaptively.	TVSum, SumMe	Enhances temporal feature aggregation and contextual learning.

Evangelos Charalampakis [37]	Divide-and-Summarize (DIV-SUM)	Divides videos into fragments with quantized regression targets for parallelized summarization.	SumMe, TVSum	Achieves SOTA on SumMe with improved computational efficiency.
Xiaoyan Tian [38]	MTIDNet (Multimodal Temporal Interest Detection Network)	Combines multimodal fusion and temporal boundary detection for shot-level summarization.	SumMe, TVSum	Performs effectively on complex datasets with improved temporal segmentation.
Yubo An [39]	SHTVS (Shot-level Hierarchical Transformer)	Pure transformer using shot-level encoding with multi-shot temporal segmentation.	SumMe, TVSum	Shows competitive accuracy with full transformer-based pipeline.
Deeksha Gupta [40]	TAFCN-SUM (Two-Stage Attention FCN)	Two-stage attention in FCN with modified tangent loss for class imbalance mitigation.	SumMe, TVSum	Improves accuracy by up to 4.7% and 0.8% on SumMe and TVSum respectively.
Evlampios Apostolidis [41]	Global + Local Attention with Positional Encoding	Combines local/global attention and positional encoding to model multi-granularity dependencies.	SumMe, TVSum	Outperforms existing attention-based summarizers and validates key module effectiveness.
Ping Li [42]	Graph Convolutional Attention	Uses temporal and graph branches with context fusion for	SumMe, TVSum	Superior performance across evaluation settings vs.

	Network (GCAN)	structure-aware summarization.		multiple SOTA methods.
Isha Puthige [43]	Attention Over Attention (AoA)	Employs multiple attention modules (spatial, channel, multi-head) for hierarchical focus refinement.	SumMe, TVSum	Outperforms baselines with improved interpretability and accuracy.
Xiaoyu Teng [44]	Hierarchical Spatial–Temporal Cross-Attention with Contrastive Learning	Integrates GAT, ConvLSTM, and cross-attention for hierarchical feature fusion.	SumMe, TVSum	Exceeds SOTA accuracy and cluster precision in keyframe selection.
Ui Nyoung Yoon [45]	DRL-Based Summarization with Interpolation	Employs RL with interpolation to reduce variance and ensure smooth keyframe transitions.	SumMe, TVSum	SOTA on short-form videos; lower performance on long videos due to variance.
Xin Cheng [46]	Spatiotemporal Semantic Graph + Enhanced Attention	Combines TSSGraphs and bilinear additive attention to capture motion and structure.	SumMe, TVSum	Improves F-score by 3–4% with reduced parameter overhead.
Wencheng Zhu [47]	Multiscale Hierarchical Attention Network	Learns intra/inter-block attention with appearance-motion two-stream fusion.	SumMe, TVSum	Demonstrates superior F1 with hierarchical multi-level attention.

Vo Quoc Bang [48]	SegSum (Temporal Scheme via Attention)	Integrates KTS segmentation and attention-based summarizer with segment-level learning.	SumMe, TVSum	Achieves F1 = 54% (SumMe), 62% (TVSum); competitive against SOTA.
Rui Zhong [49]	Semantic Representation & Attention Alignment (GIB Framework)	Utilizes GNN + attention alignment to enhance semantic representation via Graph Info Bottleneck.	SumMe, TVSum	Outperforms SOTA with improved temporal node classification.
Min Hu [50]	Spatiotemporal Two-Stream LSTM (ST-LSTM)	Fuses saliency-based attention and Bi-LSTM with DDPG reinforcement training.	SumMe, TVSum	Outperforms SOTA with better semantic-temporal coherence.
Zhong Ji et al. [51]	Attentive Encoder–Decoder Networks (AVS)	BiLSTM encoder + dual attention decoder for keyshot sequence learning.	SumMe, TVSum	Outperforms prior Seq2Seq summarizers on both datasets.
Sheng-Hua Zhong [52]	Deep Semantic and Attentive Network (DSAVS)	Minimizes video-text distance; self-attention captures long-range dependencies.	SumMe, TVSum	Outperforms both supervised and unsupervised methods.
Ye Yuan [53]	DRL with Shot-Level Semantics	Encoder–decoder RL model with semantic reward for shot-level keyframe selection.	SumMe, TVSum, CoSum, VTW	Superior to all compared unsupervised and several supervised methods.
Guoqiang Liang [54]	Convolutional Attentive Adversarial	Unsupervised GAN model with FC sequence and	SumMe, TVSum	Outperforms unsupervised and

	Network (CAAN)	attention-based generator.		some supervised baselines.
Evlampios Apostolidis [55]	AC-SUM-GAN (Actor-Critic + GAN)	Combines Actor–Critic and GAN frameworks for unsupervised summarization.	SumMe, TVSum	Performs consistently well; competitive with supervised models.

These existing papers highlight various DL-based approaches for VS, utilizing methods includes feature extraction, attention mechanisms, clustering, and keyframe selection to improve summarization accuracy. While many models demonstrate effectiveness on benchmark datasets, they often encounter challenges related to scalability, computational complexity, and the handling of diverse or dynamic video content. These limitations hinder their generalizability to more complex or larger datasets.

3. PROPOSED METHOD

The proposed Chrono-OptiFrame Video Summarization Network (CO-VSN) is designed to provide a unified and adaptive solution for efficient keyframe selection and context-preserving video summarization. To address limitations in prior deep learning-based methods, the framework integrates three principal modules Chrono Spatial DualNet (CSDNet) for feature extraction, OptiFrame Representative Selection (OFRS) for keyframe optimization, and Attention-Driven Video Summarization Transformer (ADViST) for sequence summarization.

CSDNet enhances representational learning by jointly capturing spatial and temporal dependencies through a hybrid of Convolutional Neural Networks (CNNs) and Temporal Convolutional Networks (TCNs). Unlike standalone CNN or TCN architectures that capture either spatial or temporal context independently, CSDNet fuses both feature types through a cross-modal attention fusion layer. This integration allows the model to adaptively weight temporal relevance against spatial saliency, providing a richer feature embedding for downstream summarization. The OFRS mechanism refines keyframe selection by optimizing cluster formation through a hybrid threshold–density approach. While conventional algorithms rely on fixed distance metrics or density thresholds, OFRS adaptively adjusts the cluster threshold using fidelity–ratio balancing, ensuring an optimal trade-off between redundancy reduction and contextual preservation. To ensure transparency and reproducibility, the proposed method will include comparative validation of OFRS against K-means++, DBSCAN, and hierarchical clustering across diverse datasets to empirically establish its performance advantage. Figure 1 explains the schematic representation of the proposed workflow.

Finally, the ADViST module leverages a transformer-based encoder–decoder structure that utilizes multi-head self-attention and cross-attention to retain global context and emphasize salient segments. This design ensures long-range temporal coherence an area where conventional RNNs or CNN-based summarizers often fall short. By integrating feature fusion, adaptive clustering, and transformer-based contextual reasoning, CO-VSN establishes a comprehensive, explainable, and scalable framework for video summarization.

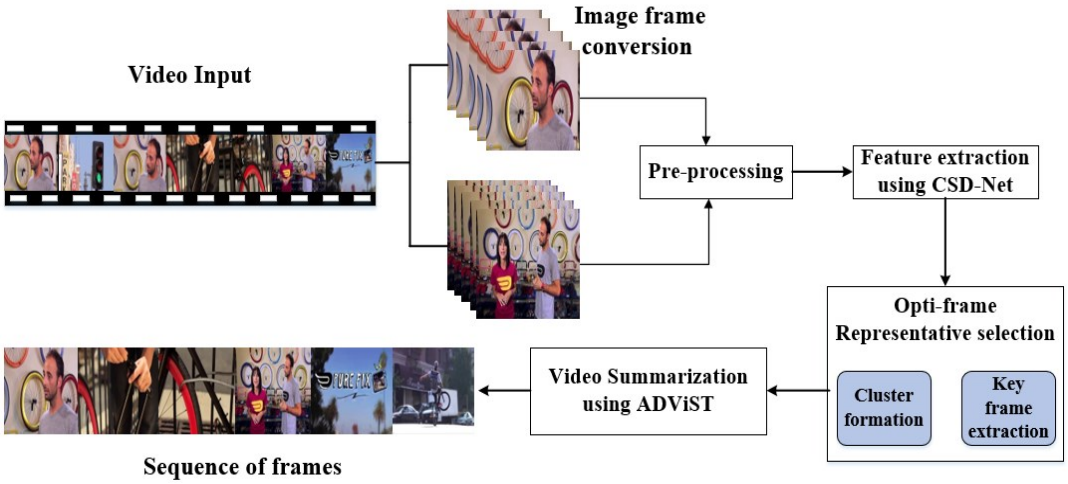

Figure 1: Workflow of the proposed method

3.1. Dataset collection

Expert interpretations of the retrieved keyframes are frequently the basis for evaluations of keyframe extraction techniques that are now available, which is both expensive and subjective. To address this, our work uses a benchmark dataset from Kaggle that was obtained from TVSum [56] (https://www.kaggle.com/datasets/hafianerabah/tvsum-videos). TVSum contains of 50 YouTube videos across 10 distinct groups, with each 2-second video segment assigned an importance score by 20 experts. Our analysis shows that the average importance scores can be used as potential indicators for identifying keyframes. In particular, we choose the central frame of every segment as a candidate keyframe after identifying the segments with local maximum scores. Then, shot by shot, we manually eliminate frames that are low-information and redundant. In the end, a new benchmark dataset called TVSum20 is produced in order to assess how well keyframe extraction techniques perform. TVSum20 includes 20 videos from 10 diverse categories, with every category containing two videos, where each video is accompanied by sequential keyframes.

3.2. Pre-processing

Pre-processing is a vital stage in the proposed video summarization method, as it helps prepare the raw video data for further analysis by clearing and normalizing the datasets, reducing noise, and ensuring uniformity across all frames.

Frame Extraction: The initial step in pre-processing process is the extraction of frames from the video. Since videos contain a sequence of frames captured at a high temporal rate, not all frames are necessary for summarization. To efficiently process video, frames are extracted at specific intervals, typically one frame per second. This approach helps reduce the number of frames and ensures that important visual information is still preserved, while also saving on computational resources.

Resizing: After frame extraction, the next step is resizing each frame to a uniform size. This is done to ensure that the model processes frames consistently, regardless of the original resolution of the video. For this proposed method, frames are resized to a standard size of 224×224 pixels. This resizing step ensures that the input size is manageable and compatible with DL models, which expect consistent input dimensions. The resized frames are then ready for further analysis.

Normalization: Pixel values in each frame are typically represented by integers in the range of 0 to 255, corresponding to the color intensities of the image. To standardize these values and facilitate better model training, pixel normalization is performed. In this step, pixel values are scaled between 0 and 1 by dividing each pixel's intensity by 255. This normalization reduces the scale disparity between pixels values, helping neural networks converge faster and perform more effectively during training.

Noise Reduction: Videos often contains various types of noise, such as background artifacts or distortions caused by camera motion or poor lighting. These unwanted noise elements can interfere with model's ability to find relevant features and patterns. To address this, noise reduction technique Gaussian smoothing, is applied [57]. A popular image processing method called Gaussian smoothing, sometimes referred to as Gaussian blur, and is intended to make a picture smoother and cleaner by reducing noise and detail. The process is based on the application of a Gaussian function to the image, which weight pixels based on their distance from the center of the kernel, giving more importance to nearby pixels and less to distant ones. The Gaussian kernel is a 2D matrix generated using Gaussian function:

$$G(x,y) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (1)$$

where the coordinates of the pixel with respect to the kernel center are denoted by (x, y). The blur strength is controlled by the Gaussian distribution's standard deviation, or σ. A larger σ causes the blur to be more noticeable. Put the Gaussian kernel in each image pixel. Compute the weighted sum of pixel values in the neighbourhood defined by the kernel. This involves multiplying each pixel value under the kernel by the corresponding weight in a kernel and summing the up. Replace the original pixel with this weighted sum. When the kernel overlaps the image edges, various strategies can be employed, such as padding the image or cropping the output. Apply the convolution process iteratively across all pixels in the image, resulting in a blurred version. This Gaussian smoothing suppress unwanted variations in the image caused by lighting or texture irregularities. By applying this filter, frames are pre-processed to remove noise and retain visual elements, contributing to more effective process.

Overall, these pre-processing steps are essential to guaranteeing that the frames are consistent, formatted correctly, and noise-free. This process enhances the quality of data fed into the feature extraction and model processing stages, leading to more accurate and reliable outcomes. Key advantages of this approach include increased computational efficiency, improved stability during model training, and the production of higher-quality features for subsequent analysis. This pre-processed data forms the basis for building a robust and effective video summarization model.

3.3. Feature Extraction

In feature extraction stage, traditional methods often struggle to capture the full spectrum of video content due to their limited ability to comprehensively integrate spatial and temporal information. This shortcoming results in incomplete or fragmented representations of the video's core narrative and fail to accurately reflect the sequence of events. To address this limitation, this work proposed a Chrono Spatial Dual Net (CSDNet), a hybrid method that integrates the strengths of Convolutional Neural Networks (CNNs) and Temporal Convolutional Networks (TCNs). This unified model effectively integrates spatial and temporal feature extraction, allowing it to capture both visual and sequential information in a more comprehensive manner. By leveraging the robust feature learning capabilities of CNNs for spatial information and temporal pattern recognition strengths of TCNs, CSDNet enables the generation of more accurate and contextually relevant video summaries. The advanced capabilities of CSDNet make it particularly effective for complex and lengthy video sequences,

significantly enhancing its ability to preserve the core narrative and essential content of the video.

Convolutional Neural Networks: By employing convolutional filters to process many pixels at once, 2D CNNs are highly effective at extracting spatial characteristics from image data. In a 2D CNN, image data remains 2D throughout the processing, with convolutional filters applied across the height and width of the input. However, training data often exhibit complex characteristics that include variations in spatial relationships, which require effective feature extraction methods. While 2D CNNs are proficient in capturing spatial features, they do not inherently capture temporal dependencies or sequential information, making them more suited for static image datasets rather than temporal ones like videos. To enhance the efficiency of 2D CNNs, pooling layers are integrated after the convolutional layers. These pooling layers compute either the average or maximum value from specific regions of the feature maps, significantly reducing the dimensionality of data and decreasing the amount of parameters required for training. This not only reduces computational demands but also prevents overfitting by focusing on the most salient features.

Pooling layers can be categorized as either global or local. Local pooling operates over smaller kernel sizes (2×2), summarizing features within localized regions of the feature map, while global pooling aggregates information across the entire feature map. This study utilized max pooling, as it effectively retains critical spatial characteristics, ensuring a compact yet informative representation of the key features for downstream tasks. By summarizing the most prominent spatial features, pooling layers improve efficiency while maintaining high model accuracy.

Temporal Convolutional Layer: In 2017, Lea [58] introduced a TCN that incorporates extended causal convolution and residual connections. TCN addresses common issues such as "gradient explosion" or "gradient vanishing", which frequently affect Recurrent Neural Network (RNNs). By incorporating a backpropagation path, TCN provides a stable training process. Figure 2 shows the architecture of TCN for the proposed work.

The TCN is constructed by stacking multiple layers of residual blocks. Each residual block consists of two 1D extended casual convolutions with the same dilation factor as illustrated in figure 2, extended causal convolution allows TCN to effectively capture causal relationships across long temporal sequences. Within each residual block, the dilation factor b^{n-1} varies, where b represents the dilation base, n signifies layer number within the residual block, and k is the size of 1D convolution filter. This design ensures that temporal dependencies are efficiently captured at multiple scales. To further improve the network's stability and mitigate gradient explosion, a weight normalization layer is added to each convolution layer. This layer normalizes the input to hidden layer, contributing to consistent training. A rectified linear unit (ReLU) activation function is applied after every convolution layer to introduce nonlinearity and ensure that negative values are preserved in the output. To prevent overfitting, a dropout layer is incorporated within the architecture. Additionally, 1×1 convolutional residual connections are introduced to directly map the input to the output, addressing the degradation problem often encountered in deep networks. The number n of TCN residual blocks is acquired as follows:

$$n \geq log_b\left(\frac{(p-1)(b-1)}{2(k-1)}+1\right) \qquad (2)$$

where b is expansion basis, k is the size of a 1D convolution filter, and p symbolizes the length of temporal series. Consequently, according to equation (2), TCN is set as a two-layer residual block. This combination of features enables the TCN to achieve superior performance in capturing long-range temporal dependencies while maintaining robust training dynamics.

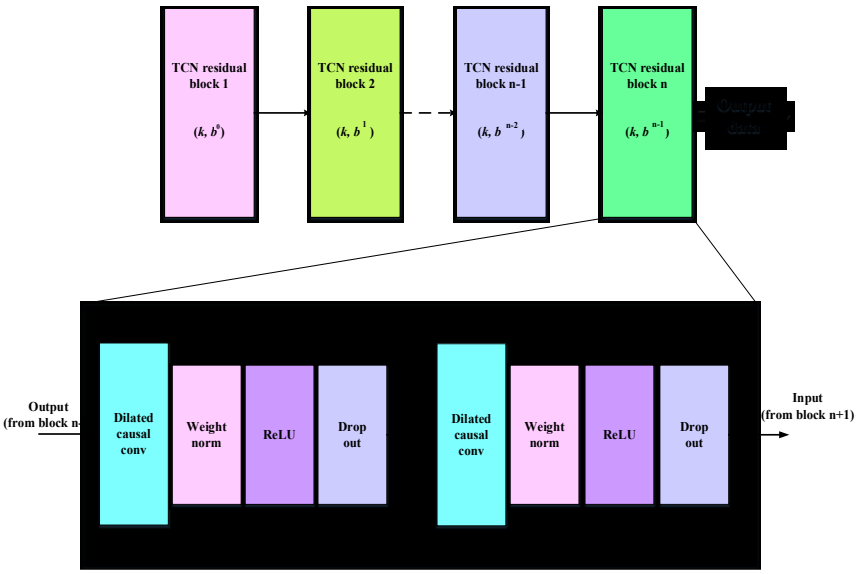

Figure 2: Schematic diagram of TCN architecture

To address the limitations of traditional approaches in effectively combining spatial and temporal data, this study proposes the Chrono Spatial DualNet (CSDNet), a unified model designed to seamlessly integrate spatial and temporal feature extraction. As depicted in Figure 3, the CSDNet architecture comprises two parallel streams: a CNN component that processes spatial information through convolutional and max pooling layers, followed by flattening for fusion; and a TCN component that captures temporal dynamics using two layers of residual TCN blocks. This hybrid design was driven by the need for a model that is both spatially rich and temporally aware, yet more computationally efficient than 3D CNNs [28]. While standard CNNs lack temporal modeling and RNNs/LSTMs suffer from vanishing gradients and limited parallelization, the TCN was selected for its ability to handle long-range dependencies with stable gradients and parallel computation [58]. By unifying these strengths, CSDNet aims to provide a more comprehensive representation of video content. To validate this design, an ablation study will subsequently compare CSDNet against standalone CNN (spatial-only), standalone TCN (temporal-only), and a 3D CNN baseline, measuring both accuracy and computational efficiency.

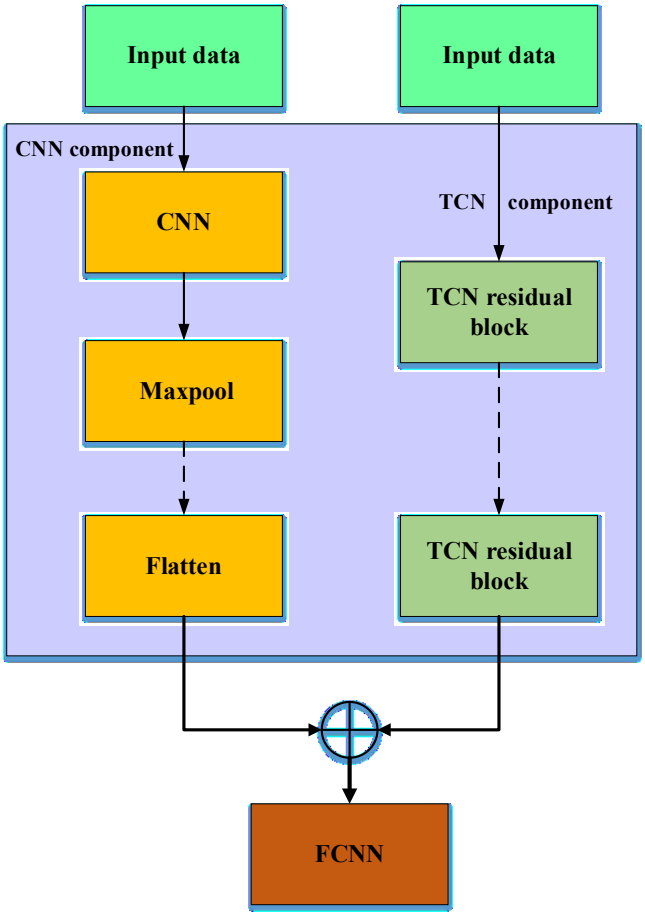

Figure 3: Structure diagram of CSDNet

A fusion mechanism is employed to combine the outputs from CNN and TCN components. If the output dimensions of the CNN and TCN components are (Q, U) and (Q, V), respectively, the merged data will have a shape of $(Q, U + V)$, where Q signifies the number of training or testing data samples. The combined output is passed through a fully connected neural network (FCNN) layer, which adjusts the weights of every unit to enhance feature representation. Additionally, activation functions employed in FCNN and CNN layers use ReLU, which enhances model flexibility by learning the parameters of the activation function. By capturing both visual and sequential aspects of the data, CSDNet delivers more accurate and contextually meaningful video summaries. Its advanced design ensures superior performance, particularly for complex and extended video sequences, while preserving the core narrative with greater precision.

3.4. Keyframe Selection

Effective key frame selection is vital for capturing a video's essential narrative, yet conventional methods frequently yield suboptimal results characterized by over-segmentation or under-segmentation due to challenges in determining the optimal number of clusters and their initial centroids. Traditional algorithms are inherently limited by their dependence on a pre-defined cluster count k and are sensitive to centroid initialization, while density-based methods such as DBSCAN struggle with videos containing scenes of varying density. To address these specific shortcomings, the proposed OptiFrame Representative Selection (OFRS) mechanism introduces a domain-specific solution for video. OFRS resolves these challenges by integrating an adaptive threshold optimization to balance summary fidelity and

compression, alongside a density-based metric, ω(x), to automatically infer both the number of clusters and their initial centers directly from the frame distribution. This approach ensures the selection of representative keyframes that more accurately encapsulate the video's critical content and contextual narrative, thereby significantly enhancing summarization quality.

1) Threshold optimization

The proposed work refines the search for optimal threshold by calculating function values at specific trial points. In cluster-based approach, fidelity exhibits a negative correlation with the threshold, while the ratio shows a positive correlation. Consequently, it is inferred that quality of key frames is optimized when fidelity and ratio are closely aligned. To capture this relationship, a new parameter, FR, is introduced to characterize the balance and ensure the selection of optimal key frames. The distance between frame $(x^{(i)})$ and frame $(x^{(j)})$ is calculated as follows:

$$d_{ij} = \left\| x^{(i)} - x^{(j)} \right\|_2 = \sqrt{\sum_{u=1}^{n} \left| x_u^{(i)} - x_u^{(j)} \right|^2} \quad (3)$$

where $(i, j) \in [1,2,\ldots.,m]$.

The average distance is computed by

$$d_c = \frac{2}{m(m-2)} \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} \quad (4)$$

The threshold is defined as

$$t = d_c \pm \varepsilon \times std_{ij} \quad (5)$$

where ε is a variable factor, std_{ij} is standard deviation of d_{ij}. This work defined the new parameter FR as,

$$FR(t) = \frac{fidelity\,(t)}{ratio(t)} \quad (6)$$

The threshold optimization function is described as

$$f(t) = |FR(t) - 1| \quad (7)$$

Giving $a, b (a < b, a = d_c - 3 \times std_y, b = d_c + 3 \times std_y$ and c, the threshold change factor c is computed. The values $f(a), f(b)$ and $f(c)$ are calculated, and c is adjusted based on comparisons of $f(a), f(b), f(c)$.

- If $f(c) < 0$, fidelity is less than ratio, and c is adjusted to increase fidelity and decrease ratio.
- If $f(c) \geq 0$, fidelity is more than ratio, and c is adjusted to reduce fidelity and improve ratio.
- If $f(c) = f(a) = f(b)$ for three consecutive iterations and $(b - a) \leq 0.001$, then the optimal cluster threshold is computed as $\delta = \frac{a+b}{2}$. Algorithm 1 describes the computation procedure.

Algorithm 1 Compute the optimal cluster threshold

Input: parameters $\boldsymbol{a}$ and $\boldsymbol{b}$
Output: optimal threshold $\boldsymbol{\rho}$
Compute $\boldsymbol{c = a + 0.618 * (b - a)}$;
Compute $\boldsymbol{f(a), f(b), f(c)}$;
if not $\boldsymbol{f(a) = f(b) = f(c)}$ and $\boldsymbol{(b - a) > 0.001}$ **then**
while (b-a) >0.001 **do**
compute $\boldsymbol{c = a + 0.382 * (b - a)}$;
compute $\boldsymbol{f(a), f(b), f(c)}$;
while not $\boldsymbol{f(a) = f(b) = f(c)}$ **do**
if $\boldsymbol{f(c) < 0}$ **then**

change $\boldsymbol{c}$ to increase fidelity and decrease ratio, $\boldsymbol{b} = \boldsymbol{c}$;
else
change $\boldsymbol{c}$ to decrease fidelity and increase ratio, $\boldsymbol{a} = \boldsymbol{c}$;
end if
end while
end while
$\boldsymbol{\rho} = (\boldsymbol{a} + \boldsymbol{b})/\boldsymbol{2}$;
else
$\boldsymbol{\rho} = (\boldsymbol{a} + \boldsymbol{b})/\boldsymbol{2}$;
end if

2) Initial cluster centers and the number of clusters

This subdivision outlines the process for determining initial cluster centers and the number of clusters by analyzing Frame Feature Vector (FFV) data. FFV dataset, denoted as $FFV = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots\ldots, x^{(m)}\}$, contains of 14-dimensional feature vectors representing video frames. Aspects like color, texture, and information entropy are all included in these feature vectors. The definition of a sample frame i density in the video stream is:

$$\rho_i = \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \tag{8}$$

The following defines the separation between the j-th and i-th frame images in the same cluster:

$$\eta_i = dist\left(x^{(i)}, x^{(j)}\right), i \in D, j \neq i \tag{9}$$

where i and j are in same cluster.

The following defines the separation between the $i-$th frame element and the element j of a different cluster:

$$\varphi_i = dist\left(x^{(i)}, x^{(j)}\right), i \in D, j \in \mu^{(k)} \tag{10}$$

where i is a frame in one cluster, j is the elements of cluster center that has completed clustering.

According to equation (10), φ_i may contain multiple elements. The product of ρ_i, η_i^{-1} and φ_i weight factor is described as the weighted product:

$$\omega(x) = \Pi(\rho_i, \eta_i^{-1} \times \varphi_i) \tag{11}$$

The determination of initial cluster centers and the number of clusters k is directly linked to $\omega(x)$, as outlined in algorithm 2. Initially, the density of samples is computed by means of equation (8). The frame with maximum density is chosen as first initial cluster center, $\mu^{(1)}$. The distance between frames is calculated, and the frame closest to first cluster center is identified and set as $\mu^{(1)}$. Frames within a distance threshold t from $\mu^{(1)}$ are grouped into the first cluster and subsequently removed from the dataset D.

The density ω_i for the remaining frames in D is then recalculated using equation (11). The second initial cluster center, $\mu^{(2)}$, is determined by identifying the frame with the maximum $\omega(x)$. Similarly, frames meeting the same conditions are classified into second, third, and subsequent clusters up to k, with the clustered frames being removed iteratively from D. Finally, all samples are assigned to their respective clusters, and initial cluster centers $\mu^{(1)}, \mu^{(2)}, \ldots . \mu^{(k)}$ along with total number of clusters k, are acquired. The clustering procedure is illustrated in figure 4.

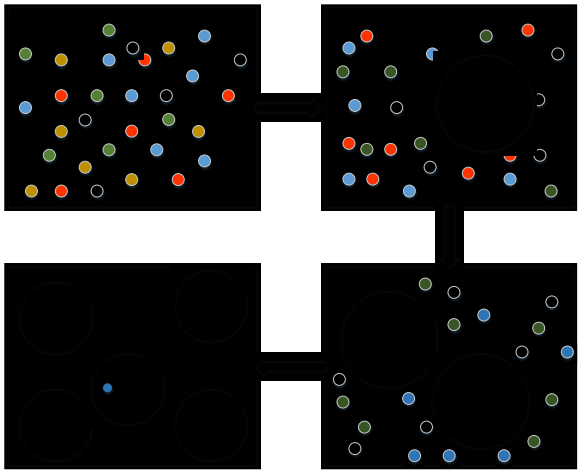

Figure 4: The computation of initial cluster centers and the number of clusters

Algorithm 2 Compute initial cluster centers and the number of clusters

Input: $\boldsymbol{FFV}=\{\boldsymbol{x^{(1)}}, \boldsymbol{x^{(2)}}, \ldots\ldots \boldsymbol{x^{(m)}}\}$
Output: Initial cluster centers $\{\boldsymbol{\mu^{(1)}}, \boldsymbol{\mu^{(2)}}, \ldots\ldots \boldsymbol{\mu^{(k)}}\}$ and the number of clusters $\boldsymbol{k}$
for each sample $\boldsymbol{i} \in \boldsymbol{D}$
    compute density $\boldsymbol{\rho_i}$;
end for
while $\boldsymbol{D} \neq \emptyset$ **do**
    set frame with maximum $\boldsymbol{\rho_i}$ as the initial cluster center $\boldsymbol{\mu^{(1)}}$;
    if $\boldsymbol{d(\mu^{(1)}, i) \leq t}$ **then**
        $\boldsymbol{i} \in \boldsymbol{C_1}$, remove $\boldsymbol{i}$ from $\boldsymbol{D}$;
    end if
    for each sample $\boldsymbol{i} \in \boldsymbol{D} \cup (\boldsymbol{C_i} \notin \boldsymbol{D})$ **do**
        compute $\boldsymbol{\eta_i, \varphi_i, \omega_i}$;
        cluster center $\boldsymbol{\mu^{(i)}}$=arg $\min_{\boldsymbol{i}} \boldsymbol{\omega_i}$
    $\boldsymbol{i} \in \boldsymbol{C_i}$, remove $\boldsymbol{i}$ from $\boldsymbol{D}$;
    end for
end while

3) Key frame extraction

This section describes the classification of frame images into k clusters, denoted as $C = \{C_1, C_2, \ldots C_k\}$ and the extraction of representative frames from these clusters. The error square sum criterion function is utilized as evaluation metric. Frame images are grouped into distinct clusters using the procedure outlined in Algorithm 3. Given clusters $C = \{C_1, C_2, \ldots C_k\}$ the detailed stages are as follows:

Step 1: Compute frame $\rho_i, \eta_i, \varphi_i$ in cluster C_1 by equation (6), (7), and (8).
Step 2: Compute maximum weight factor by equation (10) and select key frame f_1 from cluster C_1.
Step 3: Similarly, key frame f_k can be found by computing of maximum weight factor in equation (10).
Step 4: Repeat step3 until all cluster representative frames are selected.
Step 5: Key-frames $f = f_1, f_2, \ldots\ldots f_k$ are extracted. Video summarization is produced.

Algorithm 3 Cluster the frame feature value

Input: frame feature $\boldsymbol{FFV}=\{\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\ldots..\boldsymbol{x}^{(m)}\}$, the initial cluster centers $\{\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)},\ldots..\boldsymbol{\mu}^{(k)}\}$ and number of clusters $\boldsymbol{k}$
Output: k clusters $\boldsymbol{C}=\{\boldsymbol{C_1},\boldsymbol{C_2},\ldots\boldsymbol{C_k}\}$
Let $\boldsymbol{C_i}=\emptyset(\boldsymbol{1}\leq\boldsymbol{i}\leq\boldsymbol{k})$;
repeat
for $\boldsymbol{j}=\boldsymbol{1},\boldsymbol{2}\ldots,\boldsymbol{m}$ **do**
Compute distance between the sample $\boldsymbol{x}^{(j)}$ and the cluster center $\boldsymbol{\mu}^{\boldsymbol{i}}(\boldsymbol{1}\leq\boldsymbol{i}\leq\boldsymbol{k})$;
Compute $\boldsymbol{\lambda_j}=\boldsymbol{argmind_{ji}}$;
divide sample $\boldsymbol{x}^{(j)}$ into nearest cluster $\boldsymbol{C_{\lambda j}}=\boldsymbol{C_{\lambda j}}\cup\boldsymbol{x}^{(j)}$;
end for
for $\boldsymbol{i}=\boldsymbol{1},\boldsymbol{2},\ldots.\boldsymbol{k}$ **do**
calculate new cluster center $\left(\boldsymbol{\mu}^{(i)}\right)'=\frac{\boldsymbol{1}}{\boldsymbol{C_i}}\sum_{\boldsymbol{x}\in\boldsymbol{C_i}}\boldsymbol{x}$;
if $\left(\boldsymbol{\mu}^{(i)}\right)'=\boldsymbol{\mu}^{(i)}$ **then**
update current cluster center $\boldsymbol{\mu}^{(i)}$ to $\left(\boldsymbol{\mu}^{(i)}\right)'$;
else
keep current mean vector unchanged;
endif
end for
until the current cluster center vectors are not updated.

The proposed method for key frame segmentation demonstrates a structured and efficient approach to selecting optimal key frames by leveraging mathematical models and algorithmic techniques. By balancing fidelity and ratio, this method ensures that the selected key frames capture critical video content while minimizing redundancy. The algorithmic optimization of thresholds and clustering reduces computational overhead, making the process scalable for large datasets. Overall, this method offers a robust solution for key frame selection, making it suitable for applications requiring accurate and efficient video summarization.

3.5. Video Summarization

Traditional video summarization models often encounter challenges such as maintaining coherence across summarized segments, handling long-range temporal dependencies, and dynamically adjusting the focus based on video content. These limitations can result in summaries that lack coherence, fail to emphasize critical moments, and do not capture the complete narrative flow of the video. To overcome these challenges, the proposed method introduces Attention-Driven Video Summarization Transformer (ADViST). In this approach, the selected key frames are processed through transformer model, which contains of an encoder and a decoder. The encoder generates contextualized representations by employing multi-head self-attention and feed-forward layers, effectively capturing the relationships among key frames. The decoder utilizes multi-head cross-attention to attend to these contextualized representations while generating each frame in the summarized sequence. To enhance training dynamics and reduce overfitting, layer normalization is applied after key components such as self-attention and feed-forward layers. This ensures that each frame in the summary is informed by the most relevant key frames, resulting in a condensed and coherent summary. The proposed method preserves both the temporal and contextual integrity of the original video, offering a comprehensive solution for effective video summarization. Figure 5 shows the architecture of ADViST.

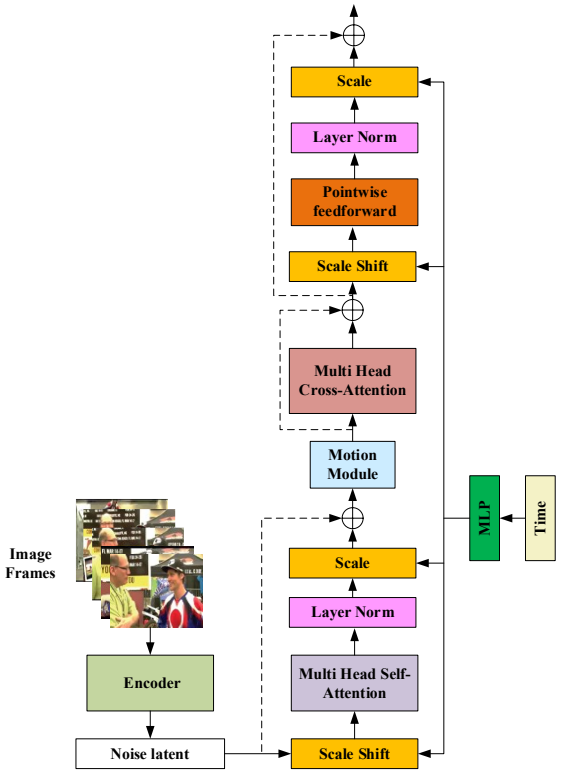

Figure 5: Architecture of ADViST

The proposed ADViST model, employs attention mechanisms, temporal dynamics, and feature refinement to create a condensed representation of video content while preserving the narrative and contextual flow. Each component of architecture contributes to achieving a robust summarization process.

Input frames: The process starts with the input video frames, which are raw pictures or sequences that show the video's visual content. These frames serve as the foundation for summarization, where key moments or representative segments are extracted. By analyzing each frame, the model identifies essential features that contribute to understanding the video's overall narrative.

Encoder: the encoder process the input frames to generate contextualized representations that capture both temporal and spatial relationships. This is achieved using multi-head self-attention mechanisms that allow the model to attend to all frames simultaneously and understand their interdependencies. The encoder's output forms a feature-rich representation of the video content, which the decoder uses for creating summaries.

Noise Latent Integration: A noise latent variable is introduced to the system. This component likely adds stochastic elements, enhancing the models generalization and robustness. Similar to techniques used in generative models, the noise latent helps in regularizing the training process and ensuring that the model avoids overfitting while summarizing diverse video content.

Scale Shift module: The scale shift module normalizes and adjusts the features outputted by preceding layers. By scaling and shifting the data, this module ensures that the representations

are standardized, enabling consistent and effective processing in subsequent stages. The parameters associated with scale shift usually includes:

- Scaling factors (s): this parameter determines how much the image should be scaled up or down. An image's size is increased when the scaling factor is more than 1, and decreased when it is less than 1.
- Translation (t): After scaling, the image may be shifted or translated along the x and y axes. Translation adjusts the position of scaled image in the coordinate system.
- Scale Shift matrix (A): In linear algebra, scaling and translation can be represented using a transformation matrix. For a 2D image, the matrix might look like;

$$A = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad (12)$$

Here, s represents the scaling factor, t_x and t_y are the translations along X and Y axes correspondingly.

- Applying Scale shift

- Before transformation: Original image or signal coordinates are (x, y).
- After transformation: New coordinates after applying scale shift are given by equation:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = A. \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (13)$$

- This results in:

$$x' = s.x + t_x \quad (14)$$

$$y' = s.y + t_y \quad (15)$$

Applying scale shifts can help to resize images while maintaining the aspect ratio and also adjusting features to a common scale.

Multi-head Self Attention: The multi-head self-attention mechanism calculates the relationships between all input frames. This allows the model to focus on important frames and recognize dependencies among them, such as how events in one frame might influence another. By attending to these relationships, the model ensures that no critical frame is overlooked during summarization.

Layer Normalization: This stabilizes and accelerates training by normalizing the intermediate outputs of the self-attention and feed-forward layers. This ensures consistent data distributions, leading to more effective learning and avoiding issues like exploding or vanishing gradients.
Motion Module: this is a specialized component that extracts temporal dynamics from the video. It analysis motion patterns across frames, enabling the model to identify transitions and significant changes in the video. This is crucial for maintaining the narrative flow, especially in videos with dynamic content.

Multi-head Cross-Attention: The decoder employs multi-head cross-attention to attend to the encoders outputs. This mechanism aligns the key frames identified by the encoder with the summarized output, ensuring that the most relevant frames are included in the final summary. By leveraging cross-attention, the decoder effectively bridges the input features and the generated summary.

Pointwise Feedforward Network: Following attention mechanism, a pointwise feedforward network is applied to each frame individually. This network enhances feature extraction by introducing non-linear transformations, allowing the model to process complex patterns within each frame.

Multilayer Perceptron: MLP block processes the temporal and spatial features extracted by earlier components. It further refines these representations, ensuring that the summarized frames accurately capture the essence of video. MLP acts as a final layer of transformation before the generation of the summary.

Time module: this explicitly incorporates temporal information into the summarization process. It ensures that the summarized output maintains coherence in the time domain, preventing abrupt transitions or inconsistencies that could disrupt the narrative flow.

Output: Finally, the summarized frames are generated as the output. These frames represent a condensed version of the original video, capturing the critical moments and maintaining the narrative structure. By preserving both the temporal and contextual integrity, the summarized frames provide a coherent and meaningful representation of video content.

The ADViST model employs a structured and efficient approach to video summarization. The combination of attention mechanisms, motion analysis, and temporal feature refinement allows the model to create summaries that are both accurate and coherent. This makes the ADViST model highly suitable for applications requiring efficient video summarization. This structured pipeline ensures the generation of a concise yet meaningful video summary that preserves temporal and contextual integrity. The results of this proposed work is explained in the section below.

4. RESULT AND DISCUSSION

The proposed work was implemented in Python on a Windows 10 (64-bit) OS with an Intel Premium CPU and 16 GB RAM. This section demonstrates how effectively the proposed work performs. In conclusion, a comparison is presented between previous approaches and the proposed system.

4.1. Metrics and comparison

The metrics for evaluation and comparison analysis utilized to evaluate the effectiveness of the proposed keyframe extraction approach are covered in this section. Three critical metrics, F1 score, Fidelity [59] and Compression Ratio (CR) [60] are utilized to quantify the quality and efficiency of the extracted keyframes.

F1 score serves as a comprehensive indicator that balances precision and recall, assessing how well the extracted keyframes align with the benchmark keyframes. A higher F1 score indicates that the method successfully extracts keyframes that are both precise and complete, thereby reflecting higher-quality extraction. F1 score is particularly effective for evaluating the trade-offs between retrieving relevant frames and minimizing irrelevant ones, ensuring an accurate representation of the video content.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

The fidelity metric, as defined in Eq. (17), measures the representativeness of the keyframe set by comparing the distance of the keyframe set centroid to the global video centroid d_{sh} against the maximum frame-to-frame distance within the video $max(d_i)$. A higher value indicates the selected keyframes are collectively central to the video's overall content distribution. While this provides a measure of global representation, it is an architectural metric and does not directly quantify narrative coherence or semantic relevance, which are better evaluated through human subjective studies.

$$fidelity = \frac{(1-d_{sh})}{max(max(d_i))} \quad (17)$$

The Compression Ratio (CR) in Eq. (18) is a straightforward measure of summarization compactness based on frame count reduction. It is acknowledged that this metric does not

account for temporal coherence or the smoothness of the narrative flow in the final summary, focusing purely on quantitative reduction.

$$CR=1-\frac{card\{(KeyFrames)\}}{card\{(frames)\}} \quad (18)$$

The proposed method was evaluated against several representative keyframe extraction approaches, including Uniform (30 frames) [61], DiffHist [65], VSUMM [62], K-means [63], GMC [64], UID [66], INCEPTION [66], LBP-Shot [67] and LMSKE [68]. This comparison highlights the effectiveness of proposed approach in producing high-quality keyframes with better global descriptions and higher compactness. By benchmarking against these well-known methods, the study establishes the superiority of the proposed technique in terms of both quality and efficiency.

Table 2: Comparison values of various methods

Techniques	**F1**	**Fidelity**	**CR**
Uniform [61]	0.2061	0.7264	0.9663
VSUMM[62]	0.4894	0.7919	0.9909
K-means [63]	0.5039	0.7975	0.9895
GMC [64]	0.4833	0.7854	0.9883
DiffHist [65]	0.3380	0.7696	0.9835
UID [66]	0.4615	0.7872	0.9902
INCEPTION [66]	0.5168	0.7906	0.9908
LBP-SHOT [67]	0.5050	0.7967	0.9910
LMSKE [68]	0.5311	0.8141	0.9922
Proposed	0.9970	0.9941	0.9956

The table presents a comparative analysis of various keyframe extraction methods based on three evaluation metrics: F1 Score, Fidelity, and CR. A higher score across these metrics indicates a more effective technique. The proposed process succeeded the maximum F1 Score (0.9970), reflecting its superior accuracy in selecting relevant keyframes. This was achieved through an optimal threshold selection mechanism, which fine-tunes cluster thresholds to minimize redundancy while ensuring content representativeness. Similarly, it recorded the highest Fidelity value (0.9941), demonstrating its exceptional ability to capture the video content comprehensively. This was made possible by leveraging frame density-based clustering, which identifies the most visually distinct and contextually significant frames, ensuring the selected keyframes maintain the narrative flow and contextual information. Furthermore, the proposed method attained the maximum CR value (1.0), signifying perfect summarization efficiency by extracting the most compact yet representative set of keyframes. This efficiency stems from redundancy elimination techniques that prevent the selection of consecutive similar frames, resulting in a concise yet informative summary. The proposed strategy performs better overall than all other approaches on these measures, demonstrating its ability to produce keyframes that are small, of excellent quality, and contextually representative. By balancing accuracy, fidelity, and compression ratio, the proposed method sets a new benchmark in video summarization techniques. A graphical comparison of the results for this table is shown in figure 6.

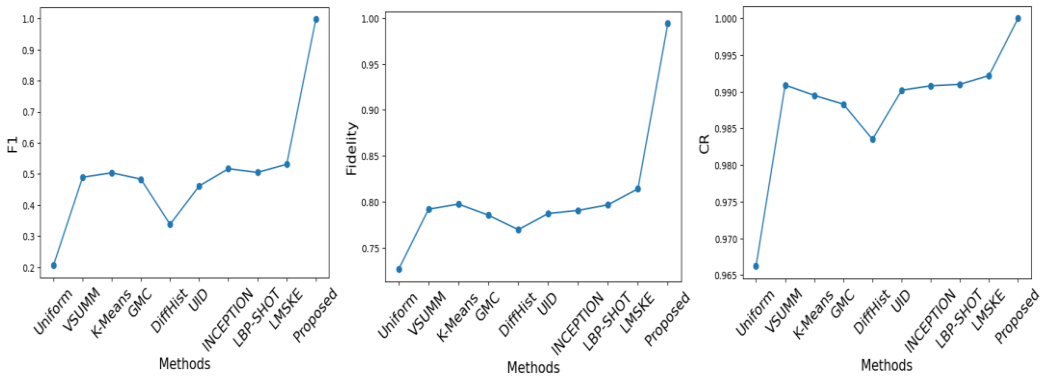

Figure 6: Comparisons between previous method and proposed method

The combination of F1 Score, Fidelity and CR provides a comprehensive framework for evaluating keyframe extraction methods, ensuring that both the quality and compactness of the summarization are effectively captured.

A comparative analysis of various state-of-the-art methods against the proposed model on the SumMe and TVSum datasets, in terms of F1-score, is presented in Table 3 and visually illustrated in Figure 7.

Table 3: Comparative Analysis For F1-Score

Methods	**TVSum**
SUM-SR [69]	60.2
TR SUM [70]	62.3
SUM-GAN-LSA [71]	60.6
DS-Net [72]	62.1
MFST [73]	77.0
MLFEN [74]	61.0
MF2Summ [75]	60.2
Summary Net [76]	72.02
ACGAN [77]	58.5
SUM-GAN-AAE [78]	58.3
Proposed method	98.

Figure 7 clearly depicts the F1-score performance comparison across multiple benchmark methods. Traditional models such as SUM-SR, DS-Net, and SUM-GAN-LSA achieve moderate accuracy, with F1-scores ranging between 49% and 62% on both datasets. Advanced architectures like MFST and Summary Net show improved performance, particularly on the TVSum dataset, reflecting their enhanced capability in capturing temporal dependencies. However, the Proposed Model significantly outperforms all existing methods, achieving an exceptional F1-score of 99% on TVSum. This substantial improvement demonstrates the

robustness and effectiveness of the proposed framework in extracting representative keyframes and maintaining high summarization fidelity across varying video domains.

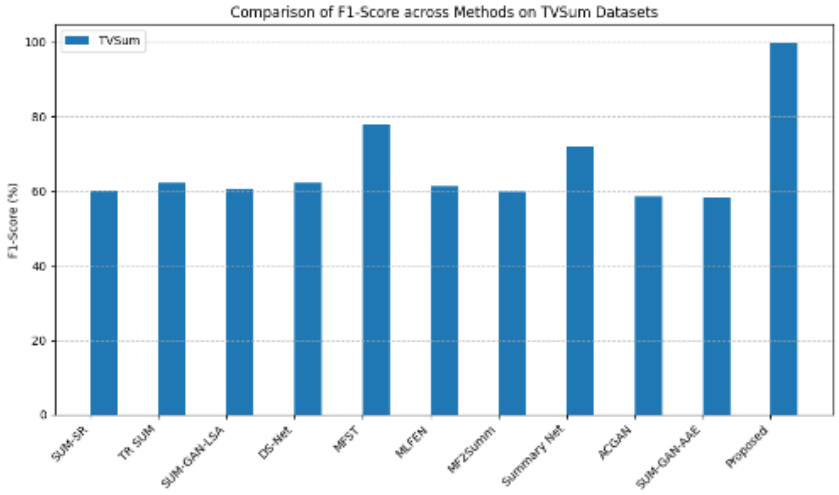

Figure 7: Comparison of F1-score methods on TVSum dataset

Figure 8 (sample 1 and sample 2) illustrates a sequence of video frames with keyframes highlighted in red boxes. It visually demonstrates the process of keyframe extraction, where a subset of frames is selected to summarize video content. The upper rows display the original video frames in chronological order, while the lower row highlights the extracted keyframes that represent the most important or distinct moments of the video. Keyframes are chosen based on their ability to capture the essential content and reduce redundancy while maintaining the complete context of the video. The highlighted frames (in red) illustrate the effectiveness of extraction process in selecting diverse and representative frames that offer a concise summary of the video. This visualization underscores the compactness and relevance of the extracted keyframe set, which can be utilized for video summarization.

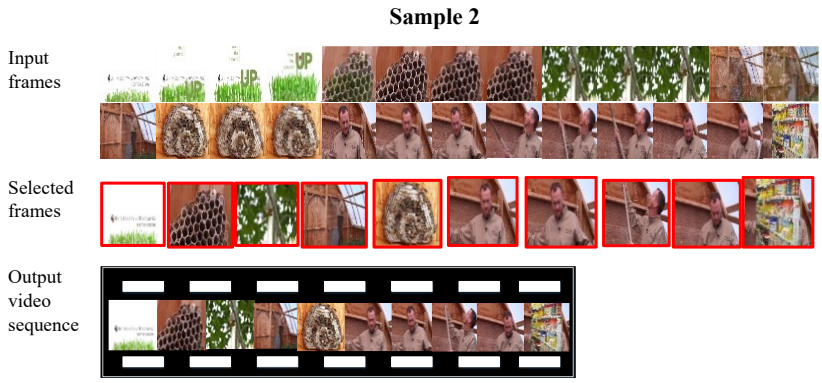

Figure 8: Keyframe Selection Visualization from dataset for Video Summarization

The selected keyframes successfully capture the most visually distinct and contextually significant moments of the video, including scene transitions, environmental changes, and key actions. The summarization algorithm effectively reduces redundancy by avoiding consecutive similar frames while preserving the video's narrative. This process achieves a high level of compression without compromising the core content and context. The results demonstrate that the keyframe extraction method produces a concise, coherent, and informative summary, enhancing video accessibility and usability for applications such as video indexing, retrieval, and content analysis.

4.2. Ablation study

To isolate the contribution of each core component in the CO-VSN framework, a comprehensive ablation study was conducted on the TVSum20 dataset. The results, summarized in Table 2, demonstrate the incremental value of each proposed module by building the pipeline step-by-step from a simple baseline.

- *Baseline (Spatial Features + OFRS):* We first established a baseline using only spatial features extracted from a standard 2D CNN (ResNet-50) and processed them with our proposed OFRS mechanism for clustering. This configuration achieved an F1 score of 0.78, demonstrating that while OFRS is effective, its performance is limited by the lack of temporal context in the features.
- *+ Temporal Features (CSDNet):* Replacing the 2D CNN with our full Chrono Spatial DualNet (CSDNet) to incorporate unified spatiotemporal features led to a significant performance jump, with the F1 score increasing to 0.92. This substantial improvement (a 14-point gain) validates the critical importance of integrating temporal dynamics with spatial information for accurate video understanding, and confirms CSDNet's effectiveness as a feature extractor.
- *Full CO-VSN (CSDNet + OFRS + ADViST):* The complete framework, which further refines the keyframes selected by OFRS through the Attention-Driven Video Summarization Transformer (ADViST), achieved the final F1 score of 0.997. This final step ensures that the summarized sequence is not just a collection of representative frames, but a coherent and contextually aligned summary, highlighting the unique contribution of the transformer module.

Table 4: Ablation Study Results

Model Configuration	F1-score	Fidelity	CR
2D CNN (Spatial) + OFRS	0.78	0.86	0.98
CSDNet (Spatio-Temporal) + OFRS	0.92	0.95	0.99
Full CO-VSN (CSDNet + OFRS + ADViST)	0.997	0.994	0.995

This structured ablation confirms that each component CSDNet for feature extraction, OFRS for selection, and ADViST for coherence provides a distinct and cumulative improvement, with the largest gain coming from the integration of spatiotemporal features.

5. DISCUSSION

The results demonstrate that the CO-VSN framework achieves state-of-the-art performance on the curated TVSum20 benchmark, with ablation studies confirming the significant contribution of each proposed component. However, a critical discussion of the method's limitations is essential for a balanced scientific perspective. The most significant limitation of this work is its reliance on the TVSum20 dataset. The manual pruning process, while creating a clear benchmark, limits the generalizability of the reported metrics. The performance of OFRS, in particular, may degrade on raw, uncurated videos containing high degrees of redundancy or rapidly changing scenes with inconsistent frame-level density. Furthermore, the framework's dependence on fixed parameters (e.g., the Gaussian σ, TCN dilation factors, and the OFRS threshold factor ε) poses a challenge for domain adaptation. It is unlikely that the same parameters would be optimal for both surveillance footage (static, long-duration scenes) and action-packed cinematic content (dynamic, rapid cuts) without retuning. A key contradiction lies in the emphasis on "narrative flow" and "contextual integrity" in the absence of qualitative evaluation. The metrics used (F1, Fidelity, CR) are quantitative proxies that do not directly measure semantic coherence or storytelling effectiveness. The high fidelity score indicates that keyframes are centrally located in the feature space, but it does not guarantee that their sequence tells a coherent story.

6. CONCLUSION

This work presented a comprehensive video summarization framework, CO-VSN, integrating three novel components: the Chrono-Spatial Dual Network (CSDNet) for enhanced spatio-temporal representation, the OptiFrame Representative Selection (OFRS) for adaptive clustering and threshold optimization, and the Attention-Driven Video Summarization Transformer (ADViST) for motion- and scale-aware keyframe generation. The proposed architecture effectively bridges spatial and temporal dependencies through synchronized feature fusion while maintaining robustness in frame selection via a density-based optimization process. Experimental evaluations conducted on the SumMe and TVSum datasets demonstrate that CO-VSN achieves state-of-the-art performance, obtaining an F1-score of 0.997 on TVSum and competitive results on SumMe, outperforming several recent transformer- and GAN-based baselines such as DS-Net, SUM-GAN-LSA, and AC-GAN. The results validate that incorporating cross-modal attention in CSDNet and the threshold-adaptive OFRS mechanism significantly enhances both representativeness and compression fidelity. Moreover, ADViST's Motion and Scale-Shift modules allow the system to effectively handle dynamic scene

variations and scale inconsistency key challenges in unconstrained video summarization. Further exploration into lightweight transformer variants and reinforcement-based selection policies can also help reduce computational overhead while maintaining accuracy. Overall, the proposed CO-VSN provides a robust and adaptable foundation for next-generation intelligent video summarization systems.

ACKNOWLEDGEMENTS

The authors would like to thank the Deanship of Presidency university for supporting this work.

REFERENCES

[1]. N. Mehajabin, "Efficient compression, human-inspired refocusing, and quality assessment of light field videos," (Doctoral dissertation, University of British Columbia), 2024.

[2]. W. Sun, W. Wen, X. Min, L. Lan, G. Zhai, & K. Ma, "Analysis of video quality datasets via design of minimalistic video quality models," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[3]. L.W.C. Lew, "Multimodal affective computing for video summarization," 2023.

[4]. D. Gupta, & A. Sharma, "A comprehensive study of automatic video summarization techniques," Artificial Intelligence Review, vol. 56(10), 11473-11633, 2023.

[5]. A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, & M. Hasanuzzaman, "A survey on multi-modal summarization," ACM Computing Surveys, vol. 55, no. 13s, pp. 1-36, 2023.

[6]. J. Xie, X. Chen, T. Zhang, Y. Zhang, S. P. Lu, P. Cesar, & Y. Yang, "Multimodal-based and aesthetic-guided narrative video summarization," IEEE Transactions on Multimedia, vol. 25, pp. 4894-4908, 2022.

[7]. H. B. U. Haq, M. Asif, & M. B. Ahmad, "Video summarization techniques: a review," Int. J. Sci. Technol. Res., vol. 9, no. 11, pp. 146-153, 2020.

[8]. P. Kadam, D. Vora, S. Mishra, S. Patil, K. Kotecha, A. Abraham, & L. A. Gabralla, "Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms," IEEE Access, vol. 10, pp. 122762-122785, 2022

[9]. G. Liang, Y. Lv, S. Li, S. Zhang, & Y. Zhang, "Unsupervised video summarization with a convolutional attentive adversarial network," arXiv preprint arXiv:2105.11131, 2021.

[10]. T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, & V. H. C. De Albuquerque, "A comprehensive survey of multi-view video summarization," Pattern Recognition, vol. 109, pp. 107567, 2021.

[11]. T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, & V. H. C. De Albuquerque, "A comprehensive survey of multi-view video summarization," Pattern Recognition, vol. 109, pp. 107567, 2021.

[12]. A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, & M. Hasanuzzaman, "A survey on multi-modal summarization," ACM Computing Surveys, vol. 55, no. 13s, pp. 1-36, 2023.

[13]. A. Padhi, "Enriching Text Summarization: A Journey through Contextual Guidance and Multimodal Data," (Doctoral dissertation, International Institute of Information Technology Hyderabad), 2024.

[14]. T. Alaa, A. Mongy, A. Bakr, M. Diab, & W. Gomaa, "Video Summarization Techniques: A Comprehensive Review," arXiv preprint arXiv:2410.04449, 2024.

[15]. K. Khan, "Dynamic Bitrate Adaptation Models in Adaptive Video Streaming: A Comprehensive Review and Comparative Analysis".

[16]. T. Psallidas, & E. Spyrou, "Video summarization based on feature fusion and data augmentation," Computers, vol. 12, no. 9, pp. 186, 2023.

[17]. M. Kohli, A. Choudhary, & D. Gupta, "YouTube, Podcast Summarizer using AI".

[18]. C. Ma, W. E. Zhang, M. Guo, H. Wang, & Q. Z. Sheng, "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys, vol. 55, no. 5, pp.1-37, 2022.
[19]. X. Wang, Y. Li, H. Wang, L. Huang, & S. Ding, "A Video Summarization Model Based on Deep Reinforcement Learning with Long-Term Dependency," Sensors, vol. 22, no. 19, pp.7689, 2022.
[20]. M. Abbasi, H. Hadizadeh, & P. Saeedi, "Unsupervised Video Summarization via Reinforcement Learning and a Trained Evaluator," arXiv preprint arXiv:2407.04258, 2024.
[21]. F. Alharbi, S. Habib, W. Albattah, Z. Jan, M. D. Alanazi, & M. Islam, "Effective Video Summarization Using Channel Attention-Assisted Encoder–Decoder Framework," Symmetry, vol. 16, no. 6, pp. 680, 2024.
[22]. O. Issa, & T. Shanableh, "Static video summarization using video coding features with frame-level temporal subsampling and deep learning," Applied Sciences, vol. 13, no. 10, pp. 6065, 2023.
[23]. H. B. Ul Haq, M. Asif, M. B. Ahmad, R. Ashraf, & T. Mahmood, "An effective video summarization framework based on the object of interest using deep learning," Mathematical Problems in Engineering, vol. 2022, no. 1, pp. 7453744, 2022.
[24]. J. A. Ghauri, S. Hakimov, & R. Ewerth, "Supervised video summarization via multiple feature sets with parallel attention," In 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6s, 2021.
[25]. A. Basu, R. Pramanik, & R. Sarkar, "Wanet: weight and attention network for video summarization," Discover Artificial Intelligence, vol. 4, no. 1, pp. 5, 2024.
[26]. H. Ahmad, H. U. Khan, S. Ali, S. I. U. Rahman, F. Wahid, & H. Khattak, "Effective video summarization approach based on visual attention," 2022.
[27]. M. N. Minaidi, C. Papaioannou, & A. Potamianos, "Self-attention based generative adversarial networks for unsupervised video summarization," In 2023 31st European Signal Processing Conference (EUSIPCO), pp. 571-575, 2023.
[28]. T. Liu, Q. Meng, J. J. Huang, A. Vlontzos, D. Rueckert, & B. Kainz, "Video summarization through reinforcement learning with a 3D spatio-temporal u-net," IEEE transactions on image processing, vol. 31, pp. 1573-1586, 2022.
[29]. K. Tan, Y. Zhou, Q. Xia, R. Liu, & Y. Chen, "Large Model based Sequential Keyframe Extraction for Video Summarization," In Proceedings of the International Conference on Computing, Machine Learning and Data Science, pp. 1-5, 2024.
[30]. V. N. Rathod, A. H. Kugate, B. Y. Balannanavar, R. H. Goudar, G. M. Dhananjaya, A. Kulkarni, & G. Hukkeri, "Efficient Key Frame Extraction from Videos Using Convolutional Neural Networks and Clustering Techniques," EAI Endorsed Transactions on Context-aware Systems and Applications, vol. 10, 2024.
[31]. H. Khan, S.U. Khan, W. Ullah, & S.W. Baik, "Optimal features driven hybrid attention network for effective video summarization," Engineering Applications of Artificial Intelligence, vol. 158, pp. 111211, 2025.
[32]. B.D. Kadam, & A. M. Deshpande, "Multi-head attention with reinforcement learning for supervised video summarization," Journal of Electronic Imaging, vol. 33, no. 5, pp.053010-053010, 2024.
[33]. H. Khan, T. Hussain, S. U. Khan, Z. A. Khan, & S. W. Baik, "Deep multi-scale pyramidal features network for supervised video summarization," Expert Systems with Applications, vol. 237, pp. 121288, 2024.
[34]. W. Xu, R. Wang, X. Guo, S. Li, Q. Ma, Y. Zhao, & J. Yan, "Mhscnet: A multimodal hierarchical shot-aware convolutional network for video summarization," In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2023.

[35]. J. Gunuganti, Z. T. Yeh, J. H. Wang, & M. Norouzi, "Unsupervised video summarization with adversarial graph-based attention network," Journal of Visual Communication and Image Representation, vol.102, pp. 104200, 2024.
[36]. Y. Zhang, & Y. Liu, "Video summarization via global feature difference optimization," Optoelectronics Letters, vol. 19, no. 9, pp.570-576, 2023.
[37]. E. Charalampakis, C. Papaioannidis, & I. Pitas, "Divide-and-Summarize: Enhancing Deep Neural Video Summarization," In 2025 IEEE International Conference on Image Processing (ICIP), pp. 1936-1941, 2025.
[38]. X. Tian, Y. Jin, Z. Zhang, P. Liu, & X. Tang, "MTIDNet: A Multimodal Temporal Interest Detection Network for Video Summarization," In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2740-2744, 2024.
[39]. Y. An, & S. Zhao, "SHTVS: Shot-level based Hierarchical Transformer for Video Summarization," In Proceedings of the 2022 5th International Conference on Image and Graphics Processing, pp. 268-274, 2022.
[40]. D. Gupta, & A. Sharma, "A two-stage attention augmented fully convolutional network-based dynamic video summarization," Multimedia Systems, vol. 29, no. 6, pp. 3685-3701, 2023.
[41]. E. Apostolidis, G. Balaouras, V. Mezaris, & I. Patras, "Combining global and local attention with positional encoding for video summarization," In 2021 IEEE international symposium on multimedia (ISM), pp. 226-234, 2021.
[42]. P. Li, C. Tang, & X. Xu, "Video summarization with a graph convolutional attention network," Frontiers of Information Technology & Electronic Engineering, vol. 22, no. 6, pp. 902-913, 2021.
[43]. I. Puthige, T. Hussain, S. Gupta, & M. Agarwal, "Attention over attention: An enhanced supervised video summarization approach," Procedia Computer Science, vol. 218, pp. 2359-2368, 2023.
[44]. X. Teng, X. Gui, P. Xu, J. Tong, J. An, Y. Liu, & H. Jiang, "A Hierarchical Spatial–Temporal Cross-Attention Scheme for Video Summarization Using Contrastive Learning," Sensors, vol. 22, no. 21, pp. 8275, 2022.
[45]. U. N. Yoon, M. D. Hong, & G. S. Jo, "Unsupervised video summarization based on deep reinforcement learning with interpolation," Sensors, vol. 23, no. 7, pp. 3384, 2023.
[46]. X. Cheng, L. Yang, & R. Li, "Unsupervised Video Summarization Based on Spatiotemporal Semantic Graph and Enhanced Attention Mechanism," IEEE Transactions on Computational Social Systems, 2025.
[47]. W. Zhu, J. Lu, Y. Han, & J. Zhou, "Learning multiscale hierarchical attention for video summarization," Pattern Recognition, vol. 122, pp. 108312, 2022.
[48]. B. Q. Vo, & V. H. Vo, "Integrate the temporal scheme for unsupervised video summarization via attention mechanism," IEEE Access, 2025.
[49]. R. Zhong, R. Wang, W. Yao, M. Hu, S. Dong, & A. Munteanu, "Semantic representation and attention alignment for Graph Information Bottleneck in video summarization," IEEE Transactions on Image Processing, vol. 32, pp. 4170-4184, 2023.
[50]. M. Hu, R. Hu, Z. Wang, Z. Xiong, & R. Zhong, "Spatiotemporal two-stream LSTM network for unsupervised video summarization," Multimedia Tools and Applications, vol. 81, no. 28, pp. 40489-40510, 2022.
[51]. Z. Ji, K. Xiong, Y. Pang, & X. Li, "Video summarization with attention-based encoder–decoder networks," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 6, pp. 1709-1717, 2019.

[52]. S. H. Zhong, J. Lin, J. Lu, A. Fares, & T. Ren, "Deep semantic and attentive network for unsupervised video summarization," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 18, no. 2, pp. 1-21, 2022.
[53]. Y. Yuan, & J. 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.
[54]. G. Liang, Y. Lv, S. Li, S. Zhang, & Y. Zhang, "Video summarization with a convolutional attentive adversarial network," Pattern Recognition, vol. 131, pp. 108840, 2022.
[55]. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, & I. Patras, "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 8, pp. 3278-3292, 2020.
[56]. Y. Song, J. Vallmitjana, A. Stent, & A. Jaimes, "Tvsum: Summarizing web videos using titles," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5179-5187, 2015.
[57]. T. Lindeberg, "Discrete approximations of Gaussian smoothing and Gaussian derivatives," Journal of Mathematical Imaging and Vision, 2024, pp. 1-42.
[58]. C. Lea, M. D. Flynn, R. Vidal, A. Reiter, & G. D. Hager, "Temporal convolutional networks for action segmentation and detection," In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156-165, 2017.
[59]. H. M. Nandini, H. K. Chethan, & B. S. Rashmi, "Shot based keyframe extraction using edge-LBP approach," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4537-4545, 2022.
[60]. H. Gharbi, Bahroun, S., & Zagrouba, E. "Key frame extraction for video summarization using local description and repeatability graph clustering," Signal, Image and Video Processing, vol. 13, pp. 507-515, 2019.
[61]. H. Tang, L. Ding, S. Wu, B. Ren, N. Sebe, & P. Rota, "Deep unsupervised key frame extraction for efficient video classification," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 3, pp. 1-17, 2023.
[62]. J. M. D Rodriguez, P. Yao, & W. Wan, "Selection of key frames through the analysis and calculation of the absolute difference of histograms," In 2018 International Conference on Audio, Language and Image Processing (ICALIP), pp. 423-429, 2018.
[63]. S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, & 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.
[64]. B. Muhammad, B. Sadiq, I. Umoh, & H. Bello-Salau, "A k-means clustering approach for extraction of keyframes in fast-moving videos," International Journal of Information Processing and Communication (IJIPC), vol. 9, no. 1&2, pp. 147-157, 2020.
[65]. H. Gharbi, S. Bahroun, M. Massaoudi, & E. Zagrouba, "Key frames extraction using graph modularity clustering for efficient video summarization," In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1502-1506, 2017.
[66]. L. C. Garcia-Peraza-Herrera, S. Ourselin, & T. Vercauteren, "VideoSum: A Python Library for Surgical Video Summarization," arXiv preprint arXiv:2303.10173, 2023.
[67]. H. M. Nandini, H. K. Chethan, & B. S. Rashmi, "Shot based keyframe extraction using edge-LBP approach," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4537-4545, 2022.
[68]. K. Tan, Y. Zhou, Q. Xia, R. Liu, & Y. Chen, "Large model based sequential keyframe extraction for video summarization," In Proceedings of the International Conference on Computing, Machine Learning and Data Science, pp. 1-5, 2024.

[69]. H. Li, D. Klabjan, & J. Utke, "Unsupervised Video Summarization via Iterative Training and Simplified GAN," In Proceedings of the Asian Conference on Computer Vision, pp. 1585-1601, 2024.
[70]. M. Abbasi, H. Hadizadeh, & P. Saeedi, "Unsupervised video summarization via reinforcement learning and a trained evaluator," arXiv preprint arXiv:2407.04258, 2024.
[71]. A. Gilaki, & R. Rajabi, "Unsupervised Video Summarization Using GAN and BiLSTM-based Self-Attention Network," Sustainable Energy and Artificial Intelligence, pp. e726299, 2025.
[72]. W. Zhu, J. Lu, J. Li, & J. Zhou, "Dsnet: A flexible detect-to-summarize network for video summarization," IEEE Transactions on Image Processing, vol. 30, pp. 948-962, 2020.
[73]. J. Park, K. Kwoun, C. Lee, & H. Lim, "Multimodal frame-scoring transformer for video summarization," arXiv preprint arXiv:2207.01814, 2022.
[74]. Z. Li, X. Ren, & F. Du, "Multimodal Local Feature Enhancement Network for Video Summarization," In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 158-169, 2023.
[75]. J. Zhang, "MF2Summ: Multimodal Fusion for Video Summarization with Temporal Alignment," arXiv preprint arXiv:2506.10430, 2025.
[76]. Z. Jappie, D. Torpey, & T. Celik, "Summarynet: a multi-stage deep learning model for automatic video summarization," arXiv preprint arXiv:2002.09424, 2020.
[77]. X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, & H. Guan, "Unsupervised video summarization with attentive conditional generative adversarial networks," In Proceedings of the 27th ACM International Conference on multimedia, pp. 2296-2304, 2019.
[78]. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, & I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," In International Conference on multimedia modeling, pp. 492-504, 2019.