http://dx.doi.org/10.62762/TSCC.2025.308540

RESEARCH ARTICLE

Dual-Pathway Sensing with Optimized Attention Network for Video Summarization in Surveillance Systems

Taimur Ali Khan[1], **Danish Ali**[2], **Zainab Ghazanfar**[3] **and Bilal Ahmad**[4,*]

[1] Department of IT, Saudi Media Systems, Riyadh 11482, Saudi Arabia
[2] Department of Electrical and Computer Engineering, Villanova University, Villanova, PA 19085, United States
[3] Department of Software and Artificial Intelligence, Gachon University, Seongnam 13120, South Korea
[4] Department of Computer Science, Govt Degree College Lal Qilla Maidan Dir Lower, Pakistan

Abstract

Video summarization (VS) aims to generate concise representations of long videos by extracting the most informative frames while maintaining essential content. Existing methods struggle to capture multi-scale dependencies and often rely on suboptimal feature representations, limiting their ability to model complex inter-frame relationships. To address these issues, we propose a multi-scale sensing network that incorporates three key innovations to improve VS. First, we introduce multi-scale dilated convolution blocks with progressively increasing dilation rates to capture temporal context at multiple levels, enabling the network to understand both local transitions and long-range dependencies. Second, we develop a Dual-Pathway Efficient Channel Attention (DECA) module that leverages statistics from Global Average Pooling and Global Max Pooling pathways. Third, we suggest an Optimized Spatial Attention (OSA) module that replaces standard 7×7 convolutions with more efficient operations while maintaining spatial dependency modeling. The proposed framework uses EfficientNetB7 as the backbone for robust spatial feature extraction, followed by multi-scale dilated blocks and dual attention mechanisms for detailed feature refinement. Extensive tests on the TVSum and SumMe benchmark datasets demonstrate the superiority of our method, achieving F1 Scores of 63.5% and 53.3%, respectively.

Keywords: video summarization, visual intelligence, surveillance systems, dual-pathway, attention network.

1 Introduction

The proliferation of visual content has fundamentally transformed the way we capture, process, and analyze multimedia data in real-time applications. Current trends indicate that social networks are reporting unprecedented levels of engagement, exemplified by Facebook's daily viewership, which reached 8 billion video interactions [1]. This exponential expansion of visual data streams presents both opportunities and substantial computational challenges for computer

Submitted: 20 October 2025
Accepted: 05 December 2025
Published: 30 December 2025

Vol. 2, **No.** 4, 2025.
10.62762/TSCC.2025.308540

***Corresponding author:**
✉ Bilal Ahmad
bilalahmadcs03@gmail.com

Citation
Khan, T. A., Ali, D., Ghazanfar, Z., & Ahmad, B. (2025). Dual-Pathway Sensing with Optimized Attention Network for Video Summarization in Surveillance Systems. *ICCK Transactions on Sensing, Communication, and Control*, 2(4), 276–289.

vision applications, particularly those that require real-time analysis of continuous visual feeds [2]. The sheer magnitude of the accumulated visual data creates significant bottlenecks in extracting actionable intelligence from video archives. Security and surveillance infrastructures exemplify this challenge, in which monitoring personnel face the daunting task of reviewing extensive footage repositories to identify critical events [3].

Traditional manual inspection approaches prove inadequate when confronted with terabytes of continuous recordings, necessitating intelligent automated frameworks capable of identifying salient patterns and events of interest. These limitations underscore the urgent need for complex computer vision methodologies that can efficiently process, interpret, and extract meaningful information from massive-scale video databases without human intervention. Modern deep learning paradigms have emerged as transformative solutions for automating visual content analysis across diverse application domains [4–6].

Contemporary approaches to intelligent video analysis can be taxonomically organized into unsupervised and supervised methodologies, each with distinct characteristics and performance profiles. Unsupervised techniques operate without explicit human guidance, instead relying on heuristic measures such as dissimilarity metrics [7], representativeness criteria [8], reconstruction-error minimization [9], and memorability assessment [10] to identify significant visual content. In contrast, supervised learning frameworks leverage human-annotated ground truth (GT) labels to guide training, enabling models to learn task-specific patterns that better reflect human judgment and domain expertise. Early supervised approaches employed recurrent architectures such as long short-term memory networks [11] to model temporal dependencies in sequential visual data. Subsequent research introduced more sophisticated hierarchical recurrent neural network architectures [12] to better capture multi-scale temporal relationships. Alternative strategies, exemplified by fully convolutional approaches [13], enabled parallel processing of temporal sequences while modeling complex structural dependencies across frames.

The advent of attention mechanisms [14, 15] has catalyzed a paradigm shift in how deep networks process sequential visual information. Attention-based architectures have been successfully integrated into various computer vision tasks, with encoder-decoder configurations [16] and dedicated attention modules [15] demonstrating remarkable capabilities in capturing long-range dependencies. Recent state-of-the-art methodologies [17–20] have further advanced the field by incorporating sophisticated attention mechanisms that enable models to selectively focus on relevant spatiotemporal regions while suppressing irrelevant information, thereby achieving enhanced performance in learning complex visual patterns across extended temporal horizons. The main contributions of our work are summarized below:

1.1 Contributions

- **Dual-Pathway Efficient Channel Attention (DECA) Mechanism:** Proposed a novel channel attention module that exploits both Global Average Pooling (GAP) and Global Max Pooling (GMP) pathways to capture complementary statistical information from feature maps. The module employs shared 1D convolutions (k=3) instead of fully connected layers for parameter efficiency, reducing computational complexity while maintaining channel interdependencies. A sigmoid-based gating mechanism with a residual connection is integrated to adaptively recalibrate channel-wise feature responses.

- **Optimized Spatial Attention (OSA) with Decomposed Convolutions:** Introduced an optimized spatial attention module that decomposes the traditional 7×7 convolution kernel into efficient 3×3 operations, significantly reducing parameters from 49 to 9 weights per position while maintaining the receptive field. This decomposition strategy achieves approximately 81% parameter reduction in spatial attention computation without sacrificing feature extraction capability.

- **Multi-Scale Dilated Convolution Architecture:** Designed a hierarchical multi-scale feature extraction network using three parallel dilated convolution blocks (D1, D2, D3) with progressively increasing dilation rates (3, 6, 9). This multi-scale sensing approach captures temporal dependencies at different granularities, enabling the network to comprehend both local frame transitions and long-range video context simultaneously.

- **EfficientNetB7-based Hierarchical Feature Learning Framework:** Leveraged EfficientNetB7 backbone (Blocks 1-7) for robust spatial feature extraction from video frames, combined with the proposed DECA and OSA modules for enhanced representational learning through optimized attention-driven feature sensing.
- **Extensive Experimental Validation:** Conducted comprehensive experiments on two benchmark video summarization (VS) datasets, TVSum and SumMe, demonstrating the effectiveness and superiority of the proposed multi-scale sensing network with dual-pathway attention mechanisms for automatic VS.

2 Related Work

Initial VS research utilized submodular functions, frame clustering, and low-rank compatibility models [21–23]. These early techniques faced significant performance limitations due to hand-engineered features and constrained model expressiveness. Contemporary approaches fall into two categories: unsupervised methods and supervised learning frameworks. Recently, attention-based mechanisms have emerged as powerful tools for supervised VS, enhancing both feature representation and selection capabilities. This section examines these methodologies, highlighting critical limitations and identifying research opportunities.

2.1 Unsupervised Video Summarization

Unsupervised methods select keyshots using heuristic principles that emphasize diversity, representativeness, and semantic relevance. Clustering-based techniques group visually similar shots into unified classes [24]. Various research efforts have integrated clustering into summarization pipelines: k-means approaches extract color-based features [25], while optimization-driven methods identify sparse correlation patterns [26]. Dictionary learning frameworks construct representative shot dictionaries from video content [27], with some work modeling videos as linear combinations of keyframes [7].

Other research directions include storyline smoothness modeling combined with importance scoring for people and objects [28], and patch-based chunk sparse representations [29]. Memorability-driven approaches score frames based on entropy and memorability metrics [10, 30, 31]. Industrial applications include resource-constrained surveillance systems [32] and static summarization using sparse autoencoders [33]. Adversarial learning has also been explored, with frameworks incorporating reward mechanisms for representativeness and diversity [8], and graph attention networks combined with bidirectional LSTMs to reduce keyframe redundancy [34]. The fundamental limitation of unsupervised approaches is their inability to leverage GT annotations, leading to lower accuracy than supervised alternatives that learn from human-labeled keyframes to maintain consistent selection criteria.

2.2 Supervised Video Summarization

Supervised methods encompass both traditional machine learning and deep learning paradigms. Traditional approaches employ handcrafted features with classical learning frameworks, including graph-based first-person summarization [35], category-specific techniques [36], and user-video-centric methods [37]. These techniques struggled with modeling long-range temporal dependencies, which are essential for effective keyframe identification.

Deep learning revolutionized sequential modeling in VS. Early work introduced bidirectional LSTM architectures with determinantal point processes for forward-backward temporal modeling [11], inspiring subsequent LSTM-based developments [16, 18, 38]. Notable contributions include quantitative loss functions for semantic relevance evaluation [18]. However, standard RNNs face computational challenges with lengthy sequences due to sequential processing requirements. Hierarchical RNN architectures address this through two-layered structures for improved long-range dependency handling [12], while integrated frameworks combine shot detection with significance prediction [39]. Sequence-to-sequence models with attentive encoders enable weighted feature importance [16], though training remains computationally intensive for large datasets.

Alternative architectures using global attention-based modules replace sequential LSTM processing and perform transformations in a single backward pass [40]. Anchor-based and anchor-free hybrid methods enhance temporal consistency [41]. Despite progress, supervised approaches require further investigation to model intricate inter-frame relationships, necessitating optimized attention mechanisms for capturing complex dependencies.

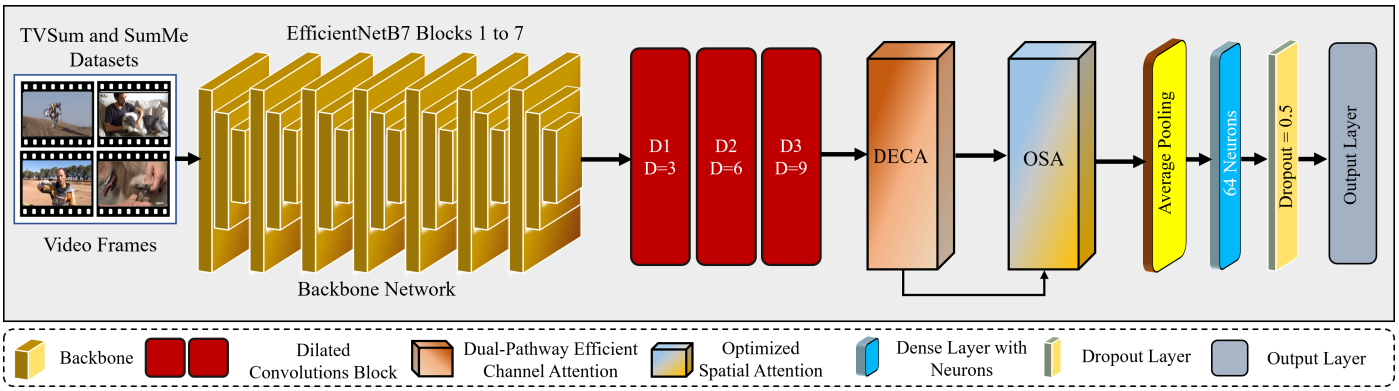

Figure 1. Overall architecture of the proposed multiscale sensing network for VS, using an EfficientNetB7 backbone Blocks 1 to 7 for spatial feature extraction, three parallel dilated convolution blocks, DECA and OSA based feature recalibration, and fully connected layers for frame importance prediction.

2.3 Attention Mechanisms

Attention mechanisms have driven substantial advances in visual intelligence, improving feature extraction and contextual reasoning [42, 43]. Foundation models have significantly enhanced summarization effectiveness [44]. Various attention-based architectures have emerged: LSTM-integrated attention layers capture frame relationships [45], while decoder-guided attention blocks utilize encoder outputs and hidden states [11]. Semantic-preserving embedding networks with specialized loss functions maintain content integrity [18], and refined self-attention mechanisms with preprocessing steps handle diverse visual content more effectively [17].

Recent work focuses on modeling multi-scale temporal structure. Multiscale hierarchical attention frameworks employ intra- and inter-block mechanisms for short- and long-range dependencies [19], achieving competitive performance through two-stream appearance-motion integration on benchmark datasets. Transformer-based networks leverage video feature patterns directly [20]. However, the prevalent reliance on GoogleNet Pool5 features limits progress, as generated attention patterns often fail to adequately capture salient content [46], indicating the need for more sophisticated attention architectures capable of modeling complex frame relationships.

3 Proposed Methodology

3.1 Network Overview

The proposed multi-scale sensing network for VS comprises several key components designed to extract discriminative features from video frames and predict their importance scores. The architecture begins with extracting video frames from benchmark datasets, followed by spatial feature encoding using an EfficientNetB7 backbone. To capture temporal dependencies at multiple scales, we introduce three parallel dilated convolution blocks with varying dilation rates. The extracted multi-scale features are subsequently refined through two attention mechanisms: DECA for channel-wise recalibration and OSA for spatial feature enhancement. Finally, the attended features pass through average pooling, dense layers with dropout regularization, and an output layer to generate frame-level importance predictions. Figure 1 illustrates the complete network architecture. Figure 2 illustrates the deployment pipeline of the proposed framework in surveillance systems.

3.2 Backbone Feature Extraction

Effective VS requires robust spatial feature representations from individual frames. We employ EfficientNetB7 as the backbone network due to its superior performance in balancing model complexity and feature extraction capability. The EfficientNetB7 architecture uses compound scaling, uniformly scaling network depth, width, and resolution, resulting in enhanced feature learning compared to conventional convolutional networks. The backbone consists of seven sequential blocks that progressively extract hierarchical features from input video frames. Each block incorporates mobile inverted bottleneck convolutions with squeeze-and-excitation optimization, enabling efficient feature propagation while maintaining computational efficiency. Given an input frame $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the backbone network produces feature maps $\mathbf{F}_{backbone} \in \mathbb{R}^{C \times H' \times W'}$, where C represents the number of channels, and H', W' denote the spatial dimensions after backbone processing. These extracted features serve as input to

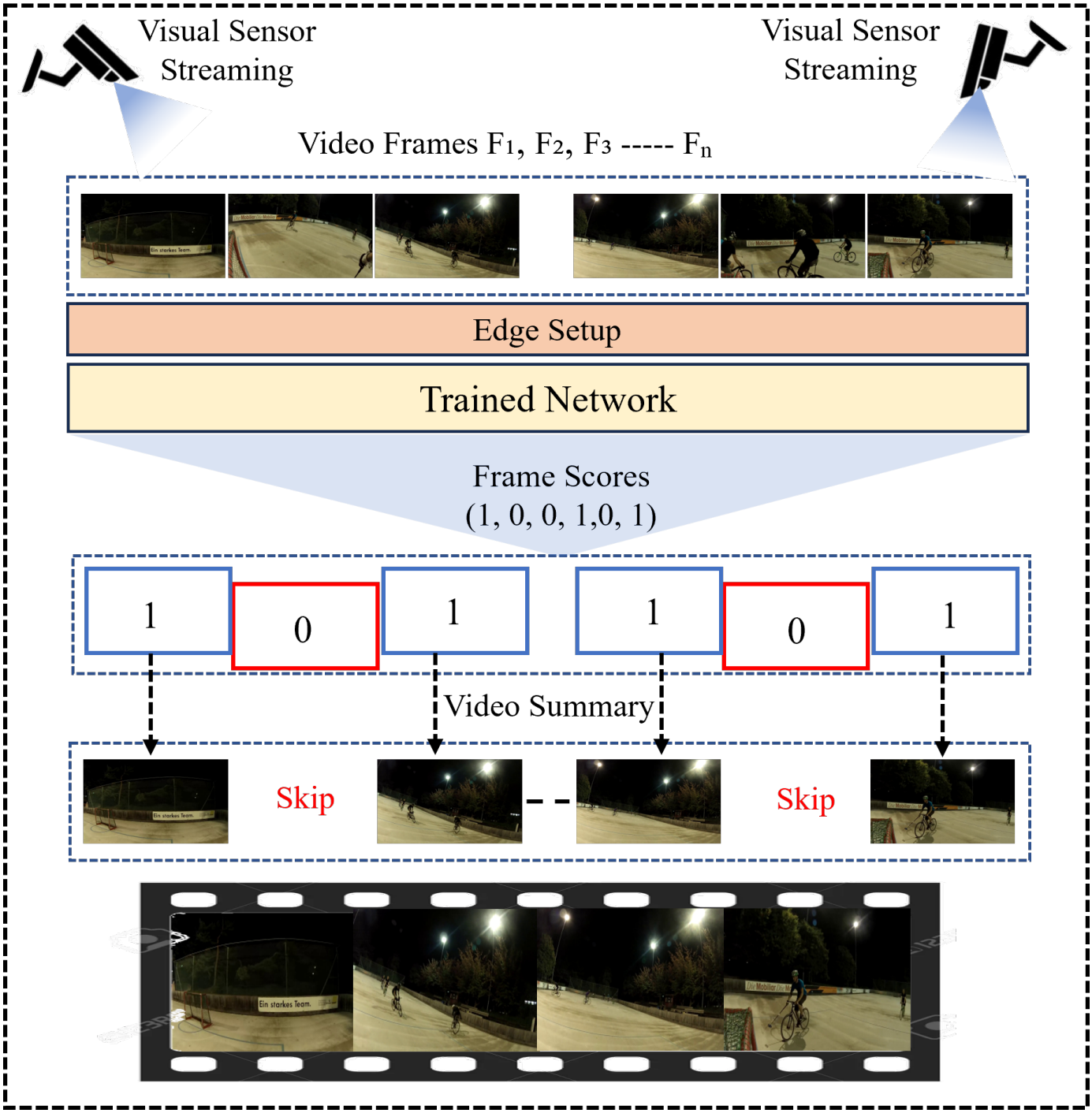

Figure 2. Suggested deployment pipeline of the proposed VS framework where incoming video frames are processed by the trained network, scores Fn, low importance frames are skipped, and only key frames are retained to form a compact video summary while preserving essential content.

subsequent multi-scale dilated convolution blocks for temporal context modeling.

3.3 Dilated Blocks for Contextual Feature Refinement

VS requires understanding temporal context across different time scales. To address this challenge, we followed the study [47] and designed a multi-scale feature extraction module comprising three parallel dilated convolution blocks (D1, D2, D3) with dilation rates of 3, 6, and 9, respectively. Unlike standard convolutions, which capture local patterns, dilated convolutions expand the receptive field without increasing computational cost or sacrificing resolution. Each dilated block applies convolutions with spacing determined by the dilation rate r. For a kernel of size

k, the effective receptive field becomes $k + (k-1)(r-1)$. The three parallel branches capture short-term, medium-term, and long-term temporal dependencies simultaneously. Mathematically, for input features $\mathbf{F}_{backbone}$, each dilated block performs:

$$\mathbf{F}_{d_i} = \text{Conv}_{d_i}(\mathbf{F}_{backbone}), \quad i \in \{1, 2, 3\} \tag{1}$$

where Conv_{d_i} denotes dilated convolution with rate $d_i \in \{3, 6, 9\}$. The multi-scale features from these three branches are concatenated to form a comprehensive temporal representation:

$$\mathbf{F}_{multi-scale} = \text{Concat}(\mathbf{F}_{d_1}, \mathbf{F}_{d_2}, \mathbf{F}_{d_3}) \tag{2}$$

This hierarchical sensing approach enables the network to capture both fine-grained frame transitions and long-range video dynamics, which are essential for identifying salient segments in videos.

3.4 Dual-Pathway Efficient Channel Attention

Channel attention mechanisms adaptively emphasize informative channels while suppressing less relevant ones [48, 49]. We followed the study [50] to propose the DECA module that leverages complementary statistical information from both average and maximum pooling operations. The architecture of the DECA module is illustrated in Figure 3. Unlike conventional channel attention that relies solely on average pooling, DECA exploits dual pathways to capture richer channel-wise statistics.

The DECA module processes input features $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$ through two parallel branches. The first branch applies Global Average Pooling (GAP) to aggregate spatial information:

$$\mathbf{F}_{gap} = \text{GAP}(\mathbf{F}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{F}_{:,:,i,j} \tag{3}$$

The second branch employs Global Max Pooling (GMP) to capture the most prominent activation:

$$\mathbf{F}_{gmp} = \text{GMP}(\mathbf{F}) = \max_{i,j} \mathbf{F}_{:,:,i,j} \tag{4}$$

Both pooled representations $\mathbf{F}_{gap}, \mathbf{F}_{gmp} \in \mathbb{R}^{B\times C\times 1\times 1}$ are independently processed through shared 1D convolutions with kernel size $k = 3$ to capture channel interdependencies. The shared weight mechanism reduces parameters while learning consistent channel relationships:

$$\mathbf{F}'_{gap} = \text{Conv1D}_{k=3}(\mathbf{F}_{gap}) \tag{5}$$

$$\mathbf{F}'_{gmp} = \text{Conv1D}_{k=3}(\mathbf{F}_{gmp}) \tag{6}$$

The processed features from both pathways are fused through element-wise addition and transformed using sigmoid activation to generate channel attention weights:

$$\mathbf{M}_{channel} = \sigma(\mathbf{F}'_{gap} + \mathbf{F}'_{gmp}) \tag{7}$$

where σ denotes the sigmoid function. Finally, the channel attention is applied to the input features through element-wise multiplication with a residual connection:

$$\mathbf{F}_{DECA} = \mathbf{F} \odot \mathbf{M}_{channel} + \mathbf{F} \tag{8}$$

where $\odot$ represents element-wise multiplication, this residual design ensures stable gradient flow during training.

3.5 Optimized Spatial Attention

While channel attention recalibrates feature maps across channels, spatial attention identifies important spatial locations within feature maps. Traditional spatial attention mechanisms employ large convolution kernels (e.g., 7×7) to capture spatial context, resulting in substantial parameter overhead. To address this limitation, we propose the OSA module that decomposes the standard 7×7 convolution into efficient 3×3 operations [51]. Given the channel-attended features $\mathbf{F}_{DECA}$, we first generate spatial statistics by applying average and max pooling along the channel dimension:

$$\mathbf{F}_{avg} = \text{AvgPool}_{channel}(\mathbf{F}_{DECA}) \in \mathbb{R}^{B\times 1\times H\times W} \tag{9}$$

$$\mathbf{F}_{max} = \text{MaxPool}_{channel}(\mathbf{F}_{DECA}) \in \mathbb{R}^{B\times 1\times H\times W} \tag{10}$$

These two spatial descriptors are concatenated along the channel dimension:

$$\mathbf{F}_{spatial} = \text{Concat}(\mathbf{F}_{avg}, \mathbf{F}_{max}) \in \mathbb{R}^{B\times 2\times H\times W} \tag{11}$$

Instead of applying a single 7×7 convolution (49 parameters per position), we employ a 3×3 convolution that reduces parameters by approximately 81% while maintaining spatial context modeling:

$$\mathbf{M}_{spatial} = \sigma(\text{Conv}_{3\times 3}(\mathbf{F}_{spatial})) \tag{12}$$

The spatial attention map $\mathbf{M}_{spatial} \in \mathbb{R}^{B\times 1\times H\times W}$ is applied to the input features:

$$\mathbf{F}_{OSA} = \mathbf{F}_{DECA} \odot \mathbf{M}_{spatial} \tag{13}$$

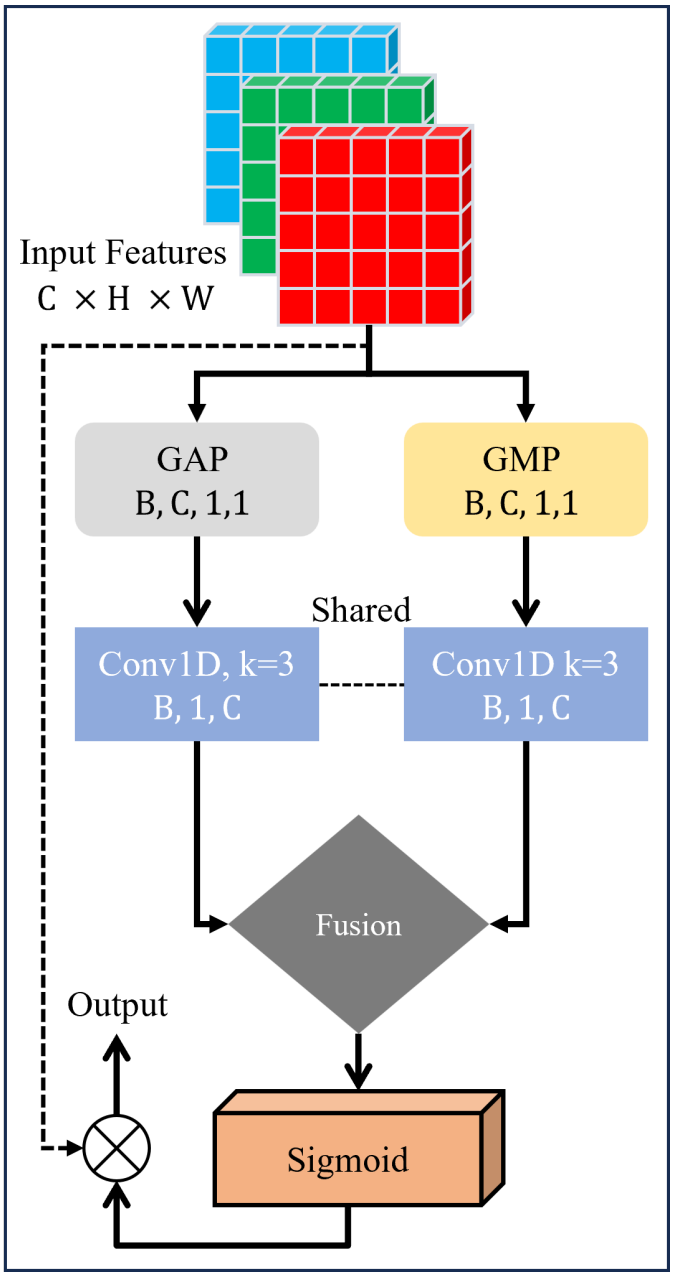

Figure 3. Dual-Pathway Efficient Channel Attention (DECA) module.

The kernel decomposition strategy significantly reduces computational complexity without sacrificing the ability to model spatial dependencies. The output features $\mathbf{F}_{OSA}$ contain both channel-wise and spatial attention refinement, producing highly discriminative representations for video frame importance prediction. These refined features are subsequently processed through global average pooling, dense layers with 1024 neurons, dropout regularization (rate = 0.5), and a final output layer to generate frame-level importance scores for VS.

3.6 Frame Importance Prediction Module

After obtaining spatially and channel-wise refined features from the DECA and OSA modules, the network requires a prediction module to transform these multidimensional representations into frame-level importance scores. The prediction module consists of sequential layers designed to aggregate spatial information, learn high-level abstractions, prevent overfitting, and generate final predictions.

The refined features $\mathbf{F}_{OSA} \in \mathbb{R}^{B\times C\times H\times W}$ first pass through a global average pooling layer that aggregates spatial information across all locations:

$$\mathbf{F}_{pooled} = \text{GAP}(\mathbf{F}_{OSA}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{F}_{OSA}[:, :, i, j] \tag{14}$$

This operation reduces the spatial dimensions to a compact feature vector $\mathbf{F}_{pooled} \in \mathbb{R}^{B\times C}$ while preserving channel-wise information. Global average pooling serves multiple purposes: it eliminates the need for flattening operations, reduces the total number of parameters, and provides spatial invariance to the final predictions. The pooled features are subsequently fed into a fully connected dense layer with 1024 neurons, which learns complex non-linear mappings between the extracted features and frame importance:

$$\mathbf{F}_{dense} = \text{ReLU}(\mathbf{W}_1\mathbf{F}_{pooled} + \mathbf{b}_1) \tag{15}$$

where $\mathbf{W}_1 \in \mathbb{R}^{1024\times C}$ and $\mathbf{b}_1 \in \mathbb{R}^{1024}$ are the weight matrix and bias vector, respectively. The Rectified Linear Unit (ReLU) activation function introduces non-linearity, enabling the network to model complex relationships between visual features and frame saliency. To mitigate overfitting and improve generalization capability, a dropout layer with a rate of 0.5 is applied after the dense layer:

$$\mathbf{F}_{dropout} = \text{Dropout}(\mathbf{F}_{dense}, p = 0.5) \tag{16}$$

During training, dropout randomly sets 50% of the neuron activations to zero, forcing the network to learn robust features that do not rely on specific neuron combinations. This regularization technique prevents co-adaptation of neurons and enhances model performance on unseen data. Finally, the dropout-regularized features pass through an output layer that generates frame importance scores:

$$\mathbf{y} = \sigma(\mathbf{W}_2\mathbf{F}_{dropout} + \mathbf{b}_2) \tag{17}$$

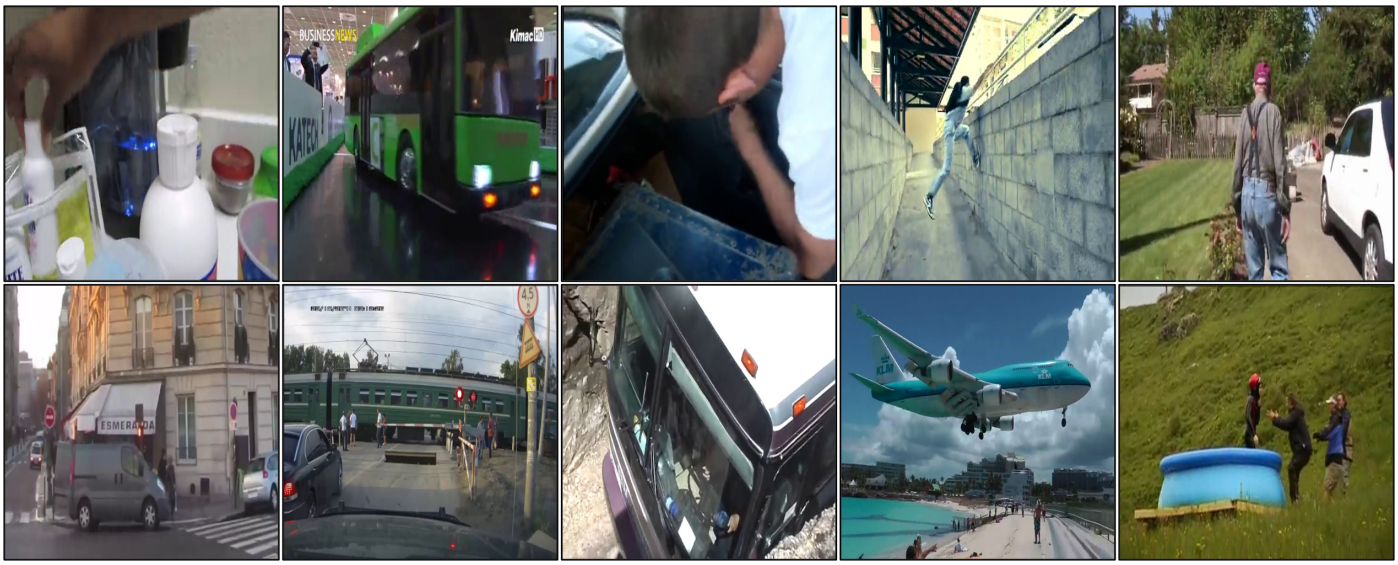

Figure 4. visual illustrations of the frames in the TVSum dataset (first two rows) and SumMe dataset (Bottom Rows).

where $\mathbf{W}_2 \in \mathbb{R}^{1\times1024}$ and $\mathbf{b}_2 \in \mathbb{R}$ are the output layer parameters, and σ represents the sigmoid activation function. The sigmoid function constrains the output to the range $[0, 1]$, which can be interpreted as the probability or importance score of each frame. Frames with higher scores indicate greater relevance to the video summary, while lower scores suggest less significant content. The predicted importance scores $\mathbf{y} \in [0, 1]^N$ for all N frames in a video are used to select key frames or segments that maximize coverage of important content while minimizing redundancy. The prediction module design balances model capacity via the 1024-neuron dense layer and generalization via dropout regularization, ensuring robust frame importance estimation across diverse video content.

4 Experimental Results

Table 1. Statistical overview of benchmark datasets used in this study.

Dataset	Content Type	Videos	Annotators
SumMe	Event recordings	25	15-18
TVSum	Professionally edited	50	20

4.1 Datasets

The proposed method is evaluated on two standard benchmarks: TVSum [52] and SumMe [37] (see Table 1 for a statistical overview). Example frames from these datasets are shown in Figure 4. These datasets provide comprehensive video collections with significant content diversity for supervised summarization tasks.

TVSum: Contains 50 videos (1-5 minutes) spanning multiple categories: educational tutorials, transportation footage, documentaries, and event recordings such as vehicle maintenance and sports. Frames are annotated with importance scores reflecting their representativeness of visual concepts, with redundant content assigned lower values to enable concise summary generation.

SumMe: Comprises 25 videos (1-6 minutes) depicting real-world events, including airplane landings, jumps, and dynamic activities. Unlike professionally edited content, most videos are raw or minimally processed, allowing for higher compression. Multiple annotators with varied backgrounds contributed GT summaries for each video.

4.2 Implementation Details and Setup

4.2.1 Ground Truth Preparation

We adopt the keyframe extraction methodology from [13] for training annotations and keyshot summaries for testing evaluation. While SumMe provides direct keyshot annotations, TVSum frame-level scores require conversion via [11]: (1) Kernel Temporal Segmentation (KTS) [36] for video segmentation, (2) interval-wise mean score computation, (3) score-based ranking, and (4) knapsack-based keyshot selection [52] under duration constraints. Training uses binary labels (0/1) to identify the most critical frames. Table 2 details annotation formats across datasets.

4.2.2 Training and Evaluation Protocol

Following established protocols [11, 13, 20], videos are uniformly sampled to 320 frames with 1×1280 channel dimensions. Training utilizes SGD (learning rate: 10^{-3}, momentum: 0.9, batch size: 8) for

Table 2. GT annotation formats and video duration statistics for benchmark datasets. Frame-level importance scores from multiple annotators are aggregated into unified representations following the methodology of [11, 20]. For each video, a single set of representative frames is generated as described in [11, 13].

Dataset	Training Annotations	Testing Annotations	Duration (min, avg, max)
TVSum	Frame-level scores	Key shot segments	32s, 146s, 324s
SumMe	Frame-level scores	Frame-level scores	83s, 235s, 647s

Table 3. Comparative performance (%) of different backbone architectures integrated with the proposed framework (Dilated Blocks + DECA + OSA) on TVSum and SumMe datasets. EfficientNetB7 achieves the highest performance, followed by ResNet-152.

Dataset	VGG-16	GoogleNet	MobileNetV2	InceptionV3	ResNet-101	ResNet-152	EfficientNetB7
TVSum	58.7	59.4	60.1	60.9	61.8	62.6	**63.5**
SumMe	49.8	50.4	50.9	51.5	52.1	52.7	**53.3**

Table 4. Ablation study showing F1-score performance (%) with progressive integration of network components on TVSum and SumMe datasets. The complete architecture achieves the best results, highlighted in **bold**.

Dataset	Backbone	Dilated Blocks	DECA	OSA	F1-Score (%)
TVSum	✓	×	×	×	60.6
	✓	✓	×	×	61.2
	✓	✓	✓	×	61.7
	✓	✓	×	✓	62.4
	✓	✓	✓	✓	**63.5**
SumMe	✓	×	×	×	50.8
	✓	✓	×	×	51.8
	✓	✓	✓	×	52.1
	✓	✓	×	✓	52.7
	✓	✓	✓	✓	**53.3**

100 epochs with frozen pre-trained EfficientNetB7 backbone and end-to-end optimization of subsequent layers. Inference predictions are resized to the original video length via nearest-neighbor interpolation, then converted to keyshots using KTS segmentation [36] and knapsack algorithm [52] with 15% length constraint. Experiments run on NVIDIA RTX 4090 (24GB).

4.2.3 *Evaluation Metrics*

We employ keyshot-based evaluation metrics following [19, 20]. Given generated summary G_S and GT H_L for video $\tilde{V}$, precision and recall are:

$$p_i = \frac{|G_S \cap H_L|}{|G_S|}, \quad r_i = \frac{|G_S \cap H_L|}{|H_L|} \tag{18}$$

The F1-score is computed as:

$$F_1 = \frac{2 \cdot p_i \cdot r_i}{p_i + r_i} \times 100\% \tag{19}$$

where higher values reflect superior summarization performance, consistent with protocols in [11, 20].

4.3 Ablation Studies

Comprehensive ablation experiments were conducted to validate the effectiveness of each component in the proposed network. Tables 3 and 4 present the results of backbone architecture comparison and progressive module integration analysis.

4.3.1 *Backbone Architecture Evaluation*

Table 3 compares seven backbone architectures integrated with the complete framework (Dilated Blocks + DECA + OSA). EfficientNetB7 achieves superior performance with F1-scores of 63.5% and 53.3% on TVSum and SumMe datasets, respectively. Its compound-scaling methodology optimally balances network depth, width, and resolution, while squeeze-and-excitation blocks enable efficient channel-wise recalibration that synergizes with the proposed attention mechanisms. ResNet-152 achieves the second-best performance with scores of 62.6% and 52.7%, attributed to its deep residual learning. ResNet-101 achieves 61.8% and 52.1%, confirming the benefits of residual connections for gradient flow. InceptionV3 obtains 60.9% and 51.5% through multi-scale feature extraction via inception

Table 5. Comparative analysis of VS methods on SumMe and TVSum datasets ranked by F1-scores (%). Superior performance is indicated in **bold**.

Technique	Features	TVSum		SumMe	
		F-1	Rank	F-1	Rank
ESSV [53]	AlexNet	–	–	40.9	14
GSF [28]	VGGNet-16	52.7	14	43.1	13
SWVT TVSum [52]	HOG+GIST+SIFT	50.0	15	–	–
VsAR [57]	GoogleNet	56.3	13	40.1	15
SeqDPP [54]	GoogleNet	58.4	10	44.3	11
TTH-RNN [39]	GoogleNet	60.2	9	44.3	11
KS-CVS [55]	CapsNet	58.0	12	46.0	10
FCSN [13]	GoogleNet	58.4	10	48.8	9
SBNT [58]	GoogleNet	61.0	7	50.7	8
LMHA (M_1) [19]	GoogleNet	61.0	7	51.1	7
LMHAD (M_2) [19]	GoogleNet	61.5	4	51.4	6
DPFN [59]	DPT-ViT	62.4	3	51.9	4
SHTVS (M_2) [20]	GoogleNet	61.4	5	52.3	3
CAVS-Net [60]	EfficientNetB1	61.4	5	51.7	5
HAVSNet [60]	EfficientNetB1	63.1	2	52.9	2
Proposed Network	**EfficientNetB7**	**63.5**	**1**	**53.3**	**1**

modules. Lightweight architectures show moderate performance: MobileNetV2 (60.1%, 50.9%) prioritizes efficiency over accuracy, while GoogleNet (59.4%, 50.4%) and VGG-16 (58.7%, 49.8%) exhibit lower scores due to their simpler architectural designs. These results empirically validate EfficientNetB7 as the optimal backbone for VS tasks.

4.3.2 *Progressive Module Integration Analysis*

Table 4 presents progressive ablation results evaluating individual component contributions. The baseline EfficientNetB7 backbone achieves 60.6% on TVSum and 50.8% on SumMe. Incorporating multi-scale dilated convolution blocks yields improvements of 0.6% and 1.0%, respectively, demonstrating effective temporal context modeling across different scales. Adding DECA to the dilated blocks provides additional gains of 0.5% on TVSum and 0.3% on SumMe, validating dual-pathway channel attention for feature recalibration. The OSA module shows stronger individual contribution, improving performance by 1.2% on TVSum and 0.9% on SumMe over dilated blocks alone. This superior performance stems from the optimized 3×3 kernel decomposition, which efficiently captures spatial dependencies while reducing the number of parameters by 81%. The complete architecture combining all components achieves the highest F1-scores of 63.5% and 53.3%, representing overall improvements of 2.9% and 2.5% over the baseline backbone. Notably, OSA demonstrates greater impact than DECA (1.2% vs 0.5% on TVSum, 0.9% vs 0.3% on SumMe), suggesting that spatial attention is more critical for predicting video frame importance. However, the synergistic combination yields an additional 1.1% improvement on TVSum and 0.6% on SumMe beyond individual attention modules, confirming that dual attention mechanisms complement each other effectively. These comprehensive ablation results validate that each proposed component contributes meaningfully to achieving state-of-the-art VS performance.

4.4 Comparison and Analysis

Table 5 presents a comprehensive performance comparison of the proposed method against 15 state-of-the-art VS techniques on TVSum and SumMe datasets. The proposed method achieves superior performance with F1-scores of 63.5% on TVSum and 53.3% on SumMe, securing rank 1 on both datasets. This represents improvements of 1.1% and 1.4% over DPFN, and 2.0% and 1.9% over LMHAD, which rank among the top-performing methods. The performance superiority stems from three key innovations: (1) EfficientNetB7 backbone provides more discriminative spatial features compared to GoogleNet or EfficientNetB1 used in competing methods, (2) multi-scale dilated convolution blocks with rates of 3, 6, and 9 effectively capture temporal dependencies across different time scales, and (3) the synergistic combination of DECA and OSA modules

enables comprehensive channel-wise and spatial feature refinement.

Analyzing performance trends reveals essential insights. Traditional handcrafted feature-based methods (SWVT TVSum, GSF) achieve F1-scores below 53% on TVSum, while early deep learning approaches (VsAR, SeqDPP, TTH-RNN) range from 56-60%. Recent attention-based methods (SBNT, LMHA, LMHAD) achieve 61-61.5%, and the transformer-based DPFN reaches 62.4%. The proposed method surpasses all competitors, demonstrating that carefully designed convolutional architectures with multi-scale temporal modeling and dual attention can outperform more complex transformer-based approaches. The consistent rank 1 performance across both datasets validates the robustness of the proposed approach. TVSum contains 50 professionally edited videos, while SumMe comprises 25 unedited event recordings with higher content diversity. Competing methods show inconsistent rankings as DPFN ranks 3 and 4, while SHTVS ranks 5 and 3 on TVSum and SumMe, respectively. In contrast, the proposed method maintains top performance on both datasets, confirming effective generalization across different video types and editing styles.

5 Conclusion

This paper presented a multi-scale sensing network for video summarization that integrates three key innovations to address limitations in existing approaches. The proposed framework combines multi-scale dilated convolution blocks, context modeling, a Dual-Pathway Efficient Channel Attention (DECA) module that exploits complementary pooling statistics for channel recalibration, and an Optimized Spatial Attention (OSA) module that achieves 81% parameter reduction through 7×7 to 3×3 kernel decomposition. Experimental results on the TVSum and SumMe datasets validate the effectiveness of the proposed approach, achieving state-of-the-art F1 Scores of 63.5% and 53.3%, respectively. The framework demonstrates improvements of 2.9% and 2.5% over the EfficientNetB7 baseline and surpasses 15 competing methods, including transformer-based approaches. Ablation studies confirm that each component contributes meaningfully, with the synergistic combination of multi-scale temporal modeling and dual attention mechanisms yielding optimal performance. Backbone architecture evaluation validates EfficientNetB7 as the superior feature extractor compared to six alternative architectures. Future research directions include extending the framework for real-time processing of longer videos, incorporating multi-modal information (audio and text) for enhanced summarization, and exploring domain adaptation techniques for improved cross-domain generalization. Moreover, we aim to investigate transformer-convolutional hybrid architectures to advance video summarization performance further while maintaining computational efficiency.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Bleu, N. (2022). 25 Latest Facebook Video Statistics, Facts, And Trends (2022). Retrieved from https://bloggingwizard.com/facebook-video-statistics/ (accessed on 29 December 2025).

[2] Ajmal, M., Naseer, M., Ahmad, F., & Saleem, A. (2017, December). Human motion trajectory analysis based video summarization. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 550-555). IEEE. [CrossRef]

[3] Wang, Z., Liu, Z., Li, G., Wang, Y., Zhang, T., Xu, L., & Wang, J. (2021). Spatio-temporal self-attention network for video saliency prediction. *IEEE Transactions on Multimedia, 25*, 1161-1174. [CrossRef]

[4] Li, H., Ke, Q., Gong, M., & Drummond, T. (2023, January). Progressive Video Summarization via Multimodal Self-supervised Learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 5573-5582). IEEE. [CrossRef]

[5] Liang, G., Lv, Y., Li, S., Wang, X., & Zhang, Y. (2022). Video summarization with a dual-path attentive network. *Neurocomputing, 467*, 1-9. [CrossRef]

[6] Zhang, Y., Zhang, T., Wang, S., & Yu, P. (2025). An efficient perceptual video compression scheme based on deep learning-assisted video saliency and

just noticeable distortion. *Engineering Applications of Artificial Intelligence, 141*, 109806. [CrossRef]

[7] Elhamifar, E., Sapiro, G., & Sastry, S. S. (2015). Dissimilarity-based sparse subset selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(11), 2182–2197. [CrossRef]

[8] Zhou, K., Qiao, Y., & Xiang, T. (2018). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1). [CrossRef]

[9] Yuan, L., Tay, F. E. H., Li, P., & Feng, J. (2019). Unsupervised video summarization with cycle-consistent adversarial LSTM networks. *IEEE Transactions on Multimedia, 22*(10), 2711–2722. [CrossRef]

[10] Muhammad, K., Hussain, T., & Baik, S. W. (2020). Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognition Letters, 130*, 370-375. [CrossRef]

[11] Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016, September). Video summarization with long short-term memory. In *European conference on computer vision* (pp. 766-782). Cham: Springer International Publishing. [CrossRef]

[12] Zhao, B., Li, X., & Lu, X. (2017). Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM International Conference on Multimedia* (pp. 863–871). [CrossRef]

[13] Rochan, M., Ye, L., & Wang, Y. (2018). Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 347–363). [CrossRef]

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

[15] Karthik, R., Hariharan, M., Anand, S., Mathikshara, P., Johnson, A., & Menaka, R. (2020). Attention embedded residual CNN for disease detection in tomato leaves. *Applied Soft Computing, 86*, 105933. [CrossRef]

[16] Ji, Z., Xiong, K., Pang, Y., & Li, X. (2019). Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(6), 1709–1717. [CrossRef]

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

[18] Ji, Z., Zhao, Y., Pang, Y., Li, X., & Han, J. (2020). Deep attentive video summarization with distribution consistency learning. *IEEE Transactions on Neural Networks and Learning Systems, 32*(4), 1765–1775. [CrossRef]

[19] Zhu, W., Lu, J., Han, Y., & Zhou, J. (2022). Learning multiscale hierarchical attention for video summarization. *Pattern Recognition, 122*, 108312. [CrossRef]

[20] An, Y., & Zhao, S. (2022). 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). [CrossRef]

[21] Ngo, C. W., Ma, Y. F., & Zhang, H. J. (2005). Video summarization and scene detection by graph modeling. *IEEE Transactions on circuits and systems for video technology, 15*(2), 296-305. [CrossRef]

[22] Zhou, H., Sadka, A. H., Swash, M. R., Azizi, J., & Sadiq, U. A. (2010). Feature extraction and clustering for dynamic video summarisation. *Neurocomputing, 73*(10–12), 1718–1729. [CrossRef]

[23] Lee, Y. J., Ghosh, J., & Grauman, K. (2012, June). Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 1346-1353). IEEE. [CrossRef]

[24] Mundur, P., Rao, Y., & Yesha, Y. (2006). Keyframe-based video summarization using delaunay clustering. *International journal on digital libraries, 6*(2), 219-232. [CrossRef]

[25] De Avila, S. E. F., Lopes, A. P. B., da Luz Jr, A., & de Albuquerque Araújo, A. (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters, 32*(1), 56–68. [CrossRef]

[26] Chu, W. S., Song, Y., & Jaimes, A. (2015, June). Video co-summarization: Video summarization by visual co-occurrence. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3584-3592). IEEE. [CrossRef]

[27] Mei, S., Guan, G., Wang, Z., Wan, S., He, M., & Feng, D. D. (2015). Video summarization via minimum sparse reconstruction. *Pattern Recognition, 48*(2), 522–533. [CrossRef]

[28] Li, X., Zhao, B., & Lu, X. (2017). A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing, 26*(8), 3652–3664. [CrossRef]

[29] Mei, S., Ma, M., Wan, S., Hou, J., Wang, Z., & Feng, D. D. (2020). Patch based video summarization with block sparse representation. *IEEE Transactions on Multimedia, 23*, 732–747. [CrossRef]

[30] Muhammad, K., Hussain, T., Tanveer, M., Sannino, G., & De Albuquerque, V. H. C. (2019). Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. *IEEE Internet of Things Journal, 7*(5), 4455–4463. [CrossRef]

[31] Fei, M., Jiang, W., & Mao, W. (2017). Memorable and rich video summarization. *Journal of Visual*

Communication and Image Representation, 42, 207–217. [CrossRef]

[32] Muhammad, K., Hussain, T., Del Ser, J., Palade, V., & De Albuquerque, V. H. C. (2019). DeepReS: A deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. *IEEE Transactions on Industrial Informatics, 16*(9), 5938–5947. [CrossRef]

[33] Mohan, J., & Nair, M. S. (2019). Static video summarization using sparse autoencoders. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1–8). IEEE. [CrossRef]

[34] Zhong, R., Wang, R., Zou, Y., Hong, Z., & Hu, M. (2021). Graph attention networks adjusted bi-LSTM for video summarization. *IEEE Signal Processing Letters, 28*, 663–667. [CrossRef]

[35] Sahu, A., & Chowdhury, A. S. (2021). First person video summarization using different graph representations. *Pattern Recognition Letters, 146*, 185–192. [CrossRef]

[36] Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014, September). Category-specific video summarization. In *European conference on computer vision* (pp. 540-555). Cham: Springer International Publishing. [CrossRef]

[37] Gygli, M., Grabner, H., Riemenschneider, H., & Van Gool, L. (2014, September). Creating summaries from user videos. In *European conference on computer vision* (pp. 505-520). Cham: Springer International Publishing. [CrossRef]

[38] Zhang, K., Grauman, K., & Sha, F. (2018, September). Retrospective Encoders for Video Summarization. In *European Conference on Computer Vision* (pp. 391-408). [CrossRef]

[39] Zhao, B., Li, X., & Lu, X. (2020). TTH-RNN: Tensor-train hierarchical recurrent neural network for video summarization. *IEEE Transactions on Industrial Electronics, 68*(4), 3629–3637. [CrossRef]

[40] Fajtl, J., Sokeh, H. S., Argyriou, V., Monekosso, D., & Remagnino, P. (2018, December). Summarizing videos with attention. In *Asian conference on computer vision* (pp. 39-54). Cham: Springer International Publishing. [CrossRef]

[41] Zhu, W., Lu, J., Li, J., & Zhou, J. (2020). Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing, 30*, 948-962. [CrossRef]

[42] Munsif, M., Khan, N., Hussain, A., Kim, M. J., & Baik, S. W. (2024). Darkness-adaptive action recognition: Leveraging efficient tubelet slow-fast network for industrial applications. *IEEE Transactions on Industrial Informatics*. [CrossRef]

[43] Amin, S. U., Abbas, M. S., Kim, B., Jung, Y., & Seo, S. (2024). Enhanced anomaly detection in pandemic surveillance videos: An attention approach with EfficientNet-B0 and CBAM integration. *IEEE Access*. [CrossRef]

[44] Samel, K., Beedu, A., Sontakke, N., & Essa, I. (2024). Exploring Efficient Foundational Multi-modal Models for Video Summarization. *arXiv preprint arXiv:2410.07405*.

[45] Lebron Casas, L., & Koblents, E. (2018). Video summarization with LSTM and deep attention models. In *International Conference on Multimedia Modeling* (pp. 67–79). Springer. [CrossRef]

[46] Zhang, Y., Wang, S., Zhang, Y., & Yu, P. (2025). Asymmetric light-aware progressive decoding network for RGB-thermal salient object detection. *Journal of Electronic Imaging, 34*(1), 013005–013005. [CrossRef]

[47] Chen, Z., Xu, Q., Cong, R., & Huang, Q. (2020, April). Global context-aware progressive aggregation network for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 10599-10606). [CrossRef]

[48] Zhang, Q., Cong, R., Li, C., Cheng, M. M., Fang, Y., Cao, X., ... & Kwong, S. (2020). Dense attention fluid network for salient object detection in optical remote sensing images. *IEEE Transactions on Image Processing, 30*, 1305-1317. [CrossRef]

[49] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision* (pp. 3-19). Cham: Springer International Publishing. [CrossRef]

[50] Liang, B., Luo, H., Wang, J., & Shark, L. K. (2025). Multi-scale attention-edge interactive refinement network for salient object detection. *Expert Systems with Applications, 275*, 127056. [CrossRef]

[51] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[52] Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015, June). TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5179-5187). IEEE. [CrossRef]

[53] Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016, June). Summary Transfer: Exemplar-Based Subset Selection for Video Summarization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1059-1067). IEEE. [CrossRef]

[54] Li, Y., Wang, L., Yang, T., & Gong, B. (2018, September). How Local Is the Local Diversity? Reinforcing Sequential Determinantal Point Processes with Dynamic Ground Sets for Supervised Video Summarization. In *European Conference on Computer Vision* (pp. 156-174). [CrossRef]

[55] Huang, C., & Wang, H. (2019). A novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(2), 577–589. [CrossRef]

[56] Zhao, B., Li, X., & Lu, X. (2018, June). HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7405-7414). IEEE. [CrossRef]

[57] Elfeki, M., & Borji, A. (2019, January). Video summarization via actionness ranking. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 754-763). IEEE. [CrossRef]

[58] Fu, H., & Wang, H. (2021). Self-attention binary neural tree for video summarization. *Pattern recognition letters, 143*, 19-26. [CrossRef]

[59] Lin, J., Zhong, S. H., & Fares, A. (2022). Deep hierarchical LSTM networks with attention for video summarization. *Computers & Electrical Engineering, 97*, 107618. [CrossRef]

[60] Alharbi, F., Habib, S., Albattah, W., Jan, Z., Alanazi, M. D., & Islam, M. (2024). Effective video summarization using channel attention-assisted encoder–decoder framework. *Symmetry, 16*(6), 680. [CrossRef]

[61] Zhang, K., Wang, W., Lv, Z., Fan, Y., & Song, Y. (2021). Computer vision detection of foreign objects in coal processing using attention CNN. *Engineering Applications of Artificial Intelligence, 102*, 104242. [CrossRef]

Taimur Ali Khan holds a Bachelor's degree in Information Technology from the University of Agriculture, Peshawar. He is currently working as a Senior Developer and IT Consultant at Saudi Media Systems. With extensive experience in software development and IT solutions, he integrates academic knowledge with real-world applications. His research interests span machine learning, deep learning, and their applications in intelligent information systems, as well as system architecture, enterprise software development, and emerging technologies. He aims to bridge practical industry expertise with cutting-edge research to develop innovative and scalable AI-driven solutions.

Danish Ali is a Ph.D. student in Electrical and Computer Engineering at Villanova University, conducting research on advanced radar systems, artificial intelligence, and secure wireless communication technologies. His work integrates signal processing, microwave engineering, and deep learning to develop next-generation radar architectures for healthcare monitoring, autonomous vehicles, and intelligent security systems. With academic experience spanning Pakistan, China, and the United States, he has built a strong foundation in wireless communications, machine learning, embedded systems, and hardware design. His technical expertise includes mmWave Studio, GNU Radio, MATLAB, STM32, Arduino, and Verilog. Danish is passionate about advancing technologies at the intersection of AI, radar, and wireless systems to enhance global connectivity, safety, and sustainability.

Zainab Ghazanfar received her M.S. degree in Computer Science from the University of Lahore. She served as a Lecturer in the Department of Computer Science at the University of Management and Technology (UMT), Lahore. She has also held a lecturer position in the Department of Computer Science at Lahore Garrison University. Currently, she is pursuing a Ph.D. in the Department of Software and Artificial Intelligence at Gachon University, South Korea. Her research interests include deep learning, medical image processing, image recognition, and the application of artificial intelligence in medical images.

Bilal Ahmad is a researcher specializing in machine learning and data analysis. He earned his Bachelor of Science in Computer Science from the University of Malakand. His research focuses on developing practical applications of machine learning techniques and enhancing the efficiency of existing algorithms. With a strong foundation in programming and a commitment to continuous learning, Bilal is devoted to advancing the field of deep learning through innovative research and development.