Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Video summarization via knowledge-aware multimodal deep networks

Jiehang Xie [a,b], Xuanbai Chen [c], Sicheng Zhao [d], Shao-Ping Lu [a,b,*]

[a] *TKLNDST, Tianjin, 300350, China*
[b] *College of Computer Science, Nankai University, Tianjin, 300350, China*
[c] *Robotics Institute, Carnegie Mellon University, United States of America*
[d] *BNRist, Tsinghua University, Beijing 100084, China*

ARTICLE INFO

Keywords:
Video summarization
Multimodal information
External knowledge

ABSTRACT

Video summarization has unprecedented importance in facilitating the rapid browsing, retrieval, and comprehension of large numbers of videos. Benefiting from possessing rich prior knowledge of the raw video and the capability to filter less crucial frames by employing multimodal information, humans can condense a lengthy video into a compact and reasonable video summary. However, existing automated video summarization approaches struggle to determine which shots in a video are significant concurrently and robustly, which is detrimental to the generation of high-quality summaries. To improve the quality of video summaries further, drawing inspiration from human abilities, we propose a novel video summarization approach based on a knowledge-aware multimodal network (KAMN). In particular, we present a knowledge-based encoder to obtain the corresponding representation for each frame. This representation is composed of captured descriptive content and affections, which are retrieved from large-scale external knowledge bases. Owing to these knowledge bases, rich implicit knowledge is provided to better understand the viewed video. Moreover, to integrate the visual, audio, and implicit knowledge features more effectively and to identify valuable information across different modalities further, we design a fusion module to learn these multimodal feature relationships more thoroughly. KAMN operates in both unsupervised and supervised training modes. Objective quantitative experiments and subjective user studies were conducted using four publicly available datasets. The results verified the effectiveness of the proposed modules and demonstrated the superior performance yielded by our framework.

1. Introduction

Owing to the remarkable surge in online videos and the desire of viewers to browse, manage, and comprehend content rapidly, the task of summarizing videos has become increasingly prevalent in both academia and industry [1,2]. The objective of video summarization is to select key frames and reduce the original video, ensuring that the frames convey crucial and represent the original untrimmed video. Summarized videos help viewers to digest, browse, and retrieve ever-growing video collections. Consequently, video summarization can be an indispensable tool in many applications [3], such as surveillance, documentaries, and recommender systems.

Over the past several years, numerous video summarization algorithms have been proposed. In the early years, only handcrafted features were involved in key-shot selection [4]. For example, some studies segmented the video into multiple clips of the same length, extracted a shot with a dialogue or action from each clip, and finally, assembled all extracted shots into a video summary [5]. Unfortunately, these efforts are both time consuming and labour intensive, and significant biases may be introduced into the generated summaries. More recently, with the drastic development of computational resources, solutions using deep-learning technology have emerged as practical options for developing more effective video summarization approaches [6]. Several studies have relied on the strong encoding ability of neural networks to extract visual features to characterize video content [7,8]. Although these visual features represent the content of a video to a certain extent, they do not completely capture valuable information. Several studies have attempted to introduce advanced techniques from various domains, such as target detection, to assist models in capturing the long-term temporal dependencies of video features [9]. However, most of them still focus solely on visual modalities by extracting either shallow or deep visual features to represent the video content, which may lead to suboptimal results. Particularly, compared to human potential, which can easily determine which parts of a video are essential, these video summarization methods remain primitive, reflecting their unreliability and lack of robustness.

* Corresponding author at: College of Computer Science, Nankai University, Tianjin, 300350, China.
E-mail addresses: jiehangxie@mail.nankai.edu.cn (J. Xie), chenxuanbai@126.com (X. Chen), schzhao@gmail.com (S. Zhao), slu@nankai.edu.cn (S.-P. Lu).

https://doi.org/10.1016/j.knosys.2024.111670
Received 13 June 2023; Received in revised form 19 February 2024; Accepted 18 March 2024
Available online 20 March 2024

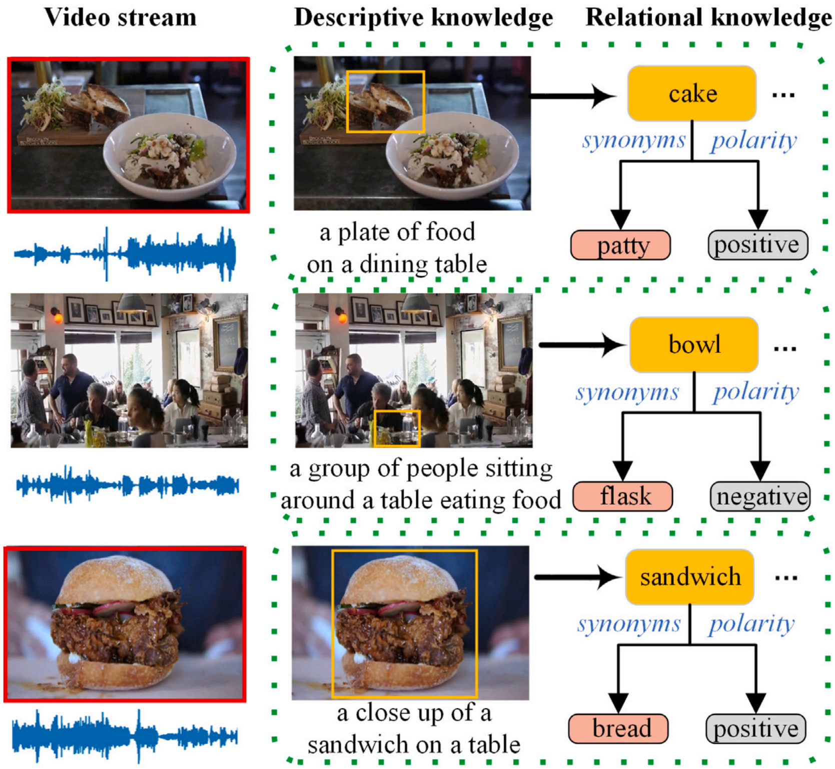


Fig. 1. An example of a video stream conveying content through audio-visual information in conjunction with implicit knowledge. Compared the image of a cluttered room, positive affections can be easily generated when humans observe the shots with food. We add the solid red rectangle to highlight the crucial shot.

We believe that the inherent sensitivity of humans to subtle variations in multimodal information contributes to the obvious discrepancy between machine and human summarization [10]. More importantly, numerous studies have demonstrated that the integration of multimodal information offers significant potential for enhancing the model performance [11]. For instance, cognitive psychology researchers have verified that human cognition is a cross-media information interaction process, indicating that cognition is an outcome of audiovisual information stimulation [12]. Compared to the attention gained by the visual modality, information from other modalities is typically neglected by existing methods, which is detrimental to generating satisfactory summaries.

In addition, humans are talented at encoding the implicit knowledge of visual content. Concretely, implicit knowledge comprises descriptive and relational knowledge. The former denotes the explanatory and semantic information of the visual content, whereas the latter encompasses the relationships between entities and affections [13]. With the assistance of implicit knowledge, humans can easily determine the parts of a video that are beneficial for a final summary [14].

It can be observed from Fig. 1 that the video stream shares immersive visual content along with the audio. Furthermore, the implicit knowledge in each shot can induce humans to generate various associations and personal affections, which can assist them in better comprehending the video content and further arouse interest in specific shots. These behaviours contribute to the capacity of humans to evaluate the importance of shots accurately. In contrast, annotating the importance score for shots is difficult for machines because the importance of a shot is an abstract concept. Moreover, the inter-dependency among frames is highly complex and inconsistent [15]. Hence, the generalizability of data-driven learning methods is poor, particularly for unseen data, which makes the classifier incapable of explicit training. In conclusion, the non-negligible knowledge gap between humans and machines and the under-utilization of audiovisual information hinder the progress of this task.

Intuitively, the inclusion of implicit semantic knowledge and multimodal features that can enhance the video representation can increase the probability of achieving a high-quality summary. However, two challenges must be overcome to achieve this. The first is acquiring implicit knowledge from video frames and incorporating it into the training process. The second is that instead of continuously increasing with the growing number of modalities, the performance of the model may suffer from exacerbation [16]. Therefore, our concern is how to fuse the acquired implicit semantic knowledge and multimodal features effectively to achieve improved performance.

The proposed KAMN, which serves as a knowledge-aware multimodal network, addresses these problems. Specifically, to capture implicit knowledge from video frames comprehensively, we attempt to generate descriptive knowledge and retrieve relational knowledge from two large-scale knowledge bases. We include the video and audio modalities by leveraging corresponding audio and visual encoders to obtain features from the input video. In addition, we propose a fusion module that considers the contextual features within a modality (intra-level) and across modalities (inter-level) to enhance the use of multimodal information.

This work is an extension of the previous conference version [17] and includes the following enhancements: First, we conduct a more comprehensive review to compare the proposed method with existing methods. Second, we elaborate on the design and setup of the proposed method in further detail. Third, we extend the vanilla supervised model to an unsupervised variant by jointly considering the diversity and representativeness of the generated summaries and achieve competitive performance. Fourth, we present canonical experimental results on an additional popular benchmark dataset, which also achieves state-of-the-art performance. Finally, we include more rigorous experiments with a systematic analysis to illustrate the superiority of the proposed method more effectively. These include additional baselines, ablation studies, user studies, statistical analyses, one-to-one transfer experiments, and runtime analyses. In Section 2, we review video summarization methods and studies corresponding to implicit knowledge. Section 3 presents the technical details of the different components of KAMN. Section 4 discusses the experimental settings and presents various quantitative and qualitative results. Finally, Section 5 concludes the paper by briefly outlining the key contributions.

Our contributions can be summarized as follows:

- We present a novel knowledge encoder that provides a high-level representation for video summarization and automatically generates the description and relational knowledge of a video based on an external knowledge base.
- We propose a multimodal fusion module for the joint modelling of intra-and inter-level feature relationships, effectively exploiting the complementary features across modalities for video summarization.
- We present extensive experimental results to demonstrate that the proposed method outperforms recently developed supervised baseline methods. Furthermore, our unsupervised variation achieves competitive results compared with classical approaches.

2. Related work

In this section, we introduce previous works on unimodal- and multimodal-based methods, and related works that have incorporated implicit knowledge. We also emphasize the specific improvements introduced by the proposed method.

Unimodal-based Methods. In the early years, researchers usually manually sampled keyframes or video segments, which is both laborious and time consuming [4]. With the emergence of deep neural networks, many attempts have been made to automate video summarization using unimodal-based methods [18,19]. The typical unimodal-based approach first acquires visual features according to an encoder and then builds a predictor under the supervision of the ground truth [20]. For example, Zhu et al. [9] provided dense sampling for the temporal interest proposals that accommodate interest variations in length and then extracts corresponding features for interest proposal location regression and importance prediction. Saquil et al. attempted to provide personalized video summaries using a multiple pairwise ranker-based video summarization model and predefined categorical user labels [21]. Li et al. proposed a joint video modelling method based on a hierarchical transformer for co-summarization by considering semantic dependencies across videos [22]. Jiang et al. [23] proposed a method for transferring samples from a correlated video moment localization task to a video summarization task.

Owing to the absence of annotated data in video summarization, researchers have proposed various unsupervised video summarization approaches to identify the most important characteristics of a good video summary [24,25]. Apostolidis et al. [26] constructed an unsupervised framework by introducing an attention mechanism into a deterministic autoencoder. Lan et al. [2] considered video summarization as a sequence-to-sequence learning problem and constructed a summarizer based on a pointer network, which yielded significant achievements. Jung et al. [27] employed a self-attention mechanism with relative position representation to perform this task without human annotation. Although unimodal data provide valuable information for video summarization, relying solely on unimodal data may yield less comprehensive and potentially biased video summaries, thereby limiting the ability of unimodal-based methods to adapt to a wide range of video content.

Multimodal-based Methods. The academic community has recently recognized the significance of multimodal learning [28]. This recognition is driven by the fact that multimedia content is often unstructured and cannot be effectively accessed or comprehended using a single modality [29]. In particular, visual, audio, and text, as the three main modalities of video content, have attracted significant research interest in the development of various models [30], resulting in outstanding achievements.

However, earlier approaches simply connected features from different modalities [31,32], and failed to simulate cross-modal interactions effectively. Therefore, subsequent studies proposed more effective multimodal fusion strategies. Zhao et al. [33] employed long short-term memory (LSTM) to integrate audiovisual information, exploring the interactions to enhance the summarization performance. Wu et al. [34] constructed a dual attention-matching module that addressed the audiovisual event localization problem and achieved remarkable success. This module enhances the modelling of event information and obtains local temporal information using a crosscheck mechanism. Wei et al. [35] developed a series of guidelines that incorporate visual and audio cues, such as audio segmentation, visual diversity, and image quality, to generate video summaries. Rhevanth et al. [36] used both visual similarity and audio signal features to detect key shots and then constitute a summary. Zhao et al. [37] proposed a hierarchical multimodal Transformer to model audiovisual information jointly, and developed a multimodal fusion mechanism to capture the global dependency and multi-hop relationships among shots. Nevertheless, compared to the ability of humans to discern important shots based on implicit information within a video frame, these methods still exhibit certain limitations. In addition, most studies are entirely driven by annotated data and do not address the common issue of handling missing annotations in practice.

Some studies have considered the text modality. For example, Wei et al. [38] generated a video summary by providing additional considerations for annotated text descriptions. Lei et al. [28] proposed a sequential multiple-instance learning model that included scripts and subtitles as input. Li et al. [39] collected textual information, including video categories and search queries, from the Internet and constructed video–text pairs to train a video summarization model in a self-supervised manner. Chen et al. [40] proposed a novel highlight extraction method that uses textual data from streamers and viewers instead of visual information to obtain highlights. However, these approaches rarely consider the visual, audio, and textual modalities simultaneously, and their performance scores reveal the limitation of not incorporating the visual and audio modalities. In addition, the process of acquiring textual data for these approaches is labour intensive.

Utilizing Implicit Knowledge. Several studies have used large-scale knowledge bases, such as ConceptNet [41] and DBpedia [42], for action recognition [43], image classification [44], and visual question-answering tasks [45], with significant progress. For example, Wu et al. [45] presented an approach for handling visual question answering by incorporating high-level semantic concepts under the association between attribute words and the image content. Zhang et al. [46] constructed a knowledge incorporation model that uses external knowledge bases to perform visual reasoning. Zhang et al. [47] described a novel model based on a knowledge graph that provides more common sense to assist in answering questions. Qi et al. [48] introduced an architecture that uses emotional knowledge for highlight detection. Wang et al. [49] proposed a deep knowledge-aware network to predict the click-through rate of news. This network enriches the news content by associating each word with related entities in the knowledge base. Considering the human manner of perceiving visual content, a promising direction is to incorporate the descriptive and relational knowledge of frames into video summarization. However, to the best of our knowledge, few studies have exploited this knowledge for video summarization.

This paper. To overcome the limitations of previous studies, we propose a video summarization method based on knowledge-aware and multimodal networks. Different from conventional multimodal-based methods, our approach incorporates not only visual and auditory modalities, but also textual modalities consisting of descriptive and relational knowledge. Particularly, the text data in our method are collected automatically. Moreover, our multimodal fusion module differs significantly from existing works [30] in that we introduce more dynamic interactions among modalities, thereby enriching the video

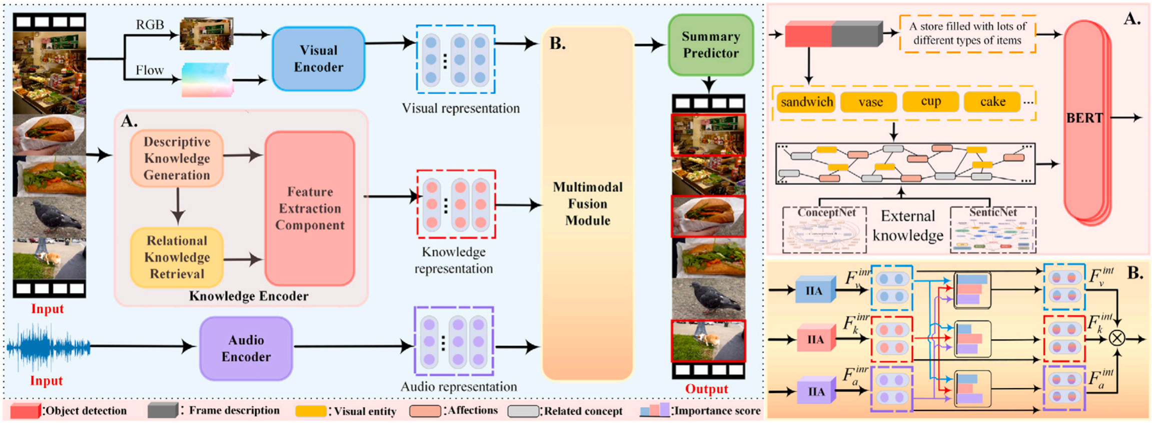


Fig. 2. The framework of proposed KAMN. The upper left part is a simplified work flow. A, B are knowledge encoder and multimodal fusion module, respectively. The right part is the detailed processes corresponding to the knowledge encoder and multimodal fusion module.

summarization with a deeper, context-aware analysis. In addition, we improve the generalization ability of the proposed method by jointly considering the diversity and representativeness of the generated summaries, which enables the model to achieve competitive performance even in the absence of annotation.

3. Methodology

In this section, we introduce the proposed KAMN for video summarization in detail. The overall architecture is shown in Fig. 2, where the input is a sequence of video frames and the corresponding audio, and the output is a shortened video that preserves the important parts of the original video. KAMN consists of a knowledge encoder, a visual encoder, an audio encoder, a multimodal fusion module, and a summary predictor. Specifically, the knowledge encoder aims to capture implicit knowledge from videos comprehensively while providing a higher-level semantic representation to guide the summarization process. A video encoder, which is a two-stream architecture that performs spatiotemporal convolutions to obtain a visual representation, is applied to capture the long-term dependencies among dynamic consecutive frames and generate a more aggregated feature representation. To exploit the valuable information in the input audio, the audio encoder was developed based on VGGish [50] to acquire audio representations. Our multimodal fusion module effectively explores the interactions within and across the modalities, fusing knowledge and visual and audio representations into final features. Finally, the obtained features are taken as input to the summary predictor and generated video summary. The motivations for and designs of each component are described in detail in the following subsections.

3.1. Knowledge encoder

Motivation. The goal of the knowledge encoder is to capture implicit semantic knowledge from the input video to imitate the behaviour whereby humans tend to select meaningful entities and content from the frames that they see when watching a video. A common intuitive approach is to perceive entities from frames that are meaningful to humans and generate text descriptions for the frames, which has been demonstrated as effective in the field of visual question answering, but has not been explored in video summarization [45]. This inspired us to add a descriptive knowledge-generation component to obtain such information. However, such a straightforward approach ignores the fact that whether humans are interested in a particular video part is largely affected by external knowledge, and this neglect may be detrimental to video summarization. We suggest simulating the influence of entities in human decision-making with the assistance of external knowledge bases and thus designing the relational knowledge retrieval component to combine entities with related concepts and affections. In addition, the obtained implicit knowledge is expected to be aggregated into explicit semantic representations for subsequent training. Therefore, we propose a feature extraction component to perform this operation.

Method. Our framework uses a pre-trained advanced object detection approach [51] and a frame description method to form a descriptive knowledge generation component to output a group of visual entities VE and a text description TD for each video frame VF. This object detection approach is based on an end-to-end teacher–student model, which learns to generate pseudo labels for unlabelled data using a teacher model for the training of the student model, and improves the student model using a semi-supervised loss. This approach can be benefit from the development of object detection research and can be replaced by newer or better alternatives. In our work, we leverage this object detection approach to detect the u non-repetitive and most conspicuous visual entities from frames, which can be formally represented as $VE = \{ve_1, ve_2, \ldots, ve_u\}$, where ve_i indicates the ith entity. Since several studies on visual question answering have demonstrated that the five most significant objects in an image are sufficient to provide the main content of the image [52], we set u to five to reduce data redundancy and the computational complexity. When the number of detected visual entities is zero, it is denoted as 'none'. The frame description approach converts visual content into human-readable text based on an architecture that connects a convolutional neural network to a recurrent neural network. Specifically, we train an Inception-v3 [53] network as an encoder on the ImageNet dataset [54], and then train an LSTM decoder on the COCO dataset [55] to learn the mapping from frames to sentences directly. To infer a sentence that a human would most likely describe based on a given frame, we utilize a beam search algorithm and set the beam size to one to retain only the best results.

The subsequent relational knowledge retrieval component consists of an affection retrieval method and a concept retrieval method, thereby retrieving affections and concepts that are related to entities from external knowledge bases. For the affection retrieval method, the external information source that we use is SenticNet [56], which associates specific video content with the affection response of humans. SenticNet is a large-scale affective knowledge base created from numerous individual desires or normative evaluations of events, objects, and behaviours in the social environment. The SenticNet knowledge base contains more than 200,000 natural language nodes, where each node represents an entity or concept in the real world. This knowledge base leverages the denotative content (i.e. semantically related concepts) and connotative information (i.e. emotion categorization values expressed in terms of four affective dimensions) associated with nodes, to provide a set of affections associated with natural language concepts

Algorithm 1: The workflow of knowledge encoder

```
   Input: Given a video frame $VF$
   Output: Knowledge representation $R_k$
1  $VE$ ← object detection approach % extract a group of visual entities
     $VE$ from $VF$, $VE = \{ve_1, ve_2, ..., ve_u\}$;
2  $TD$ ← frame description approach;
3  for each $ve$ in $VE$ do
4      if $ve$ can be retrieved from SenticNet then
5          get corresponding affection tuple $s$ using $ve$;
6      end
7      else
8          retrieve a group of related concepts $ce$ for $ve$ from
             ConceptNet, $ce = \{c_1, c_2, ..., c_o\}$ in a descending order based
             on the confidence score;
9          while $i \leftarrow 1$ to $o$ and the affection $s$ is not yet retrieved for $ve$
             do
10             do Step 5 using $c_i$;
11         end
12     end
13 end
14 Obtain the affections $S$ of all visual entities, $S = \{s_1, s_2, ..., s_u\}$;
15 Calculate knowledge representation $R_k$ according to Eq. (1);
16 final;
```

that can be used directly in affective computing, such as semantics, sentics, and polarity [57].

Specifically, the affection retrieval method takes a set of visual entities VE of given frame VF as input, and then outputs the affections S based on SenticNet, where $S = \{s_1, s_2, \ldots, s_u\}$, s_i is a affection tuple of ve_i, which is a mixture of both discrete and continuous affections, involving *emotion*, *polarity*, *sensitivity* and *intensity*. If a entity ve can be retrieved from SenticNet, we employ the corresponding s.

However, SenticNet and similar knowledge bases suffer from many problems, including sparsity of information and inconsistency of its representations. An inspection of SenticNet reveals that although the database is large, some entities are still not contained therein. To remedy this, we retrieve a related concept group for each entity according to the concept retrieval method and then use this concept group to find the relevant affections. The entity-related source used for the concept retrieval method is ConceptNet [41], which is a multilingual knowledge base containing over eight million nodes. Each node in ConceptNet is a word and phrase that is commonly used in natural language, with labelled relational edges connecting them to convey commonsense knowledge. For example, the relationship between the nodes “bread” and “food” can be bread “RelatedTo” food; this relationship can be formulated as a triplet (bread; RelatedTo; food). Each triplet in ConceptNet is correlated with a weight score ranging from 1 to 10. ConceptNet contains 32 million triplets. To minimize interference from irrelevant noise data, we perform the following steps: 1. We use English resources containing approximately 1.5 million nodes. 2. We select triplets with symmetric relations (e.g., Synonym and RelatedTo) or relations that have been extensively explored in natural language processing (e.g. SimilarTo). Next, we retrieve a related concept group ce for each entity ve, where $ce = \{c_1, c_2, \ldots, c_o\}$. The related concepts c are listed in descending order of confidence scores, we filter the concepts with low confidence scores and retain the most relevant o concepts at the end of the concept search. Here, the o is set as five. Subsequently, we retrieve the affection tuple s of c in turn until one of c in ce is retrieved s. Thus, we obtain the affections S of all visual entities.

Finally, the feature extraction component employs BERT [58] to acquire the knowledge representation by encoding the description and relational knowledge, so as to leverage TD and S efficiently:

$$R_{kn} = BERT\left([CLS] + TD + [SEP] + S + [SEP]\right). \tag{1}$$

In this equation, TD and S respectively means the text and relational knowledge, [CLS] and [SEP] are the start and connecting tokens. R_{kn} is the final represented knowledge. Algorithm 1 describes the workflow of the knowledge encoder.

3.2. Visual encoder

Motivation. As video consists of typical sequence data, the visual encoder must not only focus on individual frames, but also comprehensively understand the entire video. However, previous methods primarily employed convolutional neural networks to extract the corresponding features, which may be detrimental to the quality of the generated summary [59]. To alleviate this problem, our encoder considers the information in both the spatial and temporal domains, where the former includes the visual content of each frame and the latter captures the long-term dependencies.

Method. An I3D model [60] pre-trained on the Kinetics action classification dataset [61] is adopted as our visual encoder, which is a two-stream architecture with spatiotemporal convolutions. In particular, it inflates two-dimensional convolution filters and pooling kernels leveraged by the Inception network into three dimensions and utilizes two independent networks to extract features from a sequence of dense RGB data and the optical flows, respectively. This allows it to capture features in terms of scene changes and motion in consecutive frames. Finally, a visual representation is obtained by merging the RGB and optical flow features:

$$R_v = F_{RGB} \oplus F_{opt}, \tag{2}$$

where R_v denotes the visual representation, $\oplus$ means the matrix addition operation. F_{opt} and F_{RGB} represent the features of the optical flow and RGB, respectively.

3.3. Audio encoder

Motivation. In general, a well-optimized multimodal model is superior to a unimodal model [16]. In addition, audio has demonstrated impressive representational capabilities for video summarization [62]. These statements encouraged us to build an audio encoder for our task.

Method. First, we formulate the audio sample waveform as a representation of the original audio. To imitate the principle of how humans hear from the sound, we resample the audio as 16 kHz. To better understand and leverage the audio representation, we apply a Fourier transform using a periodic Hann window function and further square the audio to produce a canonical spectrogram. Subsequently, to maintain the effectiveness of audio further extracting, we need to compute the stabilized log-Mel spectrogram. To achieve this, we map the acquired canonical spectrogram to 64 Mel-scaled bins ranging from 125 to 7500 Hz to obtain appropriate parameters that match the logarithmic perception of frequency. Eventually, we leverage an 11-layer VGG16 architecture [63] to extract audio representations from the stabilized log-Mel spectrogram. This process can be formalized as:

$$R_a = V\left(log\left(\beta(H\left(aud\right))^2\right)\right), \tag{3}$$

where R_a, V, and log denote the audio representation, the VGGish method, and logarithm operator respectively. β means mel-scaled bin. H and aud stand for the Hanning function and sampled audio respectively.

3.4. Multimodal fusion

Motivation. Besides encoding multimodal data, combining complementary content from multiple modalities is critical for video summarization, because meaningful cues are often spread across different modalities. Therefore, it is essential to design an effective fusion scheme for integrating valuable information. To address this issue, we propose a multimodal fusion module with intra- and inter-level combination.

Method. We design a special layer to aggregate the internal information to emphasize the most significant features in each modality.

This operation can assist in modelling the intra-level combination. With the generated representation of corresponding encoder, we can obtain the intra-context feature F_m^{inr} for each modality by:

$$F_m^{inr} = IIA\left(R_m, R_m\right), \tag{4}$$

where R_m mean the intra-context feature and F_m^{inr} is the representation for m modality. $m \in \{v, kn, a\}$. IIA is set for representing this internal information aggregation layer, which can be computed as:

$$IIA\left(k, q\right) = W_1\left(as_1 \odot as_2 \odot \ldots \odot as_h\right), \tag{5}$$

$$as_i = softmax\left(\phi\left(k_i \odot q_i\right) W_{as_i}\right) k_i, \tag{6}$$

in this equation, h stands for the number of head and we set the head number as 8. W is the learnable parameters and ϕ donates the hyperbolic tangent function. $\odot$ stands for the concatenation. Finally, k, q and as_i respectively represent a key sequence, query sequence and the output of the ith head attention.

In terms of the inter-level combination, especially for every two modalities, we first learn to calculate the weights associated with each other. Then, based on learned weights, we acquires the inter-context features and further aggregates with the following equations:

$$W_{m,n} = softmax\left(\sum_{i=0}^{len} L_{m,n,2}\left(\phi\left(L_{m,n,1}\left(F_{m(i)}^{inr} \odot F_n^{inr}\right)\right)\right)\right), \tag{7}$$

$$F_m^{int} = \frac{\sum_{n\in\{v,kn,a\}} W_{m,n}}{\omega}, \tag{8}$$

where $W_{m,n}$ is the learnt associated parameters for n and m modalities, $n \in \{v, kn, a\}$. For $L_{m,n,j}$, it is the jth linear layer for $W_{m,n}$ and we name the length of extracted features as len. $F_{m(i)}^{inr}$ is the ith vector of feature F_m^{inr}, F_n^{inr} is intra-context feature of n. ω is the number of modalities and F_m^{int} is inter-context feature of m.

Eventually, we apply a neural network with two layers to operate the inter-level multichannel feature combinations and calculate the corresponding importance score for each modality. The size of hidden layer in the neural networks is half of its input size. This process is formalized as:

$$V_m = softmax\left(W_3\phi\left(W_2 F_m^{int} + b_1\right) + b_2\right), \tag{9}$$

$$\widehat{F}_m = \phi\left(W_4 F_m^{int} + b_3\right), \tag{10}$$

$$F_f = \sum_m V_m \widehat{F}_m, \tag{11}$$

where V_m means the importance score for each modality, while b means the bias. After we transform the feature vector for each modality, we use $\widehat{F}_m$ to represent and F_f is the final fused features.

As the process shown above, our multimodal fusion module differs from earlier multimodal fusion approaches [31,32], which simply connect features at the input feature level. We employ a two-layer fusion architecture to investigate the inherent correlations among various modalities while simultaneously capturing the local dependencies within each modality. In addition, different from other audiovisual fusion methods [33,34,37], our multimodal fusion module considers implicit knowledge consisting of descriptive and relational knowledge.

3.5. Summary predictor

The purpose of this component is to project the fusion features into the probability space. We design a fully connected neural network with two layers as the summary predictor. The activation functions of the 1st and 2nd layers are the ReLU and sigmoid functions, respectively. Furthermore, we present two variants of the proposed KAMN that are derived by introducing different loss functions between the predicted importance scores and ground-truth labels. These two variants are defined as follows:

Supervised KAMN can be trained when the training videos with ground truth annotations are available. For the loss function, we employ the mean square error with $L2$ regularization to minimize the difference between the predicted probability and labels, which can be written as:

$$L_S = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widetilde{y}_i\right)^2 + \lambda\|\theta\|^2, \tag{12}$$

where θ includes all the parameters and λ stands for the regularization weights for $L2$. N represents the number of training data samples and i is the index of those samples. y, $\widetilde{y}$ respectively mean ground truth and the predicted frame-level importance score. Although supervised KAMN can utilize ground-truth data to achieve the desired performance intuitively, two major problems exist: (1) Obtaining ground truth labels is time consuming and laborious because annotators are required to watch a large volume of videos and assign importance scores to every shot in a single video. (2) Different annotators may have different understandings of the same video, causing inconsistencies in the ground truths created by different annotators. To address these problems, we propose an unsupervised model based on the vanilla supervised KAMN.

Unsupervised KAMN is a modified variant that can be trained without data annotation. Specifically, we compute the length regularization loss during training to penalize the number of frames selected forming the summary video over the input video, which can be expressed as:

$$L_{lr} = \left\|\frac{1}{I}\sum_{i=1}^{I}\widetilde{y}_i - \wp\right\|_2, \tag{13}$$

where I denotes the total number of video frames, $\wp$ is a tunable hyper-parameter. In addition, instead of only containing the crucial content of the input video, the selected frames of a high quality video summary should also be visually diverse. To this end, we apply a diversity loss [64] that calculate the mean of the pairwise dissimilarities among the selected frames:

$$L_d = \frac{1}{|y|\left(|y| - 1\right)}\sum_{t\in y}\sum_{t'\in y, t'\neq t}\left(1 - \frac{x_t^T x_{t'}}{\|x_t\|_2\|x_{t'}\|_2}\right), \tag{14}$$

where y denotes the indices of the selected frames, x_t represents the visual representation of the tth selected frame, T is the number of the input video frames. Consequently, our overall objective is:

$$L_U = L_{lr} + L_d. \tag{15}$$

For brevity, we use KAMN and KAMN$_{uns}$ to denote the supervised and unsupervised forms of our model, respectively.

4. Experiments

In this section, we describe various experiments that were conducted to verify the superiority of the proposed methods. First, we explain the experimental setup, including the benchmark datasets, evaluation approaches, comparison methods, and implementation details. To demonstrate the effectiveness of the proposed method, we compare it with baselines in terms of quantitative and qualitative evaluations. In addition, we visualize the summaries generated by our methods to demonstrate its superiority. Finally, runtime and parameter analyses are performed.

4.1. Experimental settings

Datasets. We report the video summarization results for four benchmark datasets: SumMe [65], TVSum [66], OVP [67], and YouTube [67]. SumMe contains 25 videos, including a wide range of topics, such as outdoor activities, holidays, and cooking, and the lengths of the videos vary from 1.5 to 6.5 min. TVSum contains 50 edited videos ranging from 1.5 to 10.5 min. It includes 10 categories, with five videos per category. The content of the videos is diverse and included TV news,

Table 1
Comparison of F-Score (%) using supervised video summarization approaches on SumMe and TVSum under the C, A and T configurations. Approaches marked with '*' denote multimodal-based approaches. The best scores are highlighted in bold.

Supervised method	SumMe			TVSum		
	C	A	T	C	A	T
dppLSTM	38.6	42.9	41.8	54.7	59.6	58.7
SUM-GAN	41.7	43.6	–	56.3	61.2	–
VASNet	49.7	51.1	–	61.4	62.4	–
DR-DSN	42.1	43.9	42.6	58.1	59.8	58.9
SUM-FCN	47.5	51.1	44.1	56.8	59.2	58.2
Cycle-SUM	44.8	–	–	58.1	–	–
MC-VSA	51.6	53.0	48.1	63.7	64.0	59.5
DSNet	51.2	53.3	47.6	61.9	62.2	58.0
SGSN	41.5	44.9	43.8	55.7	59.1	58.7
re-Seq2Seq	–	44.9	–	–	63.9	–
CSNet	48.6	48.7	44.1	58.5	57.1	57.4
HSA-RNN	44.1	–	–	59.8	–	–
H-MAN	51.8	52.5	48.1	60.4	61.0	59.5
PCDL	43.7	–	–	59.2	–	–
M-AVS	44.4	46.1	–	61.0	61.8	–
SUM-GDA	52.8	54.4	46.9	58.9	60.1	59.0
PGL-SUM	55.6	–	–	61.0	–	–
DASP	45.5	47.0	–	63.6	64.5	–
3DST-UNet	47.4	49.9	47.9	58.3	58.9	56.1
RSGN	45.0	45.7	44.0	60.1	61.1	60.0
VJMHT	50.6	51.7	46.4	60.9	61.9	58.9
iPTNet	54.5	56.9	49.2	63.4	64.2	59.8
AVRN*	44.1	44.9	43.2	59.7	60.5	58.7
SASUM*	45.3	–	–	58.2	–	–
SMIL*	41.2	–	–	51.3	–	–
HMT*	44.1	44.8	–	60.1	60.3	–
TST*	39.4	–	–	56.2	–	–
MT*	41.5	–	–	57.9	–	–
MSVA*	54.5	–	–	62.8	–	–
KAMN*	**59.7**	**63.7**	**61.0**	**63.8**	**65.4**	**63.5**

Table 2
Comparison of F-Score (%) using unsupervised video summarization approaches on SumMe and TVSum under the C, A and T configurations. The best scores are highlighted in bold.

Unsupervised method	SumMe			TVSum		
	C	A	T	C	A	T
SUM-GAN	39.1	43.4	–	51.7	59.5	–
DR-DSN	41.4	42.8	42.4	57.6	58.4	57.8
SUM-FCN	41.5	–	39.5	52.7	–	52.9
Cycle-SUM	**57.6**	–	–	41.9	–	–
ACGAN	46.0	47.0	44.5	58.5	58.9	57.8
SGA	48.9	–	–	58.3	–	–
MC-VSA	44.6	–	–	58.0	–	–
UnpairedVSN	47.5	48.0	41.6	55.6	56.1	55.7
GL-RPE	50.2	–	–	59.1	–	–
CSNet$_{uns}$	51.3	**52.1**	45.1	58.8	59.0	59.2
PCDL$_{uns}$	42.7	–	–	58.4	–	–
SUM-GDA$_{uns}$	50.0	50.2	**46.3**	59.6	60.5	58.8
3DST-UNet$_{uns}$	44.6	49.5	45.7	58.3	58.4	58.0
RSGN$_{uns}$	42.3	43.6	41.2	58.0	59.1	**59.7**
KAMN$_{uns}$	50.2	51.4	45.3	**60.3**	**61.8**	59.0

Table 3
Comparisons of the F-score (%) between our method and baseline methods on SumMe, TVSum, Youtube datasets under canonical configuration.

Method	SumMe	TVSum	YouTube
VSUMM$_1$	32.8	–	58.7
VSUMM$_2$	33.7	–	59.9
seqDPP$_{LINEAR}$	–	–	57.8
seqDPP$_{N.NETS}$	–	–	60.3
SummaryTransfer	40.9	–	61.8
SUM-GAN$_{uns}$	39.1	51.7	60.1
SUM-GAN	41.7	56.3	62.5
SASUM	45.3	58.2	60.3
Cycle-SUM$_{uns}$	41.9	57.6	63.0
Cycle-SUM	44.8	58.1	64.2
Cycle-SUM$_{DPP}$	42.3	57.9	63.5
KAMN$_{uns}$	50.2	60.3	61.5
KAMN	**59.7**	**63.8**	**70.8**

sports activities, and dog shows. These two datasets were captured from a first- or third-person perspective and present frame-level importance scores labelled by multiple participants. With reference to previous works [9,22], we also considered two other datasets, namely OVP and YouTube, to construct more data configurations for the evaluations. The former consists of 50 videos and the latter consists of 39 videos, excluding the cartoon videos. The topics of these videos cover educational lectures and TV shows.

Compared methods. We evaluate the proposed KAMN and KAMN$_{uns}$ with the following methods. (1) Unimodal-based method: ACGAN [68], SGA [26] MC-VSA [69] DSNet [9], dppLSTM [15], MPRNet [21], SGSN [70], Cycle-SUM [71], re-Seq2Seq [8], UnpairedVSN [25], GL-RPE [27], CSNet [6], HSA-RNN [7], H-MAN [72], PCDL [24], M-AVS [73], VASNet [74], SUM-GDA [75], PGL-SUM [76], SUM-FCN [77], DASP [78], SUM-GAN [79], 3DST-UNet [59], RSGN [80], VJMHT [22], iPTNet [23], VSUMM [67], DR-DSN [64], seqDPP [81], SummaryTransfer [82]. (2) Multimodal-based method: AVRN [33] SASUM [38], SMIL [28], HMT [37], TST [37], MT [37], MSVA [83].

Evaluation approaches. We selected two types of metrics to demonstrate the superiority of the proposed method. The first type is F-score [15], which is the most commonly used evaluation approach in this field. The second is the rank correlation coefficients (RCC) [84]. Further details on these two metrics are provide below. (1) F-score: it is the harmonic mean of precision (P) and recall (R) and expressed as:

$$F = 2 \times \frac{P \times R}{P + R} \times 100\%, \tag{16}$$

where precision and recall are defined as follows:

$$P = \frac{length(gs \cap as)}{length(gs)}, R = \frac{length(gs \cap as)}{length(as)}, \tag{17}$$

where gs and as denote the generated summary and corresponding annotated summary, respectively. (2) Rank correlation coefficients: A recent study [84] proposes an approaches for assessing the Spearman's ρ [85] and Kendall's τ [86] correlation coefficients between the predicted and human-annotated importance scores, to assesses the quality of a machine-generated video summary.

Besides these two objective evaluation metrics, in the user study, building on research in both communication, psychology and the conventions of video editing [62], we conceptualized a high-quality summary as one that clearly and extensively. (1) provided diversified visual content and (2) contained as much important content of the original video as possible. The overall quality of the generated summary is a subjective metric that is used to measure the degree to which it satisfies the above definitions.

Implementation details. We conduct various experiments using four data configurations: (1) Canonical (C): We use the standard five-fold cross validation, dividing one dataset with a ratio of 8:2, where the former part is for training and the latter part is for testing. (2) Augmented (A): We still use the five-fold cross validation and utilize the other three datasets to augment the training data of canonical setting. For instance, when performing evaluation on the TVSum dataset, we randomly select 20% of the TVSum videos for testing, and all the remaining videos in the four datasets are used for training. (3) Transfer (T): For the target testing dataset, we use the other datasets as the training data in this setting. (4) One-to-one transfer: This is similar to the transfer setting, but only involve training the model on one dataset and testing on the other. For fairness, we transform the annotation formats provided by the different datasets by applying the method described in [15,74]. To reduce the temporal redundancy, we uniformly subsample the videos as 2 fps. Following a previous work [74],

Table 4
Comparison with baselines on TVSum and SumMe, using the rank correlation coefficients proposed in [84]. The best and second-best performing methods are highlighted in bold and italics, respectively.

Method	SumMe		TVSum	
	Spearman's ρ	Kendall's τ	Spearman's ρ	Kendall's τ
dppLSTM	−0.0311 ± 0.0249	−0.0256 ± 0.0214	0.0385 ± 0.0365	0.0298 ± 0.0284
SUM-GAN	−0.0122 ± 0.0504	−0.0095 ± 0.0410	−0.0701 ± 0.0444	−0.0535 ± 0.0340
DR-DSN	0.0501 ± 0.0470	0.0433 ± 0.0386	0.0227 ± 0.0666	0.0169 ± 0.0508
SUM-FCN	0.0096 ± 0.0111	0.0080 ± 0.0091	0.0142 ± 0.0042	0.0107 ± 0.0032
SGA$_{uns}$	−0.0226 ± 0.0695	−0.0180 ± 0.0558	−0.0620 ± 0.0388	−0.0472 ± 0.0299
MPRNet	0.0137 ± 0.0505	0.0108 ± 0.0407	**0.2301 ± 0.0320**	**0.1758 ± 0.0243**
VJMHT	*0.1080 ± 0.0000*	**0.1060 ± 0.0000**	0.1050 ± 0.0000	0.0970 ± 0.0000
HMT*	0.0758 ± 0.0434	*0.0563 ± 0.0258*	0.107 ± 0.0320	0.0603 ± 0.0264
KAMN$^{*}_{uns}$	0.0632 ± 0.0095	−0.0572 ± 0.0100	0.0724 ± 0.0315	0.0678 ± 0.0280
KAMN*	**0.2186 ± 0.0021**	0.0264 ± 0.0358	*0.1150 ± 0.0242*	*0.1060 ± 0.0185*

Table 5
Comparisons of F-score (%) with some baselines using different combinations of training (line 1) and testing (line 2) datasets, on one-to-one transfer setting. YT, OV, TV, SM respectively represent the YouTube, OVP, TVSum and SumMe datasets.

Method	YT	OV	TV	YT	OV	SM
		SumMe			TVSum	
VASNet	44.6	44.9	44.1	58.2	58.5	58.8
ACGAN	42.6	43.9	44.0	56.3	57.3	56.6
DSNet	47.1	48.3	44.6	57.9	58.9	58.4
DR-DSN	41.5	42.4	40.5	57.3	58.9	58.5
KAMN	**59.9**	**60.9**	**47.8**	**62.8**	**63.0**	**61.4**

the summary generated by the proposed method is limited to 15% of the input video length. Specifically, we performed zero padding to acquire the audio representations of silent videos. To obtain knowledge representation, we encode the text description and affections with a pre-trained uncased base model of BERT. The experiments are perform on an Nvidia GTX 2080Ti GPU, and 300 epochs of training are perform for each validation fold. We apply dropout for fully connected layer with a ratio of 0.3. The supervised KAMN uses Adam optimizer with a base learning rate of 5×10^{-4} to update all the parameters, and the $L2$ regularization item is equal to 10^{-3}.

After training, followed [9,15], the models that obtained the best F-scores for each test set were selected. Each reported scores are the average of these five models' performance on the test set. We initialize the hyperparameters with empirical values based on [15] and then use grid search for tuning, with the learning rate from 0.05 to 0.0005 dividing 10 each time, weight decay from 0.01 to 0.0001 dividing 10 each time, dropout from 0.1 to 0.5 with the interval of 0.1, training epoch from 100 to 1000 with the interval of 100.

4.2. Performance comparisons

Quantitative results (F-Score). We present the performances of KAMN and the supervised baselines in Table 1, KAMN$_{uns}$ and the unsupervised baselines in Table 2, and the results for an additional YouTube dataset under the canonical configuration in Table 3. The following observations can be made from Table 1: (1) Our supervised method achieves remarkable success on the SumMe and TVSum datasets, surpassing all baselines under all three configurations. (2) Previous supervised state-of-the-art methods achieve 55.6% under the C configuration on SumMe. However, KAMN directly increased this to approximately 60% for the first time. In general, it is challenging to acquire F-score enhancement on SumMe because the videos in it are shorter and contain fewer shots with dialogue or subtitles [6]. However, our knowledge encoder assists in extracting the frame descriptions and contained affections, which provides additional modality information for the entire approach. (3) The results in the A and T configurations show that KAMN exhibit superior performance compared to the baselines. For each method, the F-score in the A configuration is better than the C one, possibly because the A configuration contained more training data. With regard to the challenging T data configuration, we prove the strong generalization ability of our method. To be specific, KAMN can achieve 61.0% under T configuration on SumMe, which is 13.4% higher than that of the visual unimodal-based DSNet, and 17.8% higher than that of the audiovisual bimodal-based AVRN. This further demonstrates the effectiveness of introducing implicit knowledge into video summarization. It can be clearly observed from Table 2 that our unsupervised method still achieves competitive results. Specifically, KAMN$_{uns}$ obtains the best results in canonical, augmented configurations on the TVSum dataset and competitive performance in other configurations. The additional experiments results in Table 3 also demonstrate the superiority of both KAMN and KAMN$_{uns}$ on YouTube dataset.

Quantitative results (RCC). From the results of Spearman's ρ on the SumMe dataset in Table 4, we observe that KAMN achieves the best performance and outperforms the second-best VJMHT by a large margin i.e., 0.1106. Some methods such as dppLSTM, SGA$_{uns}$ struggled to generalize to the test datasets. Similar to SumMe, our method obtains the second-best performance on TVSum dataset, and the SUM-GAN show poor generalizability. Concerning Kendall's τ, KAMN achieves the second and third best results. The superior performance of KAMN may be owing to the introduction of implicit knowledge and appropriate fusion of multichannel representations.

One-to-one transfer. We perform experiments on the one-to-one transfer evaluation configurations listed in Table 5 to evaluate the generalization ability of KAMN further. We clearly find that our supervised method achieves the most appealing performance compared to the baselines. For example, KAMN is higher than DSNet by 12.8%, 12.6% and 3.2% when transferring to SumMe, indicating the stronger generalization ability of our model. The difference in the results of transferring to one from diverse datasets indicate the distance discrepancy among different datasets. When transferring to SumMe, the achievement of better or comparable F-scores in OVP compared to the other two datasets suggests that the OVP domain is closer to the SumMe domain.

Ablation investigation. We conduct an ablation study by comparing the supervised KAMN with several variants, as shown in Table 6, to evaluate the effectiveness of the different components of our method comprehensively. For brevity, we use KAMN$_1$ to KAMN$_{14}$ to denote the different variants, and these variant models are divided into four groups. Particularly, the KAMN$_1$ to KAMN$_4$ uses unimodal features, the KAMN$_5$ to KAMN$_8$ uses the features of two modalities, the KAMN$_9$ to KAMN$_{13}$ utilize the information of three modalities, while experiments KAMN$_{14}$ is the supervised KAMN. The KAMN$_9$ to KAMN$_{13}$ are ablate one component each time.

We can observe that the entire KAMN model achieves the best performance among different component combinations. Adequate insights can also be obtained regarding the reasons for the competitive performance under the challenging T configuration in Table 1. Specifically, from KAMN$_5$, KAMN$_6$, and KAMN$_7$, we can observe that Entity Detection + Relational Knowledge are the main contributor. Although

Table 6
Ablation study based on the F-Score (%) of different components of the proposed model, on SumMe and TVSum. The VE, FD, RK, AE and MFM represent the visual encoder, frame description, relational knowledge, audio encoder and multimodal fusion module, respectively.

Exp.	VE	FD	RK	AE	MFM	SumMe			TVSum		
						C	A	T	C	A	T
1.	✓	✗	✗	✗	✗	48.2	46.1	41.9	60.8	60.9	52.8
2.	✗	✓	✗	✗	✗	47.8	47.9	42.2	60.2	61.4	58.0
3.	✗	✗	✓	✗	✗	51.8	52.8	45.3	60.1	61.4	58.2
4.	✗	✗	✗	✓	✗	46.8	49.0	42.5	57.9	59.5	56.9
5.	✓	✗	✗	✓	✗	45.9	46.6	42.4	60.5	60.1	53.2
6.	✓	✓	✗	✗	✗	43.5	44.1	41.8	59.5	58.7	49.1
7.	✓	✗	✓	✗	✗	48.3	52.7	50.4	62.4	58.2	56.7
8.	✓	✓	✓	✗	✗	51.7	55.2	52.5	62.0	61.9	60.6
9.	✗	✓	✓	✓	✓	51.9	52.8	45.1	59.8	60.9	57.7
10.	✓	✗	✓	✓	✓	51.7	52.6	47.2	62.4	64.6	61.6
11.	✓	✓	✗	✓	✓	52.0	53.4	46.9	61.9	61.8	59.1
12.	✓	✓	✓	✗	✓	55.5	59.7	58.4	62.1	64.4	62.6
13.	✓	✓	✓	✓	✗	50.9	57.0	52.3	62.7	61.8	59.0
14.	✓	✓	✓	✓	✓	59.7	63.7	61.0	63.8	65.4	63.5

Table 7
The user study results when leveraging different video summarization methods on videos 1, 10, 11, and 14 to 16 in SumMe.

Method	SumMe					
	1	10	11	14	15	16
VASNet	4.4	4.5	4.4	4.2	**4.2**	2.3
DR-DSN	3.6	3.5	4.5	4.5	2.7	4.6
DSNet	2.0	3.0	1.6	2.6	1.9	1.6
HMT	4.1	3.5	2.8	3.4	1.5	4.8
KAMN	**5.4**	**5.3**	**5.4**	**5.4**	3.5	**5.6**
Human	6.0	5.8	6.0	4.8	5.7	5.0

Table 8
User study results when leveraging different video summarization methods on videos 1, 10, 11, and 14 to 16 in TVSum.

Method	TVSum					
	1	10	11	14	15	16
VASNet	3.6	3.6	4.8	3.1	2.1	3.6
DR-DSN	4.2	2.8	4.8	4.4	4.0	2.0
DSNet	2.8	3.6	1.0	4.2	4.4	4.4
HMT	3.7	2.5	3.3	2.5	1.2	4.8
KAMN	**5.8**	**4.8**	**5.4**	**5.1**	**4.5**	**5.6**
Human	6.2	5.2	5.0	6.1	5.1	6.2

Table 9
The statistical significance of the improvement of KAMN over other methods in term of user study (Wilcoxon test).

Dataset		VASNet	DR-DSN	DSNet	HMT
Sum Me	1	1.93e−18	1.94e−18	1.95e−18	1.93e−18
	10	2.04e−13	1.98e−18	1.90e−18	6.57e−18
	11	1.90e−17	4.45e−13	9.89e−19	1.98e−18
	14	1.94e−18	1.93e−18	1.94e−18	1.94e−18
	15	4.31e−1	3.51e−17	1.93e−18	1.90e−18
	16	1.93e−18	1.90e−18	1.93e−18	3.01e−18
TV Sum	1	1.98e−18	2.08e−18	9.43e−19	1.95e−18
	10	1.94e−18	1.93e−18	1.94e−18	1.90e−18
	11	8.20e−17	7.29e−17	1.93e−18	1.94e−18
	14	1.98e−18	8.50e−16	1.26e−17	1.97e−18
	15	1.94e−18	2.30e−15	2.29e−2	1.96e−18
	16	1.97e−18	1.93e−18	2.41e−18	3.02e−18

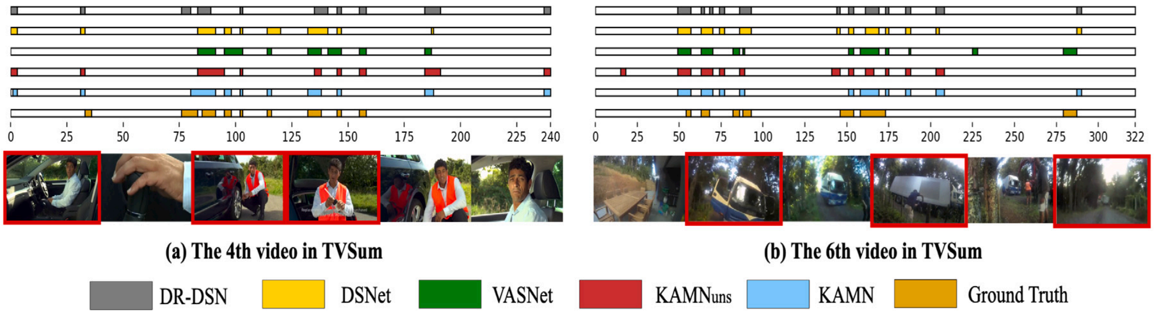
the topics of videos in different datasets are not the same, the overall affections are finite and similar. Therefore, detecting and adding affections in the model training process can reduce the gap between datasets.

Remarkably, the results from KAMN$_{9}$ to KAMN$_{13}$ show that even in the absence of a certain modal feature, the variant models can achieve competitive performance by benefiting from other modalities. For example, even the KAMN$_{12}$ without audio encoder, the performance of KAMN$_{12}$ in the canonical configuration is still better than the multimodal-based HMT model, the F-scores of KAMN$_{12}$ on the SumMe and TVSum datasets are 11.4% and 2% higher than those of HMT, respectively. This proves the robustness of the proposed method It is worth noting that the performance of these variants does not necessarily improve with an increasing number of modalities, which may be because different modalities generalize with one another at different rates, so simply concatenating them is not optimal [16]. For instance, KAMN$_{5}$ with audio underperforms KAMN$_{1}$ without audio on both datasets in the C configuration. In comparison, KAMN$_{12}$ outperforms KAMN$_{8}$ on both datasets when the proposed fusion module is added, demonstrating the effectiveness of the fusion module.

Qualitative evaluation. Owing to the subjective nature of video summarization, we design a user study to evaluate our approach further, and the results are reported in Tables 7 and 8. Specifically, we recruited 100 human subjects with at least two years of experience in computer vision to participate in the study, including 68 males and 32 females, with ages ranging from 22 to 27. During these tests, the subjects were allowed to view each summary multiple times before making a decision and each summary is rated using a 7-point Likert scale. The viewing conditions conformed to the international standard procedure multimedia subjective test [87]. Due to space constraints, we randomly selected several test results, as listed in Tables 7 and 8. It can be observed that KAMN consistently achieved the best performance in most videos, except for video 15 on the SumMe dataset. Another interesting observation is that the results of the subjective evaluation are not always in accordance with those of the objective evaluation. For instance, in the SumMe dataset, the DR-DSN achieves a score of 3.9 for all videos on average, whereas DSNet only achieves a score of 2.12. However, Table 1 shows that DSNet outperforms DR-DSN in all three configurations on the SumMe dataset. This reflects the gap between the subjective and objective evaluations in the video summarization task, indicating that video summarization methods that perform well in objective criteria do not always guarantee satisfactory performance in practice.

Statistical analysis. Table 9 displays the Wilcoxon test results for the user studies in Tables 7 and 8 from a statistical perspective. From Table 9, we can observe that the P-values between the KAMN and these baselines are mostly less than 0.05, which proves that our KAMN significantly outperforms the baselines.

Case Study. We visualize the selected frames of the generated summary and compared them with the three baselines and the ground



(a) The 4th video in TVSum

(b) The 6th video in TVSum

Fig. 3. Selected frames visualization among the whole video by DR-DSN, DSNet, VASNet, KAMN$_{uns}$, KAMN, and the ground truth for 4th and 6th videos in TVSum. The colourful bars are selected frames, while the value of x coordinate stands for second.

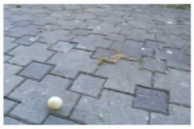
Num	A	B	C	D
Frame				
Text description	A dog and a bird are playing with a ball	A ball is shown on the ground	a close up of a sandwich on the dinning table	a close up of a computer screen with a sky background
Visual entities	dog bird ball plant	ball	sandwich	none
Polarity	positive*4	positive	positive	negative
Intensity	0.817, 0.843*2, 0.835	0.843	0.9	-0.64
Emotion	joy*3 eagerness*3, pleasantness*2	pleasantness, eagerness	joy, eagerness	sadness, disgust
Sensitivity	0.966, 0.885, 0.7, 0	0.7	0.9	-0.97
Important score	0.5625 **(min. 0, max. 0.5625)**	0 **(min. 0, max. 0.5625)**	3.55 **(min. 1.1, max. 3.6)**	1.4 **(min. 1.1, max. 3.6)**

Fig. 4. Some basic information and corresponding implicit knowledge of different frames. A, B come from video 25 in SumMe. C, D come from video 20 in TVsum. The maximum and minimum important scores of each video, which contains the frame displayed, are shown in the important score line. The value in red means corresponding frame is the critical frame, while the value in green means not.

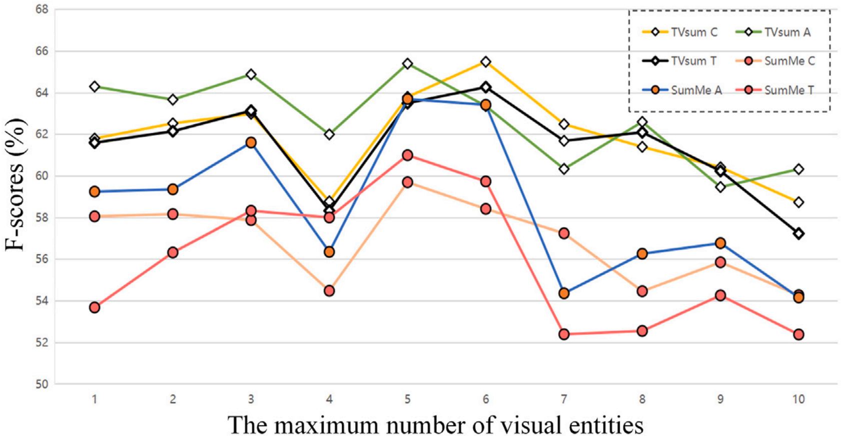


Fig. 5. KAMN parameter analysis of the maximum number of visual entities u in the entity detection part of knowledge encoder under canonical, augmented and transfer settings.

truth, as shown in Fig. 3, to demonstrate the effectiveness of our method. The visualization results intuitively demonstrate that our method can achieve the best performance because of the most similarity between its selected frames and the ground truth one. Fig. 4 provides some examples of the implicit knowledge captured by the knowledge encoder. Every two adjacent sample pairs belong to the same video. We can observe the following from the figure: (1) Frames rich in content can be perceived as more visual entities. For instance, frame A, which describes a relatively more complex scene, contains four entities, whereas frame B only contains only one. Frames that cannot be perceived as entity typically appear at the opening and end of the video. Properly denoting them as ‘none’ helps KAMN to skip or ignore them. (2) Each entity can inspire four kinds of affective attributes and frames with more positive affective attributes often correspond to higher important scores. For example, compared with D, frame C owns the scores of intensities and sensitivity above zero, relative positive

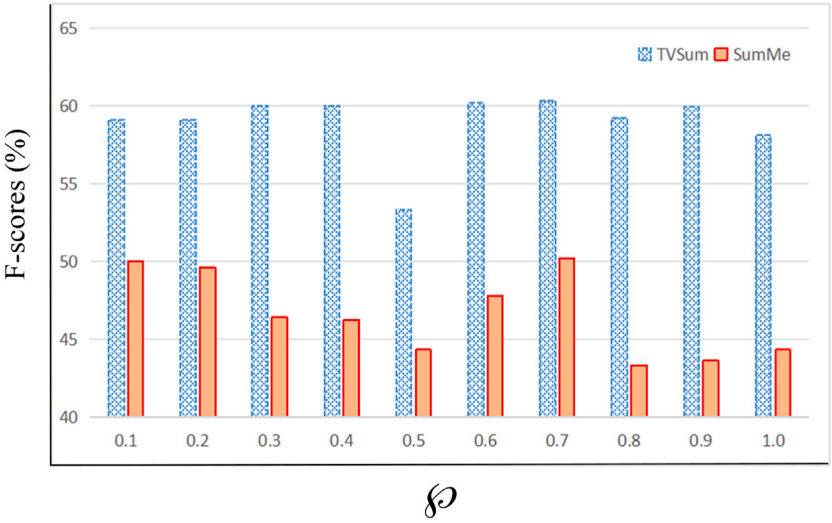


Fig. 6. KAMN$_{uns}$ parameter analysis of $\wp$ in Eq. (13) from 0 to 1 under canonical setting on TVSum and SumMe datasets.

Table 10
The average number of frames per video, and the per epoch training time (seconds) between our method and some baselines.

	Method	SumMe	TVSum
Average frames	–	293	470
Training time	DSNet	0.32	0.86
	VASNet	0.38	0.80
	SUM-GAN	11.85	38.95
	DR-DSN	0.33	0.98
	SGA	16.39	54.23
	CSNet	28.43	89.85
	KAMN$_{uns}$	2.36	3.36
	KAMN	1.30	4.12

emotions and higher important score. Moreover, the text description of frames can provide more semantic information. Therefore, we can conclude that adding the implicit knowledge captured by video frames in the training process can enhance the capability of the model in selecting important frames.

Running Speed. Table 10 presents the average training times per epoch for the SumMe and TVSum datasets. It can be observed that our supervised method can achieve 1.30 and 4.12 s on the SumMe and TVSum datasets, respectively, which is better than the majority of the baselines. Although our method is slower than DSNet, VASNet, and DR-DSN, we claim that the additional time is a reasonable compromise given the improvements in quality of the generated summaries.

Parameter analysis. We conduct a parameter analysis of the tunable hyper-parameter in our method, which are the maximum number of visual entities u in the entity detection part of the knowledge encoder in Fig. 5, and the $\wp$ in Eq. (13) in Fig. 6. From the former figure, we can observe that our method obtains the best performance on average when the number is five. This phenomenon is consistent with the results of previous studies [45,52]. Therefore, we set the default value (the maximum number of visual entities) as five in our experiments. From the latter figure, we can find our method achieves the best performance when $\wp$ is 0.7, which is set as the default value.

5. Conclusion and discussion

In this paper, we propose a novel approach for both supervised and unsupervised video summarization by fusing the representations of three modalities: visual, audio, and implicate knowledge. We develop a knowledge encoder that considers implicate knowledge and provides additional descriptive and relational knowledge for the entire approach. The visual and audio encoders are designed to exploit audio-visual information that naturally captures spatiotemporal information and sound clues. Particularly, we present a multimodal fusion module that combines complementary and valuable information from multiple modalities. The experiments show that our supervised approach achieves significantly better results than the baseline methods. To address the challenge of inadequate training data, we extended KAMN to KAMN$_{uns}$. Even with unlabelled data, the unsupervised variant of the KAMN model exhibit competitive performance. More importantly, the evaluation of the proposed method highlights its advantages over existing video summarization models.

5.1. Limitations

Despite achieving remarkable success in this task, some limitations still exist. (1) Although both audio and visual information are important in the video summarization task, to reflect a particular artistic style, part of the audiovisual information may be absent, which may present a challenge for our video summarization model. (2) From the experimental results, we can observe an obvious gap between the subjective and objective evaluations. We believe that two main constraints lead to this discrepancy. First, most existing video summarization methods learn frame importance using a neural network to model the videos. However, large-scale annotated datasets that can be useful for improving the generalization ability of video summarization methods are lacking. Second, the aesthetic elements of a video also greatly affect humans when selecting important shots. Although previous studies [88] have proposed various methods to produce summaries, they ignored factors that can influence the personal sense of audiences, such as the aesthetic quality of videos. Consequently, in some cases, video summarization methods that perform well based on objective criteria are not satisfactory in practice.

5.2. Future work

We consider the following future research directions to address the abovementioned issues The first is to develop effective and audiovisual alignment strategies to reduce the negative influence due to the presence of low-quality data on the model. In addition, it may be advantageous to formulate rules based on film-making guidelines to assist the model in selecting shots with particular artistic expressions. Second, to overcome the disadvantages of prevailing datasets, we plan to construct an annotated video summarization dataset containing more videos and scenes (such as movies and documentaries), with longer durations (10 to 60 min). In addition, we will investigate the effects of subjective audience preferences and cinematographic aesthetics guidelines, such as the shot length, colour continuity, and shot stability on video summaries to reduce the gap between subjective and objective evaluations. Finally, we will design a visual aesthetics encoder to capture the aesthetic elements of the shots and a more effective multimodal fusion module to capture valuable information.

CRediT authorship contribution statement

Jiehang Xie: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Xuanbai Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Data curation. **Sicheng Zhao:** Writing – review & editing, Writing – original draft. **Shao-Ping Lu:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [Grant No. 62132012 and No. 61972216], and CCF-DiDi GAIA Collaborative Research Funds for Young Scholars.

References

[1] E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, I. Patras, Video summarization using deep neural networks: A survey, Proc. IEEE 109 (11) (2021) 1838–1863.

[2] L. Lan, C. Ye, Recurrent generative adversarial networks for unsupervised WCE video summarization, Knowl.-Based Syst. 222 (2021) 106971.

[3] Y. Yuan, J. Zhang, Unsupervised video summarization via deep reinforcement learning with shot-level semantics, IEEE Trans. Circuit Syst. Video Technol. 33 (1) (2023) 445–456.

[4] X. Zhu, X. Wu, J. Fan, A.K. Elmagarmid, W.G. Aref, Exploring video content structure for hierarchical summarization, Multimedia Syst. 10 (2) (2004) 98–115.

[5] S. Pfeiffer, R. Lienhart, S. Fischer, W. Effelsberg, Abstracting digital movies automatically, J. Vis. Commun. Image Represent. 7 (4) (1996) 345–353.

[6] Y. Jung, D. Cho, D. Kim, S. Woo, I.S. Kweon, Discriminative feature learning for unsupervised video summarization, in: AAAI, 2019, pp. 8537–8544.

[7] B. Zhao, X. Li, X. Lu, HSA-RNN: Hierarchical structure-adaptive rnn for video summarization, in: CVPR, 2018, pp. 7405–7414.

[8] K. Zhang, K. Grauman, F. Sha, Retrospective encoders for video summarization, in: ECCV, 2018, pp. 383–399.

[9] W. Zhu, J. Lu, J. Li, J. Zhou, DSNet: A flexible detect-to-summarize network for video summarization, IEEE Trans. Image Process (2021) 948–962.

[10] A. Hanjalic, Adaptive extraction of highlights from a sport video based on excitement modeling, IEEE Trans. Multimed. 7 (6) (2005) 1114–1122.

[11] M. Narasimhan, A. Rohrbach, T. Darrell, CLIP-It! language-guided video summarization, in: NIPS, 2021, pp. 13988–14000.

[12] Z. Li, J. Tang, X. Wang, J. Liu, H. Lu, Multimedia news summarization in search, ACM Trans. Intell. Syst. Technol. 7 (3) (2016) 1–20.

[13] P. Cao, X. Zuo, Y. Chen, K. Liu, J. Zhao, Y. Chen, W. Peng, Knowledge-enriched event causality identification via latent structure induction networks, in: IJCAI, 2021, pp. 4862–4872.

[14] J. Gao, T. Zhang, C. Xu, Watch, think and attend: End-to-end video classification via dynamic knowledge evolution modeling, in: ACM MM, 2018, pp. 690–699.

[15] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, in: ECCV, 2016, pp. 766–782.

[16] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard? in: CVPR, 2020, pp. 12692–12702.

[17] J. Xie, X. Chen, S.-P. Lu, Y. Yang, A knowledge augmented and multimodal-based framework for video summarization, in: ACM MM, 2022, pp. 740–749.

[18] Z. Ji, K. Xiong, Y. Pang, X. Li, Video summarization with attention-based encoder–decoder networks, IEEE Trans. Circuit Syst. Video Technol. 30 (6) (2020) 1709–1717.

[19] M. Ma, S. Mei, S. Wan, Z. Wang, D.D. Feng, M. Bennamoun, Similarity based block sparse subset selection for video summarization, IEEE Trans. Circuit Syst. Video Technol. 31 (10) (2021) 3967–3980.

[20] T.-J. Fu, S.-H. Tai, H.-T. Chen, Attentive and adversarial learning for video summarization, in: WACV, 2019, pp. 1579–1587.

[21] Y. Saquil, D. Chen, Y. He, C. Li, Y.-L. Yang, Multiple pairwise ranking networks for personalized video summarization, in: ICCV, 2021, pp. 1718–1727.

[22] H. Li, Q. Ke, M. Gong, R. Zhang, Video joint modelling based on hierarchical transformer for co-summarization, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–14.

[23] H. Jiang, Y. Mu, Joint video summarization and moment localization by cross-task sample transfer, in: CVPR, 2022, pp. 16388–16398.

[24] B. Zhao, X. Li, X. Lu, Property-constrained dual learning for video summarization, IEEE Trans. Neural Netw. Learn. Syst. 31 (10) (2019) 3989–4000.

[25] M. Rochan, Y. Wang, Video summarization by learning from unpaired data, in: CVPR, 2019, pp. 7902–7911.

[26] E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, I. Patras, Unsupervised video summarization via attention-driven adversarial learning, in: MMM, 2020, pp. 492–504.

[27] Y. Jung, D. Cho, S. Woo, I.S. Kweon, Global-and-local relative position embedding for unsupervised video summarization, in: ECCV, 2020, pp. 167–183.

[28] J. Lei, Q. Luan, X. Song, X. Liu, D. Tao, M. Song, Action parsing-driven video summarization based on reinforcement learning, IEEE Trans. Circuit Syst. Video Technol. 29 (7) (2019) 2126–2137.

[29] R.R. Shah, Y. Yu, A. Verma, S. Tang, A.D. Shaikh, R. Zimmermann, Leveraging multimodal information for event summarization and concept-level sentiment analysis, Knowl.-Based Syst. 108 (2016) 102–109.

[30] X. Wang, L. Zhu, Z. Zheng, M. Xu, Y. Yang, Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision, IEEE Trans. Multimed. 25 (2023) 6079–6089.

[31] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, YouTube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53.

[32] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: ICDM, 2016, pp. 439–448.

[33] B. Zhao, M. Gong, X. Li, AudioVisual video summarization, IEEE Trans. Neural Netw. Learn. Syst. 34 (8) (2023) 5181–5188.

[34] Y. Wu, L. Zhu, Y. Yan, Y. Yang, Dual attention matching for audio-visual event localization, in: ICCV, 2019, pp. 6291–6299.

[35] W. Jiang, C. Cotton, A.C. Loui, Automatic consumer video summarization by audio and visual analysis, in: ICME, 2011, pp. 1–6.

[36] M. Rhevanth, R. Ahmed, V. Shah, B.R. Mohan, Deep learning framework based on audio–Visual features for video summarization, in: Advanced Machine Intelligence and Signal Processing, 2022, pp. 229–243.

[37] B. Zhao, M. Gong, X. Li, Hierarchical multimodal transformer to summarize videos, Neurocomputing 468 (2022) 360–369.

[38] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, C. Yao, Video summarization via semantic attended networks, in: AAAI, 2018, pp. 216–223.

[39] H. Li, Q. Ke, M. Gong, T. Drummond, Progressive video summarization via multimodal self-supervised learning, in: WCAV, 2023, pp. 5584–5593.

[40] C. Chen, L. Lo, S. Lin, COHETS: A highlight extraction method using textual streams of streaming videos, Knowl.-Based Syst. 258 (2022) 110000.

[41] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, in: AAAI, 2017, pp. 4444–4451.

[42] A. Sören, B. Christian, K. Georgi, L. Jens, C. Richard, I. Zachary, DBpedia: A nucleus for a web of open data, in: ISWC, 2007, pp. 722–735.

[43] C.X. Junyu Gao, I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs, in: AAAI, 2019, pp. 8303–8311.

[44] K. Marino, R. Salakhutdinov, A. Gupta, The more you know: Using knowledge graphs for image classification, in: CVPR, 2017, pp. 20–28.

[45] Q. Wu, C. Shen, P. Wang, A. Dick, A.v.d. Hengel, Image captioning and visual question answering based on attributes and external knowledge, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1367–1381.

[46] Y. Zhang, M. Jiang, Q. Zhao, Explicit knowledge incorporation for visual reasoning, in: CVPR, 2021, pp. 1356–1365.

[47] L. Zhang, S. Liu, D. Liu, P. Zeng, X. Li, J. Song, L. Gao, Rich visual knowledge-based augmentation network for visual question answering, IEEE Trans. Neural Netw. Learn. Syst. 32 (10) (2021) 4362–4373.

[48] F. Qi, X. Yang, C. Xu, Emotion knowledge driven video highlight detection, IEEE Trans. Multimed. (2020) 1–15.

[49] H. Wang, F. Zhang, X. Xie, M. Guo, DKN: Deep knowledge-aware network for news recommendation, in: WWW, 2018, pp. 1835–1844.

[50] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, K. Wilson, CNN architectures for large-scale audio classification, in: ICASSP, 2017, pp. 131–135.

[51] X. Mengde, Z. Zheng, H. Han, W. Jianfeng, W. Lijuan, W. Fangyun, B. Xiang, L. Zicheng, End-to-end semi-supervised object detection with soft teacher, in: ICCV, 2021, pp. 3060–3069.

[52] Q. Wu, P. Wang, C. Shen, A. Dick, A. van den Hengel, Ask me anything: Free-form visual question answering based on knowledge from external sources, in: CVPR, 2016, pp. 4622–4630.

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Rethinking the inception architecture for computer vision, in: CVPR, 2016, pp. 2818–2826.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.

[55] A. Karpathy, F.-F. Li, Deep visual-semantic alignments for generating image descriptions, in: CVPR, 2015, pp. 3128–3137.

[56] E. Cambria, Y. Li, F.Z. Xing, S. Poria, K. Kwok, SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis, in: CIKM, 2020, pp. 105–114.

[57] S. Ji, S. Pan, E. Cambria, P. Marttinen, P.S. Yu, A survey on knowledge graphs: Representation, acquisition, and applications, IEEE Trans. Neural Netw. Learn. Syst. 33 (2) (2022) 494–514.

[58] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: ACL, 2019, pp. 4171–4186.

[59] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, B. Kainz, Video summarization through reinforcement learning with a 3D spatio-temporal u-net, IEEE Trans. Image Process. 31 (2022) 1573–1586.

[60] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: CVPR, 2017, pp. 6299–6308.

[61] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017, CoRR abs/1705.06950.

[62] J. Xie, X. Chen, T. Zhang, Y. Zhang, S.-P. Lu, P. Cesar, Y. Yang, Multimodal-based and aesthetic-guided narrative video summarization, IEEE Trans. Multimed. 25 (2023) 4894–4908.

[63] S. Karen, Z. Andrew, Very deep convolutional networks for large-scale image recognition, 2014, pp. 1–14, arXiv preprint arXiv:1409.1556.

[64] K. Zhou, Y. Qiao, T. Xiang, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, in: AAAI, 2018, pp. 7582–7589.

[65] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: ECCV, 2014, pp. 505–520.

[66] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, in: CVPR, 2015, pp. 5179–5187.

[67] S.E.F. de Avila, A.P.B. Lopes, A. da Luz, A. de Albuquerque Araújo, VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognit. Lett. 32 (1) (2011) 56–68.

[68] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, H. Guan, Unsupervised video summarization with attentive conditional generative adversarial networks, in: ACM MM, 2019, pp. 2296–2304.

[69] Y.-T. Liu, Y.-J. Li, Y.-C.F. Wang, Transforming multi-concept attention into video summarization, in: ACCV, 2020, pp. 1–16.

[70] Z. Li, L. Yang, Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward, in: WACV, 2021, pp. 3239–3247.

[71] L. Yuan, F.E.H. Tay, P. Li, J. Feng, Unsupervised video summarization with cycle-consistent adversarial lstm networks, IEEE Trans. Multimed. 22 (10) (2020) 2711–2722.

[72] Y.-T. Liu, Y.-J. Li, F.-E. Yang, S.-F. Chen, Y.-C.F. Wang, Learning hierarchical self-attention for video summarization, in: ICIP, 2019, pp. 3377–3381.

[73] Z. Ji, K. Xiong, Y. Pang, X. Li, Video summarization with attention-based encoder–decoder networks, IEEE Trans. Circuit Syst. Video Technol. 30 (6) (2019) 1709–1717.

[74] J. Fajtl, H.S. Sokeh, V. Argyriou, D. Monekosso, P. Remagnino, Summarizing videos with attention, in: ACCV, 2018, pp. 39–54.

[75] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, L. Shao, Exploring global diverse attention via pairwise temporal relation for video summarization, Pattern Recognit. 111 (2021) 1–12.

[76] E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, Combining global and local attention with positional encoding for video summarization, in: ISM, 2021, pp. 226–234.

[77] M. Rochan, L. Ye, Y. Wang, Video summarization using fully convolutional sequence networks, in: ECCV, 2018, pp. 347–363.

[78] Z. Ji, F. Jiao, Y. Pang, L. Shao, Deep attentive and semantic preserving video summarization, Neurocomputing 405 (2020) 200–207.

[79] B. Mahasseni, M. Lam, S. Todorovic, Unsupervised video summarization with adversarial lstm networks, in: CVPR, 2017, pp. 202–211.

[80] B. Zhao, H. Li, X. Lu, X. Li, Reconstructive sequence-graph network for video summarization, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2021) 2793–2801.

[81] B. Gong, W.-L. Chao, K. Grauman, F. Sha, Diverse sequential subset selection for supervised video summarization, in: NIPS, vol. 27, 2014, pp. 1–9.

[82] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Summary transfer: Exemplar-based subset selection for video summarization, in: CVPR, 2016, pp. 1059–1067.

[83] J.A. Ghauri, S. Hakimov, R. Ewerth, Supervised video summarization via multiple feature sets with parallel attention, in: ICME, 2021, pp. 1–6.

[84] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkila, Rethinking the evaluation of video summaries, in: CVPR, 2019, pp. 7596–7604.

[85] D. Zwillinger, S. Kokoska, CRC Standard Probability and Statistics Tables and Formulae, Crc Press, 1999.

[86] M.G. Kendall, The treatment of ties in ranking problems, Biometrika 33 (3) (1945) 239–251.

[87] document Rec. ITU-R, Methodology for the subjective assessment of video quality in multimedia applications, 2007, pp. 1–13, BT.1788.

[88] T. Hu, Z. Li, W. Su, X. Mu, J. Tang, Unsupervised video summaries using multiple features and image quality, in: BigMM, 2017, pp. 117–120.