sensors

MDPI

Article

A Hierarchical Spatial–Temporal Cross-Attention Scheme for Video Summarization Using Contrastive Learning

Xiaoyu Teng [1,2], Xiaolin Gui [1,2,*], Pan Xu [1,2], Jianglei Tong [1,2], Jian An [1,2], Yang Liu [3] and Huilan Jiang [4]

[1] Department of Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2] Shaanxi Province Key Laboratory of Computer Network, Xi'an Jiaotong University, Xi'an 710049, China
[3] Medical College, Northwest Minzu University, Lanzhou 730030, China
[4] ONYCOM Co., Ltd., Seoul 04519, Korea
* Correspondence: xlgui@mail.xjtu.edu.cn; Tel.: +86-157-2196-0091

Abstract: Video summarization (VS) is a widely used technique for facilitating the effective reading, fast comprehension, and effective retrieval of video content. Certain properties of the new video data, such as a lack of prominent emphasis and a fuzzy theme development border, disturb the original thinking mode based on video feature information. Moreover, it introduces new challenges to the extraction of video depth and breadth features. In addition, the diversity of user requirements creates additional complications for more accurate keyframe screening issues. To overcome these challenges, this paper proposes a hierarchical spatial–temporal cross-attention scheme for video summarization based on comparative learning. Graph attention networks (GAT) and the multi-head convolutional attention cell are used to extract local and depth features, while the GAT-adjusted bidirection ConvLSTM (DB-ConvLSTM) is used to extract global and breadth features. Furthermore, a spatial–temporal cross-attention-based ConvLSTM is developed for merging hierarchical characteristics and achieving more accurate screening in similar keyframes clusters. Verification experiments and comparative analysis demonstrate that our method outperforms state-of-the-art methods.

Keywords: video summarization; spatial–temporal features; cross-attention

Citation: Teng, X.; Gui, X.; Xu, P.; Tong, J.; An, J.; Liu, Y.; Jiang, H. A Hierarchical Spatial–Temporal Cross-Attention Scheme for Video Summarization Using Contrastive Learning. *Sensors* **2022**, *22*, 8275. https://doi.org/10.3390/s22218275

Academic Editors: Kang Ryoung Park, Sangyoun Lee and Euntai Kim

Received: 23 September 2022
Accepted: 25 October 2022
Published: 28 October 2022

1. Introduction

With the rapid development of multimedia information technology and intelligent terminal equipment, video data have emerged as a critical medium of information transmission due to its lack of reading threshold and high data-carrying capacity. However, the openness and informality of video production result in the accelerated growth of video data and several undesirable phenomena, such as widespread data redundancy [1], unclear content emphasis, and blurred video theme boundaries. Therefore, it is becoming vital to provide effective and efficient tools for the management, browsing, and retrieval of these videos. Video summarization, which uses a subset of the most informative frames to create a condensed version of the original video by removing redundant information [2–4], is an effective tool for addressing these issues.

Recent methods for video summarization rely heavily on the superior performance of deep learning, particularly in feature extraction. In addition, feature extraction is a fundamental component of video summarization algorithms that extract time series [5,6] or spatial–temporal features from video data [7,8]. From the perspective of a video feature, the performance of the video summary is dependent on the feature extraction technique. These deep learning video summarization algorithms constantly increase the depth and breadth of video feature extraction to improve its performance. The most important criterion for measuring video summarization performance is user satisfaction. User satisfaction is contingent upon their requirements for video summarization performance. Furthermore, user requirements can be translated into property constraints of algorithms [7]. These

property constraints can be categorized as representativeness [5], content coverage [8], redundancy [3], diversity [5], interestingness [9], importance [10], etc. The variety of user requirements continues to expand, while their feature definitions are more hazy. Consequently, video summarization algorithms focusing on video salient characteristics extraction are incapable of satisfying the multi-source user requirements. In addition, with the rise in popularity of video terminal equipment and the evolution of multimedia technology, hand-held and fragmented time-created videos have become the predominant sources of new created video data. Certain more prominent properties of the new video production, such as significant redundancy, a lack of strong focus, and a fuzzy theme boundary, disrupt the video summarization's initial thinking mode based on video feature information and present it with new challenges. With the evolution of video characteristics and user requirements for video summarization, the demand for keyframe accuracy screening has increased. Some traditional methods are no longer applicable, such as clustering [11].

To be more precise, existing algorithms can meet a portion of the user-centered requirements and capture good summarization performance. However, the following challenges remain: contradiction between breadth extraction of salient video characteristics and multi-source of user diversified requirements; the contradiction between depth extraction of salient video characteristics and unbounded new video productions; the contradiction between similarity frames and more accurate keyframe screening.

To address the issues mentioned above, this paper proposes a hierarchical spatial–temporal cross-attention scheme based on contrastive learning, as shown in Figure 1. The central idea of this article is to extract features and relationships between frames that account for coarse and fine-grained, global and local, depth and breadth, to fuse hierarchical features while increasing the difference between similar frames, and then screen keyframes and generate summaries by evaluating their significance. From the perspective of video feature extraction, the solution to diverse user requirements for video summary lies in the extraction of the frame's own characteristics, relationship features between frames, and relationship features between frames and the entire video. This study uses DB-ConvLSTM and multi-head attention mechanisms to design multi-conv-attention cells and joint GAT to acquire the spatial–temporal connection of keyframes to extract fine-grained spatial–temporal feature information from video frames. The GAT adjusted DB-ConvLSTM to extract the global and breadth features. In addition, to amplify the difference of similar keyframes, a spatial–temporal cross-attention-based ConvLSTM is constructed for merging hierarchical characteristics. Finally, video summarization is generated by CB-ConvLSTM through possibility. Therefore, the major contributions of this work can be summarized as follows:

1. A hierarchical spatial–temporal video feature extraction approach is developed. The purpose is to ensure as much characteristic information as possible for generating video summarization;
2. A cross-attention cell that combines the local and global features information based on DB-ConvLSTM is proposed. It seeks to emphasize the difference between related frames and achieve more accurate screening in similar keyframes clusters for video summary generation;
3. Verification experiments and comparative analysis are performed on two benchmark datasets (TVSum and SumMe) for this paper's algorithm. The results demonstrate that the proposed algorithm is extremely rational, effective, and usable.

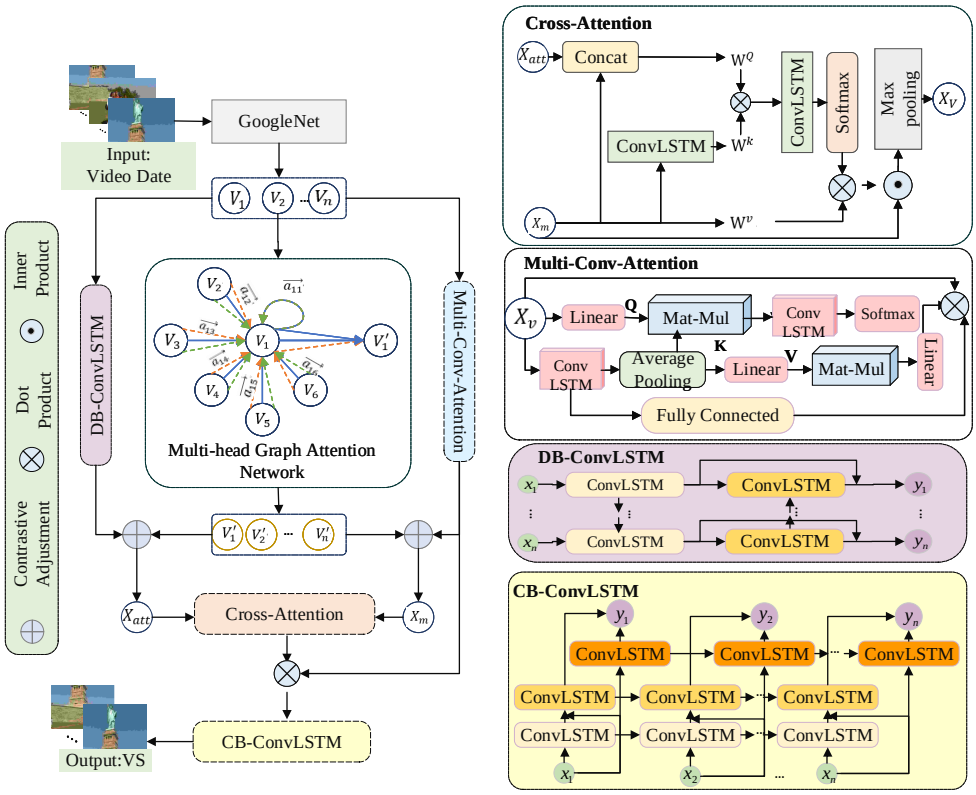

Figure 1. Overview of our approach.

2. Related Work

In this section, we briefly overview some state-of-the-art video summarization approaches and correlation techniques pertaining to our hierarchical spatial–temporal cross-attention scheme.

2.1. Video Summarization

Generally speaking, pre-processing, feature extraction, post-processing, and VS creation comprise the video summary generating procedures. The post-processing can be left out. In particular, feature extraction is the central stage of the algorithm. The initial algorithm is based on time series techniques such as vsLSTM/dppLSTM [5]. The initial method of similar keyframes decision is based on clustering [11]. Zhao et al. [12] develop an extended bidirectional LSTM (Bi-LSTM) for extracting both structure and information characteristics from video data. To acquire a more precise extraction of video features, refs. [3,13] offer a keyframe-selection strategy based on video spatial–temporal characteristics. In addition, graph neural networks are employed to implement this notion [1,6]. However, the aforementioned algorithms are all video-centric and lack comprehensive analysis of video topics and user demands. In [14], first-person (egocentric) videos-based models are proposed. A model of characterizing egocentric video frames uses a graph-based center-surround model. User requirements impose certain restrictions on the feature extraction results. The video summarization algorithms [15] are based on attention technologies, mimicking human keyframe filtering. Ji et al. [16] solve the problem of short-term contextual attention insufficiency and distribution inconsistency. Köprü [17] proposes two new architectures based on temporal attention (TA-AVSUM) and spatial attention (SA-AVSUM).

Additionally, for the video summarization algorithm, both video feature information and video frame relational are crucial [18]. Continuously improving the performance of the user-requirements-driven algorithm fundamentally necessitates more comprehensive and accurate feature extraction. This scheme is based on the concept of creating stereoscopic

modeling using spatial–temporal feature information, relationship information, and other multi-elements.

2.2. Cross Attention

Refs. [19–22] have conducted substantial study on how to more properly and completely extract video features and the relationship features between video frames. Contextual information is vital in visual understanding problems [19] and is also applicable to generating video summarization. Huang et al. [19] proposes a Criss-Cross Network (CCNet) based on attention for obtaining video information in a more effective and efficient way. Lin et al. [20] presents a universal Cross-Attention Transformer (CAT) module for accurate and efficient semantic similarity comparison in one-shot object detection. In [22], the attention mechanism is incorporated at two main levels: a self-attention module leverages global interactions between encoder features, while cross-attention in the skip connections allows fine spatial recovery in the U-Net decoder by filtering out non-semantic features. It can be seen that cross-attention has the ability to simultaneously extract the depth and breadth characteristics of video data. This study uses cross-attention to merge the hierarchical spatial–temporal characteristics, and it aims to accentuate the distinctions between video frames.

2.3. Graph Attention Networks (GATs)

Veličković et al. [23] give a novel neural network architecture that operates on graph-structured data, leveraging masked self-attentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations. GATs provide distinct weights to each neighbor based on their importance, effectively filtering the neighbors. Zhong et al. [1] build a method for video summarizing utilizing graph attention networks and Bi-LSTM. However, it does not take into account information loss throughout the confrontation process. This paper makes use of GATs to capture spatial–temporal relational attention between video frames and comparative-adjusting feature extraction.

3. Materials and Methods

Figure 1 shows an overview of our hierarchical spatial–temporal cross-attention scheme for video summarization. DB-ConvLSTM, multi-conv-attention, and multi-head attention GAT are all used for video feature extraction. The DB-ConvLSTM is employed to extract coarse-grained global spatial–temporal video characteristics. Effective fine-grained local features are extracted using multi-conv-attention networks and spatial–temporal relational feature extraction using multi-head attention GAT. This research derives hierarchical spatial–temporal feature information on the basis of cross-attention, taking into consideration both global and local characteristics and coarse-grained and fine-grained features. In particular, this scheme promotes comparative learning for acquiring local feature information for multi-conv-attention and GAT, and obtaining global feature knowledge for DB-ConvLSTM and GAT. The local and global characteristics are combined using spatial–temporal cross-attention. Finally, CB-ConvLSTM obtains the video summary.

Following the algorithm phases, this part elaborates the DB-ConvLSTM and CB-ConvLSTM, contrastive adjustment learning, and spatial–temporal cross-attention for the keyframes screening module. The contrastive adjustment learning is adjustment learning based on contrastive learning. Finally, we will introduce the loss function used in our framework.

3.1. DB-ConvLSTM and CB-ConvLSTM

Both DB-ConvLSTM and CB-ConvLSTM are founded on the technology of ConvLSTM. ConvLSTM is not only designed for extracting spatial–temporal information features but also for inferring saliency information concurrently. Then, suppose there are n frames in a video, the whole video can be written as $f = \{f_1, \cdots f_n\}$, c_t is the memory cell, f_t is

the forget gate, and i_t is the input gate. From [24], we can obtain that the ConvLSTM is defined as:

$$\begin{aligned} i_t &= \sigma(W_i^{\chi} \times X_t + W_i^{H} \times H_{t-1}) \\ f_t &= \sigma(W_f^{\chi} \times X_t + W_f^{H} \times H_{t-1}) \\ o_t &= \sigma(W_o^{\chi} \times X_t + W_o^{H} \times H_{t-1}) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_c^{\chi} \times X_t + W_c^{H} \times H_{t-1}) \\ H_t &= o_t \circ \tanh(c_t) \end{aligned} \tag{1}$$

3.1.1. DB-ConvLSTM

In the video information processing methods, the DB-ConvLSTM [25] network is suggested to extract spatial–temporal video characteristics more deeply and precisely. DB-ConvLSTM is a bidirectional two-layer architecture, one forward-oriented and one backward-oriented. The forward-oriented and backward-oriented have information interaction. The deeper layer is composed of backward-cells, its input is the output features of forward-cells, and the output is $\{Y_t\}_{t=1}^{t}$. The backward-ConvLSTM is defined as:

$$i_t^b = \sigma(W_i^{H^f} \times H_t^f + W_i^{H^b} \times H_{t+1}^b) \tag{2}$$

$$f_t^b = \sigma(W_f^{H^f} \times H_t^f + W_f^{H^b} \times H_{t+1}^b) \tag{3}$$

$$o_t^b = \sigma(W_o^{H^f} \times H_t^f + W_o^{H^b} \times H_{t+1}^b) \tag{4}$$

$$c_t^b = {f_t}^b \circ c_{t+1}^b + i_t^b \circ \tanh(W_c^{H^f} \times H_t^f + W_c^{H^b} \times H_{t+1}^b) \tag{5}$$

$$H_t^b = o_t^b \circ \tanh(c_t^b) \tag{6}$$

where W are the training parameters, denoting the learnable weights, H is the hidden state, σ is the activation function, $\times$ denotes the convolution operator, and $\circ$ denotes the hadamard product. In the VS algorithm, the DB-ConvLSTM can be written as:

$$Y_t = \tanh(W_y^{H^f} \times H_t^f + W_y^{H^b} \times H_{t-1}^b) \tag{7}$$

tanh is the activation function to normalize Y_t, and the loss function of training DB-ConvLSTM is distance minimization.

3.1.2. CB-ConvLSTM

CB-ConvLSTM is capable of extracting not only the characteristics of a single video frame but also the spatial–temporal relationships between different frames [7]. From [7], we can obtain the definition of CB-ConvLSTM, based on Equations (2)–(7), and replace the content in Equation (1) by ConvLSTM; then, CB-ConvLSTM is defined as follows:

$$H_t^f = ConvLSTM(X_t, H_{t-1}^f) \tag{8}$$

$$H_t^b = ConvLSTM(X_t \oplus H_{1,t}, H_{t+1}^b) \tag{9}$$

$\oplus$ is the operation of fusing two vectors, $H_{1,t}$ is the first hidden state, and the loss function of training CB-ConvLSTM is distance minimization. In this paper, the three layers in the network cell aim to extract and aggregate the features, and the final outputs are the possibility of whether a frame will be selected as a keyframe for video summarization.

3.2. Contrastive Adjustment Learning

Contrastive learning [26] introduces a novel idea of features derived from many perspectives: the learning algorithm does not have to concentrate on every element of the sample itself, as long as it learns enough traits to differentiate it from others. In our

study, the use of contrastive learning serves three purposes: (1) to overcome the diversity theme of video, (2) to extract elastic traffic feature information, and (3) to increase feature extraction with surface breadth and detail while enlarging the difference between video frames. As shown in Figure 2, the specific application of our strategy is to use the GATs-obtained data as the primary line and generate positive and negative pairs from the results of DB-ConvLSTM and multi-conv-attention, respectively. D_m is supposed as the results of the two sections of the comparative learning. DDPG [27] is used to train the D_m adjusted DB-ConvLSTM, which is the same as [1].

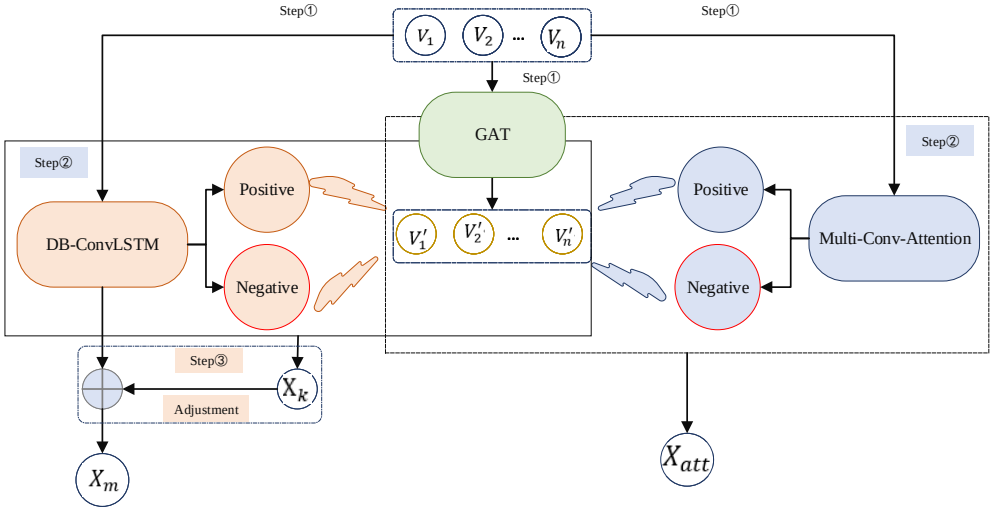

Figure 2. Step of contrastive adjustment learning.

x^+ is the positive sample, and x^- is the negative sample, S is the function for measuring the samples' similarity, and similar to [26], the rule for setting positive pairs is:

$$S(Y(x), Y(x^+)) \gg S(Y(x), Y(x^-)) \tag{10}$$

D_t is the video characteristics, which are extracted by multi-conv-attention and DB-ConvLSTM. D^+ is the keyframe sets, D^- is the non-keyframe sets, Q_j is feature mapping of the labeled data, and Q^+ is the annotated manually keyframe-sets. Then, the positive pairs include: $Y^A = \{D^+(x) \cap Q^+(x)\}$. Moreover, the loss function of a negative sample is InfoNCE in this paper, and it can be written as:

$$\mathcal{L}_{adj} = \sum_{x,x^+,x^-} \left[-log \left(\frac{e^{Y(x)^T Y(x^+)}}{e^{Y(x)^T Y(x^+)} + Y(x)^T Y(x^-)} \right) \right] \tag{11}$$

3.3. Multi-Conv-Attention and Cross-Attention

3.3.1. Multi-Conv-Attention

The temporal, spatial, and multi-element video properties are all important parts of our approach. As a consequence, a new network cell is constructed using ConvLSTM and multi-head attention. It uses convolution to improve the attention mechanism's ability to get as much video information as possible. In our multi-conv-attention cell, we first adopt a set of projections to obtain query Q. Additionally, it employs ConvLSTM and average pooling to produce two sets of projections of key K and value V, enhancing the K and V dimensions of the attention mechanism while also boosting the performance and consistency of feature information extraction. Finally, the attention is calculated as:

$$M_c(Q, K, V) = Softmax(ConvLSTM(\frac{QK^T}{\sqrt{d_k}}))V \tag{12}$$

In this scheme, we employ $n = 8$, and $d_k = d_v = d_{model}/n = 64$.

3.3.2. Cross-Attention

The Cross-Attention module is shown in Figure 3. $F(\cdot)$ and $G(\cdot)$ are projections to align dimensions using interpolation function. Then, the module performs cross-attention between X_m and X_{att}, which can be expressed as

$$q = G(X_m) \cdot W^Q \tag{13}$$

$$k = ConvLSTM(G(X_{att})) \cdot W^k \tag{14}$$

$$v = F(X_m) \cdot W^v \tag{15}$$

Finally, calculate the cross-attention using Equation (12).

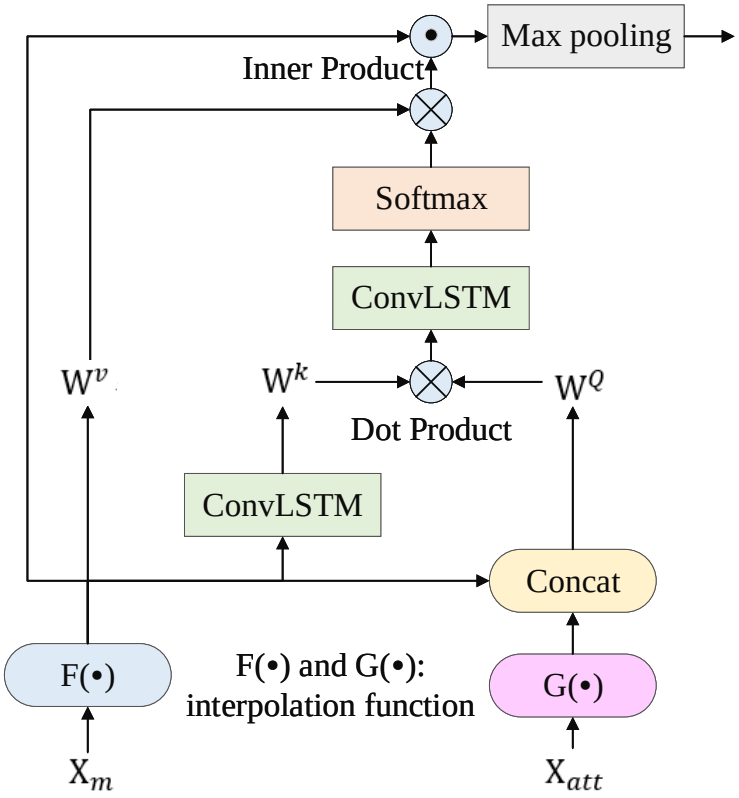

Figure 3. Spatial–temporal cross-attention cell.

3.4. Loss Function

The total loss is primarily made up of three components, and all the loss functions of these parts are based on cross-entropy. In items of supervised learning, the selection of keyframes is ultimately intended to decrease the discrepancy between predicted and background data. The cross-entropy is used to approximate the distribution of the learnt model to the background data. The lower the value is, the more similar the probability distributions of the anticipated and background data. p is the probability distribution of background data, and q is the predicted probability distribution, and the cross-entropy $H(p,q)$ is:

$$H(p,q) = \sum_{i=1}^{n} p_i log \frac{1}{q_i} = -\sum_{i=1}^{n} p_i log q_i \tag{16}$$

In our network, the softmax is used to normalize the cross-entropy, y_i is the output of network cells, $\hat{y}_i$ is the category i of background data, $\hat{y}_i \in \{0,1\}$, and the loss function is:

$$\mathcal{L} = -\frac{1}{m}\left[\sum_{i=1}^{m} \hat{y}_i log \frac{e^{z_i}}{\sum_{i=1}^{k} e^{z_k}}\right] = -\frac{1}{m}\left[\sum_{m}^{i=1} \hat{y}_i log y_i\right] \tag{17}$$

$\mathcal{L}_{mat}$ is the loss function of the model of multi-conv-attention contrastive GAT. $\mathcal{L}_{dat}$ is the loss function of the model of DB-ConvLSTM contrastive adjustment GAT. $\mathcal{L}_{cro}$ is the loss function of cross-attention. Both $\mathcal{L}_{mat}$ and $\mathcal{L}_{dat}$ are cross-entropy, as defined by Equation (17). To resolve the centralization issue and reduce the ambiguity problem in key frame filtering, we use $\mathcal{L}_{cen}$ for centralization keyframe scores:

$$\mathcal{L}_{cen} = \lambda \cdot \frac{min(\mathcal{L}_{dat}, \mathcal{L}_{mat})}{max(\mathcal{L}_{dat}, \mathcal{L}_{mat})} \tag{18}$$

In Equation (17), λ balances the function of global and local domains. Formally, the objective function $\mathcal{L}_{obj}$ is written as

$$\mathcal{L}_{tol} = \mu \cdot \mathcal{L}_{cro} + \mathcal{L}_{cen} \tag{19}$$

μ balances the loss of cross-attention and multi-conv-attention.

4. Experiments Analysis

4.1. Datasets

Each database has its focus, so before the experiment, the two databases TVsum [28] and SumMe [29] should be analyzed, and the results are shown in Table 1. Additionally, we use two other public datasets, OVP (Open Video Project) [30] and YouTube [11], to augment the training sets.

Table 1. Analysis of TVsum and SumMe dataset.

Datasets	Description
TVsum	The title-based video summarization dataset contains 50 videos of various genres (e.g., news, documentary, egocentric) and 1000 annotations of shot-level importance scores (20 user annotations per video). The duration varies from 2 to 10 min.
SumMe	The SumMe dataset consists of 25 videos, each annotated with at least 15 human annotated summaries. The duration of videos varies from 1.5 to 6.5 min.

4.2. Evaluation Metrics

To facilitate a comparison study of the experimental influence on current research findings, the Precision, Recall, and F-score are used as measurement standards, similar to the literature [3]. S is the video summarization generated by the algorithm, G denotes the ground user-marked ground truth, and the following definitions apply to Precision, Recall, and F-score:

$$Precision = \frac{|S \cup G|}{|S|} \tag{20}$$

$$Recall = \frac{|S \cup G|}{|G|} \tag{21}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{22}$$

As shown in [31], randomly generated video summaries may achieve equivalent performance when using the F-score measure. To avoid this problem, we evaluated our method as seen in Table 2. Under our comparison method, the comparing parameter is F-score. Furthermore, to be more precise, these datasets are randomly split into different training and testing sets five times, and the final measure is produced by averaging the five results.

Table 2. Datasets setting used for evaluation (C: Canonical; A: Augmented; T: Transfer).

Datasets	Setting	Training Phase	Testing Phase
TVSum	C	80% TVSum	The rest 20% of TVSum
	A	80% TVSum+SumMe+ OVP+YouTube	The rest 20% of TVSum
	T	SumMe+OVP+YouTube	TVSum
SumMe	C	80% SumMe	The rest 20% of SumMe
	A	TVSum+80% SumMe+ OVP+YouTube	The rest 20% of SumMe
	T	TVSum+OVP+YouTube	SumMe

4.3. Experimental Environment and Parameters Settings

The deep learning platform for operating our approach is Pytorch. The hidden states are with dimensionality of 256 for ConvLSTM, and other parameters settings are as follows: similar to other algorithms, we use the pool5 layer of GoogleNet to extract the visual features for each video frame. The number of ConvLSTMs hidden layers is 256, the learning rate initialized is $le-5$, the batch size is 5, the kernel size is set as (5,1), and the maximum training epoch is set as 100. Furthermore, considering that the training epochs are critical to summarization performance, after increasing for five epochs continuously, their influence on the validation set is plotted in Figures 4–6. The horizontal coordinate is the training epochs, and the vertical coordinates are the values the of F-scores, Recall, and Precision.

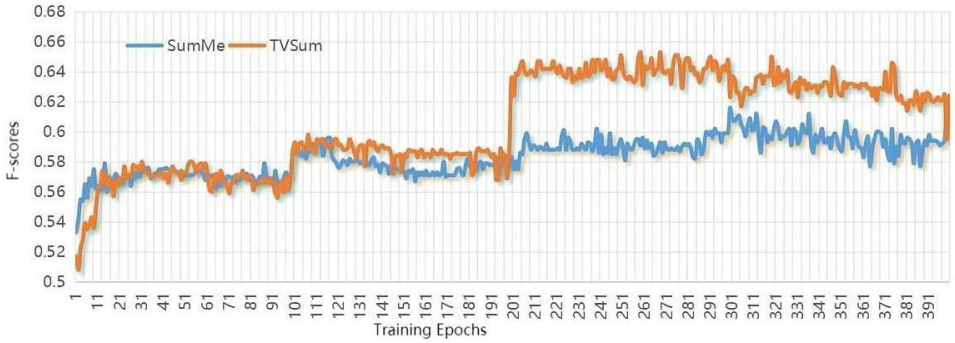

Figure 4. Plots show the influence of training epochs on the value of F-scores.

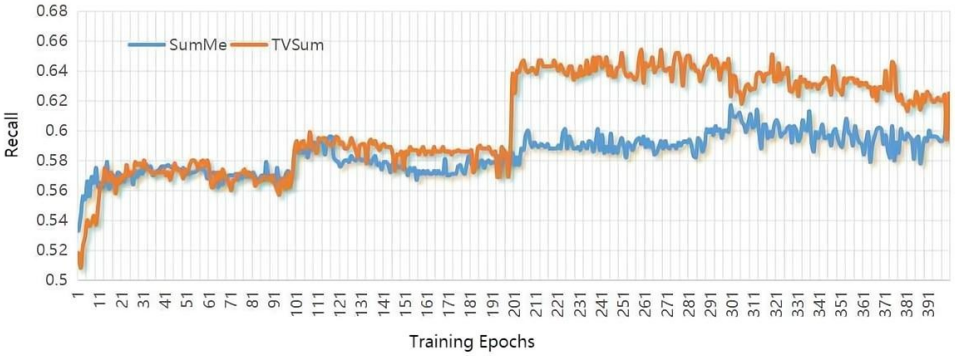

Figure 5. Plots show the influence of training epochs on the value of Recall.

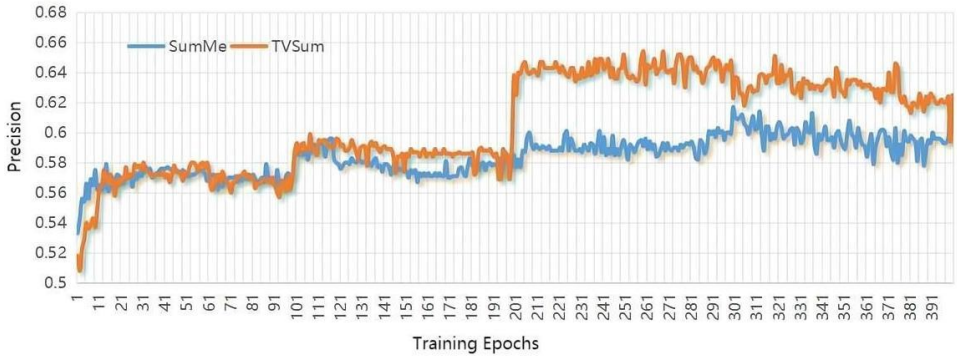

Figure 6. Plots show the influence of training epochs on the value of Precision.

4.4. Comparative Analysis of Schemes

This section verifies the feasibility and effectiveness of the proposed strategy through two ways: one is validation of the algorithm itself, and the other one is comparative analysis with state-of-the-art video summarization approaches.

4.4.1. Self-Verification

Before comparing the scheme to other state-of-the-art algorithms, it is vital to validate the scheme's performance itself. Table 3 and Figure 7 show the results of evaluating the performances of our methods on the SumMe and TVSum datasets. From the perspective of result stability, the test variation curves of C, T, and A on SumMe and TVSum datasets are shown in Figures 8 and 9. The horizontal coordinate of both figures is training epochs. Figure 7 gives an example of generating video summarization on the SumMe and TVSum datasets by our approach; the yellow lines show the annotation importance scores of ground truth summarization marked by the user, and the blue lines show the prediction score of our method. We clearly observe that our models achieve very competitive results against state-of-the-art methods.

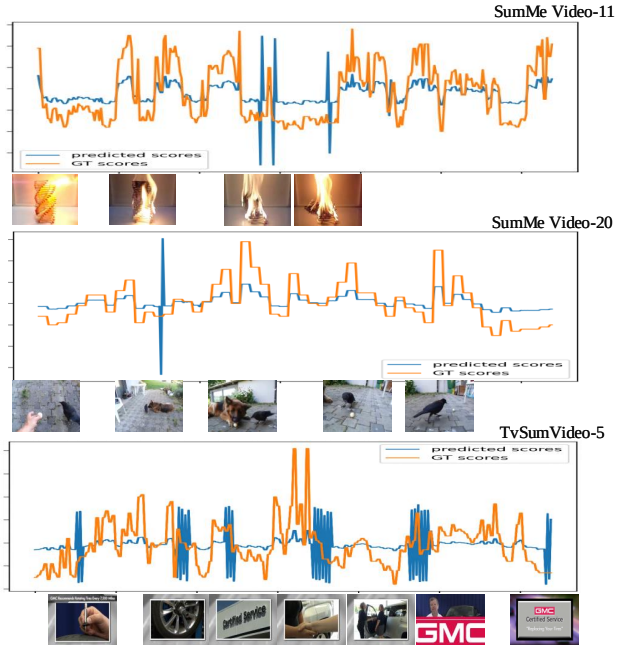

Figure 7. An example of generating video summarization on SumMe and TVSum datasets; the first two are samples from SumMe datasets and the last one is from TVSum datasets. The yellow lines show the annotation importance scores of ground truth summarization marked by the user, and the blue lines show the prediction score of our method.

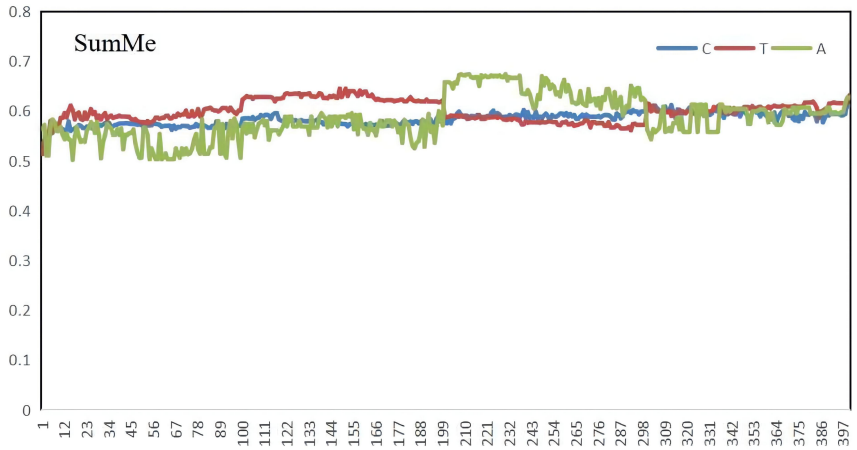

Figure 8. The A, C, and T results of SumMe.

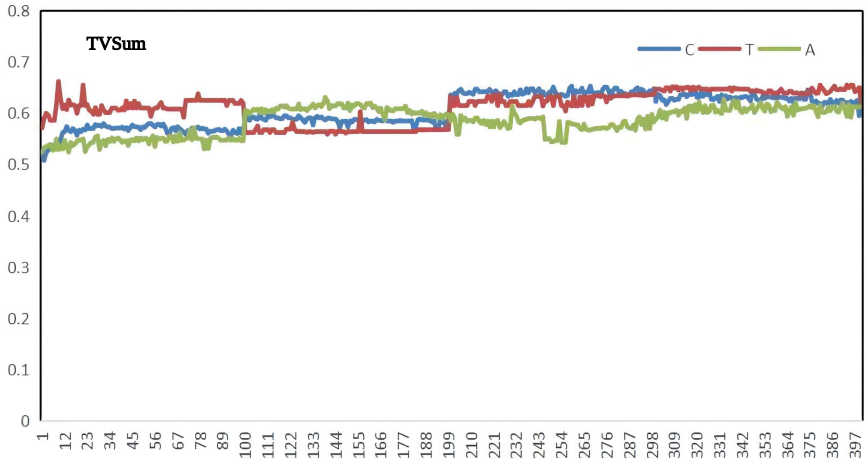

Figure 9. The A, C, and T results of TvSum.

Table 3. Performance analysis of self-verification (F-scores).

Data Sets	**TVSum**			**SumMe**		
Metric	**C (%)**	**A (%)**	**T (%)**	**C (%)**	**A (%)**	**T (%)**
MAX	65.3	67.4	66.2	61.6	63.1	64.5
MIN	50.8	50.2	55.9	53.3	52.4	51.3
AVERAGE	60.57	58.62	61.26	58.4	58.4	60.01

4.4.2. Comparative Analysis with Relative Approaches

The primary components of our algorithm consist of the attention mechanism, ConvLSTM, and GATs. In this section, we compared our approach with some state-of-the-art video summarization methods on SumMe and TvSum. Comparison methods can be classified into three categories: based on "LSTM+", based on "Attention+", and based on GATs methods.

(1) Comparison With "Bi-LSTM+" Methods

Due to the few research results on the summary algorithm based on ConvLSTM, this section compares our scheme to the Bi-LSTM based algorithms. Some classic algorithms are compared, as shown in Table 4.

Table 4. Performance analysis of methods based on "Bi-LSTM+".

Data Sets	TVSum			SumMe		
Metric	**C (%)**	**A (%)**	**T (%)**	**C (%)**	**A (%)**	**T (%)**
vsLSTM [5]	54.2	57.9	56.9	37.6	41.6	40.7
dppLSTM [5]	54.7	59.6	58.7	38.6	42.9	41.8
H-RNN [12]	57.9	61.9	–	42.1	43.8	–
HAS-RNN [32]	58.7	59.8	–	42.3	42.1	–
DHAVS [33]	60.8	61.2	57.5	45.6	46.5	43.5
Ours	65.3	67.4	66.2	58.4	58.4	60.01

H-RNN [12] and HAS-RNN [32] are based on hierarchical architecture. According to the findings of the comparison, we observe that our method outperforms state-of-the-art video summarization methods on both datasets.

(2) Comparison With "Attention+" Methods

Since the scheme in this paper involves not only the combination of ConvLSTM and attention but also the graph neural network, we will analyze it separately. The results of comparison with "Attention+" methods are shown in Table 5. SABTNet [15] is based on attention and a binary neural tree. Liang et al. [34] proposes a video summarization method based on dual-path attention, while Zhu et al. [35] is based on hierarchical attention. Table 5 demonstrates that the cross-attention method has clear benefits over the SumMe database.

Table 5. Performance analysis of methods based on "Attention+".

Data Sets	TVSum			SumMe		
Metric	**C (%)**	**A (%)**	**T (%)**	**C (%)**	**A (%)**	**T (%)**
M-AVS [36]	61.0	61.8	–	44.4	41.6	–
SABTNet [15]	61.0	–	–	51.7	–	–
[34]	61.58	61.2	58.9	51.7	52.1	44.1
[35]	61.5	62.8	56.7	51.1	52.1	45.6
Interp-SUM [2]	59.14	–	–	47.7	–	–
3DST-UNet [3]	58.3	58.9	56.1	47.4	49.9	47.9
Ours	65.3	67.4	66.2	58.4	58.4	60.01

(3) Comparison With "Graph Attention+" Methods

The extraction of spatial–temporal characteristics and frame–relationship features is facilitated by a graph neural network. Table 6 shows the results of comparing our method with some "Graph Attention+" video summarization methods including RSGN [13], GCAN [37], Bi-GAT [1] and SumGraph [38]. From the experimental results in Table 6, our method outperforms other approaches, which are based on "Graph Attention+".

Table 6. Performance analysis of methods based on "Graph Attention+".

Data Sets	TvSum (F-Score %)	SumMe (F-Score %)
RSGN [13]	60.1	45.0
GCAN [37]	60.1	53.0
Bi-GAT [1]	59.6	51.7
SumGraph [38]	63.9	51.4
Ours	65.36	58.48

4.4.3. Comparison Results

Following the comparison tests outlined above, it can be seen that the proposed method has certain advantages over existing approaches, most notably in the SumMe

database set. Specifically, the hierarchical spatial–temporal cross-attention scheme in this research enhances the algorithm's stability, scalability, and other performance characteristics.

5. Conclusions

This paper proposes a hierarchical spatial–temporal cross-attention scheme for video summarization using contrastive learning. The scheme solves the contradictions of diversification user requirements, depth and breadth of features extraction and new creation videos. The hierarchical architecture is divided primarily into depth and breadth feature extraction and spatial–temporal cross-attention feature merging. This paper extracts local and depth features using a graph attention network and multi-head attention mechanism, and it extracts global and breadth features using a GAT adjusted DB-ConvLSTM. Furthermore, merging hierarchical characteristics via spatial–temporal cross-attention cells is used for more precise keyframe screening. Finally, video summarization is generated by CB-ConvLSTM. In practice, results from the TVSum and SumMe datasets indicate that the proposed algorithm is highly rational, effective, and usable. Nevertheless, the analysis of similarity keyframe screening is still insufficiently detailed.

Author Contributions: Conceptualization, X.T., X.G., P.X., J.T., J.A., Y.L. and H.J.; methodology, X.T. and X.G.; software, P.X. and J.T.; validation, X.T., X.G. and P.X.; formal analysis, J.A.; investigation, Y.L.; resources, H.J.; writing—original draft preparation, X.T.; writing—review and editing, X.T. and P.X.; project administration, X.G. and J.A.; funding acquisition, X.G. and J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was partially supported by the National Key Research and Development Project Under Grant (2018YFB1800304), National Natural Science Foundation of China (61472316), Fundamental Research Funds for the Central Universities (xzy012020112).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhong, R.; Wang, R.; Zou, Y.; Hong, Z.; Hu, M. Graph attention networks adjusted bi-LSTM for video summarization. *IEEE Signal Proc. Lett.* **2021**, *28*, 663–667. [CrossRef]
2. Yoon, U.-N.; Hong, M.-D.; Jo, G.-S. Interp-SUM: Unsupervised Video Summarization with Piecewise Linear Interpolation. *Sensors* **2021**, *21*, 4562. [CrossRef] [PubMed]
3. Liu, T.; Meng, Q.; Huang, J.-J.; Vlontzos, A.; Rueckert, D.; Kainz, B. Video summarization through reinforcement learning with a 3D spatio-temporal u-net. *IEEE Trans. Image Proc.* **2022**, *31*, 1573–1586. [CrossRef] [PubMed]
4. Li, W.; Pan, G.; Wang, C.; Xing, Z.; Han, Z. From coarse to fine: Hierarchical structure-aware video summarization. *ACM Trans. Mult. Comput. Commun. Appl. TOMM* **2022**, *18*, 1–16. [CrossRef]
5. Zhang, K.; Chao, W.-L.; Sha, F.; Grauman, K. Video summarization with long short-term memory. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 766–782.
6. Zhao, B.; Li, H.; Lu, X.; Li, X. Reconstructive sequence-graph network for video summarization. *IEEE Trans. Patt. Anal. Mach. Intell.* **2021**, *44*, 2793–2801. [CrossRef] [PubMed]
7. Teng, X.; Gui, X.; Xu, P. A Multi-Flexible Video Summarization Scheme Using Property-Constraint Decision Tree. *Neurocomputing* **2022**, *506*, 406–417. [CrossRef]
8. Ji, Z.; Zhang, Y.; Pang, Y.; Li, X.; Pan, J. Multi-video summarization with query-dependent weighted archetypal analysis. *Neurocomputing* **2019**, *332*, 406–416. [CrossRef]
9. Rafiq, M.; Rafiq, G.; Agyeman, R.; Choi, G.S.; Jin, S.-I. Scene classification for sports video summarization using transfer learning. *Sensors* **2020**, *20*, 1702. [CrossRef]
10. Zhu, W.; Han, Y.; Lu, J.; Zhou, J. Relational Reasoning Over Spatial-Temporal Graphs for Video Summarization. *IEEE Trans. Image Proc.* **2022**, *31*, 3017–3031. [CrossRef]
11. De Avila, S.E.F.; Lopes, A.P.B.; da Luz, A., Jr.; de Albuquerque Araújo, A. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Patt. Recognit. Lett.* **2011**, *32*, 56–68. [CrossRef]
12. Zhao, B.; Li, X.; Lu, X. Hierarchical recurrent neural network for video summarization. In Proceedings of the 25th ACM International Conference on Multimedia, New York, NY, USA, 23–27 October 2017; pp. 863–871.

13. An, Y.; Zhao, S. A Video Summarization Method Using Temporal Interest Detection and Key Frame Prediction. *arXiv* **2021**, arXiv:2109.12581.
14. Sahu, A.; Chowdhury, A.S. First person video summarization using different graph representations. *Patt. Recognit. Lett.* **2021**, *146*, 185–192. [CrossRef]
15. Fu, H.; Wang, H. Self-attention binary neural tree for video summarization. *Patt. Recognit. Lett.* **2021**, *143*, 19–26. [CrossRef]
16. Ji, Z.; Zhao, Y.; Pang, Y.; Li, X.; Han, J. Deep attentive video summarization with distribution consistency learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 1765–1775. [CrossRef]
17. Köprü, B.; Erzin, E. Use of Affective Visual Information for Summarization of Human-Centric Videos. *arXiv* **2021**, arXiv:2107.03783.
18. Mi, L.; Chen, Z. Hierarchical Graph Attention Network for Visual Relationship Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
19. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
20. Lin, W.; Deng, Y.; Gao, Y.; Wang, N.; Zhou, J.; Liu, L.; Zhang, L.; Wang, P. CAT: Cross-Attention Transformer for One-Shot Object Detection. *arXiv* **2021**, arXiv:2104.14984.
21. Sanabria, M.; Precioso, F.; Menguy, T. Hierarchical multimodal attention for deep video summarization. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 7977–7984.
22. Petit, O.; Thome, N.; Rambour, C.; Themyr, L.; Collins, T.; Soler, L. U-net transformer: Self and cross attention for medical image segmentation. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Strasbourg, France, 27 September 2021; pp. 267–276.
23. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
24. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal ON, Canada, 7–12 December 2015; Volume 28.
25. Song, H.; Wang, W.; Zhao, S.; Shen, J.; Lam, K.-M. Pyramid dilated deeper convlstm for video salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 715–731.
26. Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv* **2021**, arXiv:2104.08821.
27. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the International Conference on Learning Representations 2016, San Juan, Puerto Rico, 2–4 May 2016; pp. 1–15.
28. Song, Y.; Vallmitjana, J.; Stent, A.; Jaimes, A. Tvsum: Summarizing web videos using titles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5179–5187.
29. Gygli, M.; Grabner, H.; Riemenschneider, H.; Van Gool, L. *Creating Summaries from User Videos*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2014; pp. 505–520.
30. Open Video Project. Available online: https://open-video.org/ (accessed on 22 September 2022).
31. Otani, M.; Nakashima, Y.; Rahtu, E.; Heikkila, J. Rethinking the evaluation of video summaries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7596–7604.
32. Zhao, B.; Li, X.; Lu, X. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7405–7414.
33. Lin, J.; Zhong, S.-h.; Fares, A. Deep hierarchical LSTM networks with attention for video summarization. *Comput. Electr. Eng.* **2022**, *97*, 107618. [CrossRef]
34. Liang, G.; Lv, Y.; Li, S.; Wang, X.; Zhang, Y. Video summarization with a dual-path attentive network. *Neurocomputing* **2022**, *467*, 1–9. [CrossRef]
35. Zhu, W.; Lu, J.; Han, Y.; Zhou, J. Learning multiscale hierarchical attention for video summarization. *Patt. Recognit.* **2022**, *122*, 108–312. [CrossRef]
36. Ji, Z.; Zhao, Y.; Pang, Y.; Li, X.; Han, J. Video summarization with attention-based encoder–decoder networks. *IEEE Trans. Circ. Syst. Video Technol.* **2019**, *30*, 1709–1717. [CrossRef]
37. Li, P.; Tang, C.; Xu, X.Video summarization with a graph convolutional attention network. *Front. Inform. Technol. Electr. Eng.* **2021**, *22*, 902–913. [CrossRef]
38. Park, J.; Lee, J.; Kim, I.-J.; Sohn, K. Sumgraph: Video summarization via recursive graph modeling. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 647–663.