Multimodal-based and Aesthetic-guided Narrative Video Summarization

Jiehang Xie, Xuanbai Chen, Tianyi Zhang, Yixuan Zhang,
Shao-Ping Lu, *Member, IEEE,* Pablo Cesar, *Senior Member, IEEE,* and Yulu Yang

***Abstract*—Narrative videos usually illustrate the main content through multiple narrative information such as audios, video frames and subtitles. Existing video summarization approaches rarely consider the multiple dimensional narrative inputs, or ignore the impact of shot artistic assembly when directly applied to narrative videos. This paper introduces a multimodal-based and aesthetic-guided narrative video summarization method. Our method leverages multimodal information including visual content, subtitles and audio information through our specified key shot selection, subtitle summarization, and highlight extraction components. Furthermore, under the guidance of cinematographic aesthetic, we design a novel shot assembly module to ensure the shot content completeness and then assemble the selected shots into a desired summary. Besides, our method also provides the flexible specification for shot selection, to achieve which it automatically selects semantically related shots according to the user-designed text. By conducting a large number of quantitative experimental evaluations and user studies, we demonstrate that our method effectively preserves important narrative information of the original video, and it is capable of rapidly producing high-quality and aesthetic-guided narrative video summaries.**

***Index Terms*—Narrative video summarization, multimodal information, aesthetic guidance**

I. Introduction

Narrative videos, such as documentaries, movies and scientific explainers, share the immersive visual information along with the narrated story-telling subtitles, voiceover and background musics (BGM) [1, 2]. As the huge number of narrative videos are uploaded on various online social platforms, there is an urgent need in creating narrative video summaries which can help viewers browse and understand the content quickly, and presenting them in knowledge popularization platforms and many other applications [3, 4, 5].

Creating a high-quality narrative video summary is currently a very challenging issue. A common workflow for creating a narrative video summary manually begins with the outline writing stage [6], in which the user (video producer) writes a story outline by repeatedly watching the given video. The story outline records the narrative threads and the sequence of important events. The next stage is usually shot selection, where the user chooses the matched shots with the outline content from the given video, and decides the cut points and timing (beginning and ending time points, and shot length) for each selected shot. Experienced users intend to capture some highlight and preference/personalized shots that are not mentioned in the outline, to ensure that the generated summary is diverse enough. The last stage is the shot assembly, where the user composes the selected shots into a summary under some reasonable order. In this stage, experienced users try to balance multiple criteria based on video editing conventions and aesthetic guidelines (e.g., avoiding too short shots, including complete shots content, ensuring color continuity between adjacent shots, etc.). Thus, manually creating a narrative video summary is labor-intensive, even for experienced users.

In this context, various automatic video summarization approaches have been introduced in the research community [7, 8, 9]. However, managing to automatically produce both short and coherent summaries for long narrative videos is extremely difficult [10]. Conventional methods relying solely on visual features find it difficult in capturing narrative threads and highlights, letting alone showing a personalized visual content [11, 12]. What is more, many works rely on the strong encoding ability of neural networks to help skip the outline writing stage, and the shot selection stage is usually completed by constructing a classifier. However, these works all assume that the selected shots conform to aesthetic guidelines, thus their shot assembly is merely a process of stitching the shots together in chronological order. In practice, there are often cases where a model cannot obtain a good or complete shot in the shot selection stage. For instance, the selected shots are either too short to cover a complete voiceover or too long to be interested by the viewers, resulting in the limited quality of generated summaries. Therefore, in this work, we propose a multimodal-based and aesthetic-guided narrative video summarization framework, named MANVS, to generate high-quality video summaries. This framework selects meaningful and personalized shots based on multimodal information, and considers the completeness of shot content and the aesthetics of selected shots in the shot assembly stage, to assemble them into a video with smooth visual transitions while preserving an overall pleasing aesthetics.

We assume that the user desires to generate a condensed version of the given narrative video, which can allow viewers to acquaint a comprehensive overview of a given narrative video quickly. Moreover, the time-aligned subtitles related to the given video are assumed to be available, gathered from online or personal resources. However, there are several technical challenges need to be addressed in our approach. Firstly, the method should determine, based on the multimodal

Jiehang Xie, Xuanbai Chen, Yixuan Zhang, S-P. Lu and Yulu Yang are with TKLNDST, CS, Nankai University, China (email: {jiehangxie; 1711314; 2011432}@mail.nankai.edu.cn; {slu; yangyl}@nankai.edu.cn. Corresponding author: Shao-Ping Lu)

Tianyi Zhang and Pablo Cesar are with Centrum Wiskunde & Informatica. (email: {Tianyi.Zhang; P.S.Cesar}@cwi.nl)

Manuscript received 15 Sep, 2021; revised 23 Feb, 2022 and 6 Apr, 2022; accepted xx xxxx; date of current version April 6, 2022.

information, which shots contain important content and need to be selected. Secondly, if a user wants to select the personalized visual content from the input video and preserves it into the summary, how does the model locate this personalized content. Finally, professional videos usually have a group of features in accordance with the conventions and aesthetic guidelines of video editing, which can be utilized to distinguish them from amateur ones and make videos more visually attractive. Thus, when assembling the selected shots into a video summary, both the cinematographic aesthetics and the completeness of shots content need to be jointly considered. Note that though the completeness of shot content is preserved well, this shot may not be visual attracted enough. Similarly, an engaging shot does not necessarily guarantee containing a complete voiceover.

To address the challenges above, our proposed MANVS consists of two main modules. The first module is the multimodal-based shot selection module, which integrates visual, audio and time-aligned subtitles into the shot selection process to capture meaningful narrative content from consecutive sequences of shots. Furthermore, this module provides a flexible way to acquire the shots that users are interested in, which allows users to choose a shot by inputting text. Next, the aesthetic-guided shot assembly module firstly filters repetitive and low-quality shots, and then automatically checks whether the selected shots are content-complete. If a shot fails in these checks, through following cinematographic aesthetic guidelines and developing a series of shot completion strategies, we achieve a good trade-off between aesthetics and the completeness of shots, and finally assembles these shots into a video with smooth visual transitions.

In summary, the contributions of this paper are as follows.

- We design an aesthetic-guided shots assembly module, which establishes a series of strategies to preserve the shot content completeness and aesthetics. To the best of our knowledge, we are the first to consider aesthetic guidelines in the shots assembly stage.
- We present a multimodal-based shot selection module that comprehensively analyzes subtitle, image, and audio information to capture narrative, representative, and highlight shots. Besides, we provide a flexible way for shot selection in order that users can choose the shots they desire to watch.
- We conduct extensive quantitative evaluations and user studies to evaluate the effectiveness of MANVS. The results demonstrate that our method can generate video summaries with the quality comparable to that produced by experts and is much less time-consuming.

II. Related Work

In this section, we briefly introduce some techniques on video summarization, text summarization, language video localization, highlight detection and computational cinematography, which are most relevant to our work.

Video summarization. The main objective of general video summarization is to produce a shortened video containing the most representative visual information of a given one [14, 15]. A typical video summarization solution usually begins with selecting key frames or video segments [16, 17], which can mainly be divided into supervised and unsupervised styles. The former supposes to own human annotations of key frames in the original videos [7]. It is noticeable that constructing datasets with sufficient labels is very difficult in practice. There thus appears some unsupervised learning based approaches [18, 19]. However, although the aforementioned methods can obtain some important visual information from original videos, there are some common disadvantages. For example, some image information are just considered by searching for shot boundaries [20] in the shot selection process, where the switched shots are regarded as important content and the multimodal information of the original video are ignored. Consequently, the generated video summary loses a lot of information, which makes it look like a truncated version of the original video without coherent narrative information. The current research shows that human cognition is a process of cross-media information interaction [21]. Therefore, multimodal information such as video frames, audios, and subtitles should be leveraged to select crucial shots and provide viewers with vivid and comprehensive content. Which shots should be chosen, and how to assemble them into a video with smooth visual transitions while preserving an overall good aesthetics, deserve our attention.

Text summarization. Approaches to text summarization can be briefly classified into two categories: abstractive or extractive [22]. The former usually generates new sentences to express the crucial information [23]. However, the state-of-the-art methods of this class are likely to generate abstracts that are not fluent or introduce grammatical errors [24]. In contrast, the extractive methods focus on selecting some subsets sentences containing important contextual information from the source text and assembling them to form a text summary [25]. The main advantage of this kind of methods is that grammatical errors can be avoided for the sentences of the generated text summary. Pavel *et al.* [26] use crowdsourcing to make a text summary for the original video, and find the corresponding crucial video frames according to the content of the text summary. However, crowdsourcing is not only time-consuming and laborious, but also can not ensure the overall consistency of the summary generated by different staffs. Inspired by the above methods, we further consider that the subtitle and video shots are semantically relevant [27], and our work concentrates on automatically dividing subtitle document to extract the text summary, which effectively helps generate topically coherent video summaries.

Language video localization. The goal of language video localization is to locate a matching shot from the video that semantically corresponds to the language [28, 29] which has attracted the attention of a large number of researchers [30, 31]. Xu *et al.* [31] propose a multilevel model by doing text features injection, modulating the processing of query sentences at the word level in a recurrent neural network. Li *et al.* [32] propose a deep collaborative embedding method for multiple image understanding tasks. It is the first approach attempting to solve this issue under the framework of deep factor analysis and acquiring fruitful achievements.

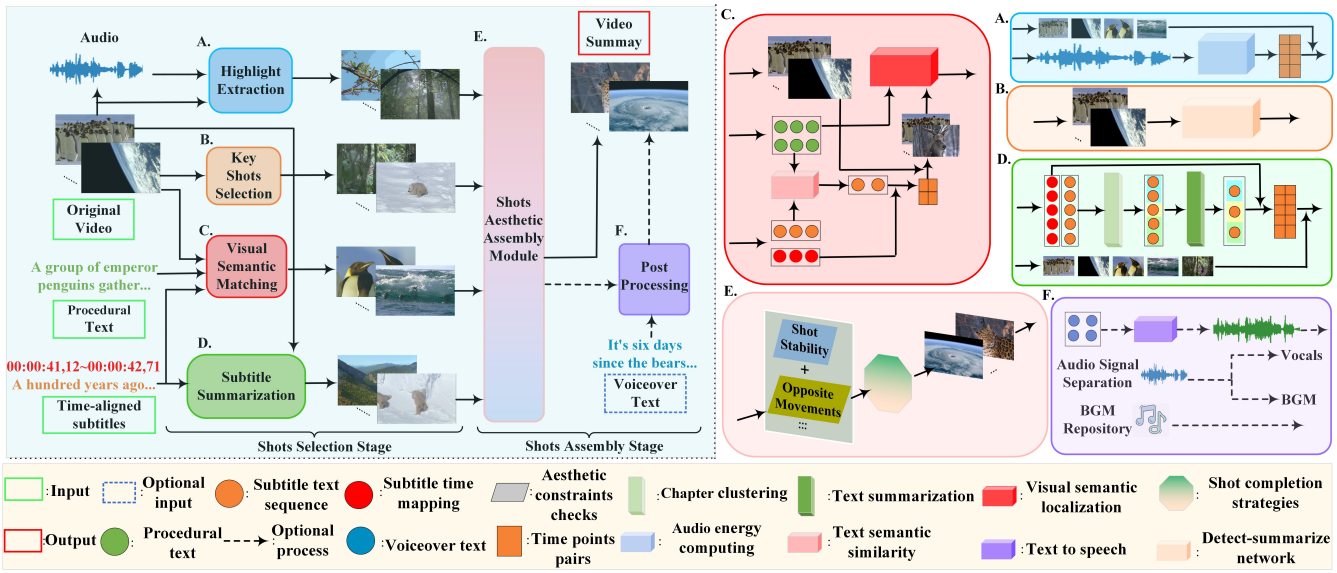

Fig. 1. The proposed MANVS framework for narrative video summarization. The upper left part is a simplified flow chart and the upper right part is the detailed processes corresponding to the annotated components on the left. The detect-summarize network is established based on [13]. Best viewed in color.

Chen *et al.* [30] introduce a semantic activity proposal which utilizes the semantic information of sentence queries to get discriminative activity proposals. To fill the gap between question-answering and language localizing, a query guided highlighting approach was proposed in [33]. In our work, we design a semantic similarity component to reduce redundant and irrelevant shots, and then leverage the language video localization method to select the personalized shots for every user in shot selection.

Highlight detection. The purpose of highlight detection is to select video clips that can attract viewer's attention [34, 35]. In the past research, the audio information is widely used in this task, as audio-based modeling is computationally easier than that of videos when representing remarkable contextual semantics [36]. Highlight shots in sports videos, e.g. goals in soccer games and home runs in baseball games, can cause the sports commentators excited voices and cheers from audience [37]. While in narrative videos such as movie and documentary, highlight shots would also cause the change of audio information [38]. The louder and denser of video sounds in a period of time, the more possibility it is wonderful [39]. Based on this observation, we take the change of video sounds into consideration for our shot selection.

Computational cinematography. A desired narrative video needs not only a fluent narration, but also shots of high-quality and artistic expression [40]. Niu *et al.* [41] discuss the aesthetic differences between professional and amateur videos by modeling image attributes including the color and noise, and video attributes such as camera motion and shots duration. Huber *et al.* [42] have some observations by analyzing a large number of Vlogs on Youtube: shots for story-telling are usually A-roll, it is reasonable to insert B-roll when the speaker stops for a long time, and B-roll usually distributes uniformly in the whole video. Wang *et al.* [43] utilize the text information to connect multiple shots, and the well preserved attributes of videos achieve that both the color continuity is ensured and the movement of shot is avoided. Hu *et al.* [44] propose to generate the informative and interesting summary using a set of aesthetic features such as saturation and brightness to select shots. Without training and human annotation, it is convenient to process videos. Our work mainly focuses on introducing the computational cinematography to automatically improve the quality of selected shots in the shots assembly stage.

III. Overview

Fig. 1 shows the overall architecture of our MANVS. The input of MANVS includes an original video, the corresponding time-aligned subtitles and the user-designed procedural text. Particularly, the time-aligned subtitles include subtitle text sequence and corresponding time maps with the video shots. We use the term of procedural text [45] to emphasize that it should focus on describing a specific visual content that users are interested in. In the shots selection stage, our method employs the multimodal-based shot selection module, which consists of subtitle summarization, visual semantic matching, highlight extraction, and key shot selection components to seek the narrative, personalized, highlight, and key shots.

In the shots assembly stage, our method passes these shots obtained by shot selection module to the aesthetic-guided shot assembly module, to obtain a good summary. Note that shot selection is an ill-posed problem, and incomplete or repeated shots could be introduced in this process. If incomplete or repeated shots appear, the overall quality would significantly degrade even if the visual aesthetic of the summary itself is satisfactory. Similarly, an engaging shot does not necessarily guarantee containing a complete voiceover. We thus design our aesthetic-guided shot assembly module so that it preserves the completeness of the shot content and the overall aesthetics. Additional, a user-designed voiceover text is allowed as an optional input to the post-processing component, to provide

customized effects for the generated summary, such as custom voiceover. Finally, MANVS outputs a high-quality video summary.

IV. Multimodal-based Shot Selection

In this section, we describe the multimodal based shot selection module that are completed the four components. All the components presented in this section output a series of shots and corresponding timelines with a beginning and ending time point pairs.

A. Subtitle Summarization

In order to increase the narration capacity for video summarization, we design a subtitle summarization component. It is comprised of two cascaded components: chapter clustering component and text summarization component.

Fig. 2 illustrates the whole pipeline of subtitle summarization component. The input data includes the given narrative video and the corresponding time-aligned subtitles. Firstly, the chapter clustering component, which utilizes the term frequency inverse document frequency (TF-IDF) [46] and K-means algorithm, automatically organizes the storytelling structure of the input subtitle text sequence and divides it into different chapters. Next, the text summarization component is conducted in every single chapter to generate text summary. Finally, our method extracts narrative shots from the input video according to the time mapping corresponding to every generated text summaries.

Formally, the time-aligned subtitle of a narrative video is represented by a two-tuple (S, T), where $S = \{s_1, s_2, ...s_n\}$ denotes the subtitle text sequence, s_i represents the $i-th$ subtitle in the text, $T = \{t_1, t_2, ...t_n\}$ is the time mapping, $t_i = (b_i, e_i)$, b_i and e_i respectively denotes the start and end time point of s_i, n is the total number of sentences in S. The objective of the chapter clustering component is to divides S into a chapter sequence $D = \{d_1, d_2, ...d_m\}$. Here m is the total number of chapters d, which can be defined as follows:

$$D = \varsigma\left(\vartheta\left(S\right)\right), \quad s.t. \begin{cases} \sum_{h=0}^{m} d_h = \sum_{i=0}^{n} s_i, \\ m << n, \end{cases} \tag{1}$$

where $\vartheta\left(\cdot\right)$ represents the TF-IDF similarity score, $\varsigma\left(\cdot\right)$ denotes K-means clustering. In this process, we use the NLTK toolkit to exclude 'or', 'the', and other stop words. Our chapter clustering component ensures that the text of every chapter describes the coherent stories of such chapter.

After that, a pretrained extractive text summarization method [47] is employed as the backbone of our text summarization component, aiming to obtain the representative context semantic information and meaningful narrative clues. Concretely, this component leverages the robust transformer encoder [48] to map a chapter to a semantic feature space. Then, this component uses the auto-regressive decoder pointer network with attention mechanism [49] to extract a subset to form text summary from each chapter. This process and objective function can be formulated as follows:

$$r\left\{s_b, s_o, ...s_q\right\} = \lambda\left(\varepsilon\left(d\right)\right), \tag{2}$$

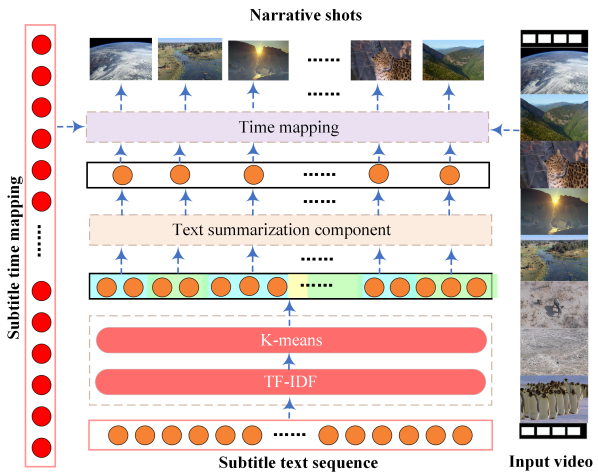

Fig. 2. The pipeline of the subtitle summarization component. The green, yellow, and blue blocks represent different chapters. The text summarization component is established based on [47].

$$P(r|d) = \underset{s_q}{argmax} \prod_{b}^{q} P\left(s_q|\{s_b, ..., s_{q-1}\}, d\right), \tag{3}$$

where $\{s_b, s_o, ...s_q\}$ denotes the sentences that form a text summary r, the first and last sentence of r is represented by s_b and s_q, respectively. $\lambda\left(\cdot\right)$ means the transformer encoder, $\varepsilon\left(\cdot\right)$ denotes pointer network decoder, P is the probability.

Finally, our subtitle summarization component searches narrative shots corresponding to the time mappings t from the given video based on r, ensuring that the video summary can effectively cover those significant narrative information of the input video.

B. Visual Semantic Matching

To find some specific shots that users might be delighted to watch, we construct a visual semantic matching component to search for those shots that match the user-designed procedural text. Our visual semantic matching component is consisted of two cascaded components, which are text semantic similarity component and visual semantic localizing component.

Fig. 3 presents the workflow of visual semantic matching component. The input data includes the given video, the time-aligned subtitles and the user-designed procedural text. By calculating the semantic similarity weight between procedural text and the text of all subtitles, the text semantic similarity component obtains some subtitles semantically related to the procedural text. This component then creates a sub-video by extracting corresponding shots from the input video based on the time mapping of these subtitles. The semantic similarity component is designed to reduce redundant and irrelevant shots. Next, the sub-video and procedural text are fed into the visual semantic localizing component. Finally, in the generated sub-video, this component locates the shot that corresponds to the scene described by the procedural text.

The flow of our text semantic similarity component is as follows. Firstly, it computes the word co-occurrence for procedural text x and every subtitle text y_i, where $i \in \{1, 2, ...n\}$.

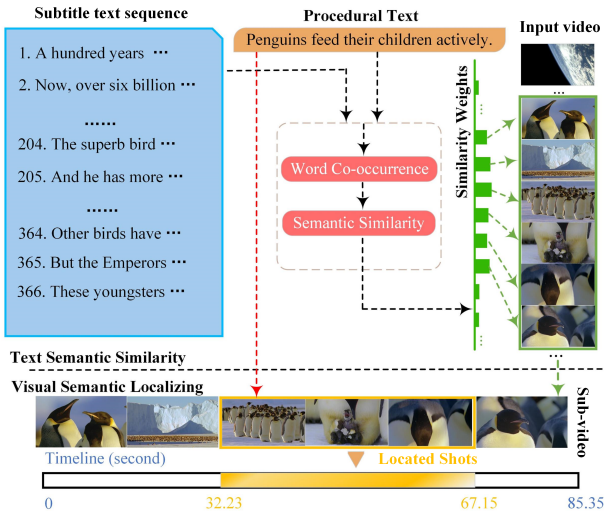

Fig. 3. The workflow of visual semantic matching component. Generated sub-video and located shots are outlined in green and yellow, respectively. The visual semantic localizing is established based on [50].

This is to measure the number of same words for every text-subtitle pair. Secondly, it selects the top k_1 subtitles which own relatively high word co-occurrence rate. Then, LSTM is employed to extract the semantic features for text x and k_1 subtitles, and the semantic similarity weight for every selected text-subtitle pair is calculated. This assists us to find k_2 subtitles owning high weights from k_1, and we take them as the final candidates. Subsequently, a sub-video is created by assembling video shots extracted from input video corresponding to k_2 subtitle candidates. In order to increase the story completeness of the sub-video, we lengthen the duration of each subtitle, some content before the beginning (for l_1 seconds) and after the end (for l_2 seconds) are also included in such sub-video. This enables us to involve some more supplementary shots (B-roll), and in our work l_1 and l_2 are both set as 6 according to detailed B-roll analysis provided in [42].

Formally, our text semantic similarity component is defined as:

$$SC = T_{k_2}(\Gamma_{i=0}^{g} \frac{LSTM(x) * LSTM(y_i)}{\|LSTM(x)\| \, \|LSTM(y_i)\|}), \tag{4}$$

where SC denotes the final subtitle candidates, $T_{k_2}(\cdot)$ represents selecting top k_2 subtitles which are similar to the procedural text x. $\Gamma_{i=0}^{g}$ is the value range of i from 0 to g, the operator '$*$' means dot product, and $\|\cdot\|$ represents computing the norm of the vector. Here g is the number of selected subtitles G, which are obtained by

$$G = T_{k_1}(\Gamma_{j=0}^{n} W(x, y_j)), \tag{5}$$

where $T_{k_1}(\cdot)$ represents selecting top k_1 subtitles which are similar to the procedural text x, and n is the total number of subtitles in a whole narrative video. $W(\cdot,\cdot)$ means the computing method of word co-occurrence. k_1 and k_2 are set as 12 and 6, respectively.

Since the procedural text is a set of semantic phrases combination $F(\cdot)$, its different levels of semantic information can be used to match the visual feature of the created sub-video, and the visual content can be located by the text. Therefore, the objective of our localization task is defined as maximizing an expected log-likelihood:

$$\theta^* = \underset{\theta}{argmax}\, \mathbb{E}\left\{logp_\theta \left[SV\,(C)\,|F\,(P)\right]\right\}, \tag{6}$$

where θ means the parameters that need to be optimized, SV denotes our generated sub-video, C represents a time interval of the target region within SV, and P is the procedural text.

Our visual semantic localizing component based on the state-of-the-art temporal language grounding method, i.e. LG4 [50]. This method firstly uses a sequential query attention network [51] to explore the sequence relationships between sentences. Next, it takes video-text interaction in three different levels, where the first level is the segment-level modality fusion. This encourages that the segment features relevant (or irrelevant) to the semantic phrase features should be highlighted (or suppressed). The second level interaction is for local context modeling, where the neighbors of individual segments are considered by leveraging the Hadamard product [52]. The last level interaction deals with contextual and temporal relations between semantic phrases by employing the non-local block presented in [53]. After that, the shots that are the most semantically related to the procedural text are located and selected from the sub-video.

C. Crucial shot selection

To effectively capture those representative and highlight shots, we design the key shot selection and highlight extraction components.

Key shot selection: Here we use a pretrained detect-summarize network [13] to achieve this purpose by only inputting the given video. The temporal consistency is formulated to ground the representative contents of the given video in this approach. In detail, our method firstly samples a series of temporal interest proposals with different scales of intervals. After that, long-range temporal features are extracted for both predicting important shots and selecting the relatively representative ones. Finally, a set of correlated consecutive frames within a temporal slot are considered for shot selection.

Highlight extraction: Audio owns remarkable representation of the corresponding semantic content, and its processing is computationally easier than that of video, so it has been widely used in the area of highlight extraction. Following the above observations, here we utilize the fluctuation of the sound energy as a supervised prior to extract highlight shots and the sound energy is determined by the volume level of the audio track in video. The input data includes the given video and the audio extracted from the video. Firstly, we divide the audio into clips with the same length and then compute the value of sound energy for those clips. Then, some clips with larger audio energy are selected. Finally, desired shots are extracted from the video based on the time mapping corresponding to the selected audio clips. Formally, it is constructed as follows:

$$HS = T_x(\Gamma_{k=0}^{l}(\sum_{i=0}^{w} E_{k+i}^2)), \tag{7}$$

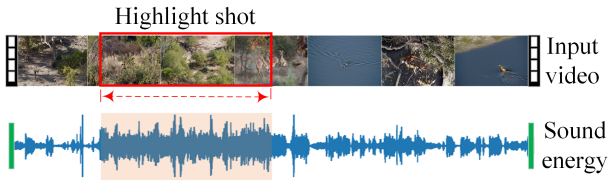

Fig. 4. A example of highlight extraction. The audio segment with higher sound energy and the corresponding highlight shot are marked with red block and line respectively.

where HS is the desired highlight shots to be selected, $T_x(\cdot)$ denotes top x (we fix it as 10) percent of the calculated sound energies of all audio clips, Γ_k means the value range of k, and l is the duration of the video. Suppose E_k is the value of the audio signal in time k, for each audio clip from time k to $k+w$, w is the clip size (e.g. 5 seconds), then $\sum_{i=0}^{w} E_{k+i}^2$ is the value of the sound energy of such clip. Fig. 4 shows a example of highlight extraction.

V. Aesthetic-guided shot assembly

Although some works successfully apply aesthetic guidelines to the video summarization task [44], they only take the brightness and saturation features of frames as a supervisory signal to select shots in the shot selection stage, while ignoring maintaining the completeness of the shot content and following the aesthetics guidelines in the shot assembly stage. Thus, existing methods cannot be directly applied to solve the following issues: (1) Whether the aesthetic quality of the selected shot itself (e.g., shot length and shot stability) meets the cinematographic aesthetic guidelines, or further optimization is needed. (2) Whether there is an overlap between the timelines of the selected multiple shots, or filtration is needed. (3) Whether the timeline of a selected shot with a beginning and ending time point pair is perfectly in accordance with a complete narrative subtitle or content, or completion is needed. (4) How to solve the above situations while preserving an overall pleasing aesthetic feeling as well as the smooth visual transitions among consecutive shots.

We believe that the quality of a single shot itself is highly correlated with the overall one of the summary, i.e., both aesthetic and content integrity factors of each shot can affect the overall quality. For example, for a narrative video summary, most of the attention should be paid to narrative shots and the voiceover. If the meaningful narrative shots and voiceover are incomplete (cut off in spoken sentences) or repetitive, the overall quality would significantly degrade even if the visual aesthetic of the shots themselves is satisfactory. Similarly, shots that are complete but do not conform to aesthetic guidelines (such as too long shots) could result in a limited quality of the generated summary. We thus establish the aesthetic-guided shot assembly module to effectively solve the issues above. To the best of our knowledge, we are the first to use aesthetic guidelines and keep the content completeness of the selected shots in the shots assembly stage.

Specifically, this module takes the shots selected from the previous module as input, and refers to the classic cinematographic aesthetic guidelines and time-aligned subtitles, to preserve the aesthetic quality and content integrity. The guidelines provide a set of predefined constraints for ensuring the aesthetic quality, while the time-aligned subtitles serve as reference cutting points for selecting complete narrative shots.

Firstly, we make a simple shots aesthetic evaluation for those selected shots. Though converting the shots aesthetic evaluation into a regression task is a straightforward solution, it brings great difficulties in training the model. The reason lies in the subjectivity to determine whether a shot is aesthetic, and the annotations are generally not available. To simplify the problem, our method automatically checks whether the selected shots meet some simple predefined aesthetic constraints, such as shot stability, opposite movement. If a shot fails these simple checks, we abandon it to avoid introducing low-quality shots. Secondly, our method checks the content redundancy and completeness of those remaining shots. Generally, if there is an overlap between the timelines of multiple selected shots, it is considered as redundant. If the shot timeline with the beginning and ending time points can be aligned with a complete subtitle, the content of the shot is regarded as complete, which can bring the viewer an enjoyable audio-visual experience. Otherwise, incomplete shots can make some cut off for spoken sentences appearing in the summary. Once a shot fails the above checks, our method analyses its eight possible situations and leverages corresponding strategies of shot complement, to remove the redundant shots as well as ensure the completeness of the content. Meanwhile, our method automatically expands or filters shots so that the selected shots satisfy some predefined aesthetic constraints, such as color continuity and shot length. In this way, MANVS achieves a good trade-off between ensuring the completeness of shots and preserving an excellent aesthetic quality of the generated video. Our method preserves smooth visual transitions among consecutive shots, by automatically adjusting the saturation and brightness of adjacent shots.

A. Cinematographic aesthetic constraints

Considering that our focus is to generate summaries of professional narrative videos, some cinematography rules are not always applicable for our task. For instance, for some specific artistic expressions or authenticity, shots with low saturation and brightness do not necessarily make a low-quality video. Therefore, we select several classical aesthetics guidelines which are suitable for our task, and explain below how to apply them in our setting:

Shot Stability. The high-quality video shots should move smoothly and stably. The lower the local acceleration of the shot content is, the more stable the shot is, and conversely, it will reach the opposite [54]. In our setting, we calculate the local shake value of a shot according to the homography transformation matrices from some consecutive sampled frames in the shot:

$$F(f_i) = ||H(f', f'')\, p_{f'}(i) - p_{f'}(i) - H(f, f')\, p_f(i)\,||_2, \quad (8)$$

$$F_{SS}(s) = -\frac{1}{4N_s^f} \sum_{f \in s} \sum_{i}^{4} F(f), \quad (9)$$

where f, f' and f'' represent three consecutive frames in the shot s. The symbol f_i is the corner i of f, $p_f(i)$ is the position f_i in pixels, where $i = 1, ..., 4$. We use $H(f, f')$ to denote the homography transformation matrix between f and f'. $F(f_i)$ means the local shake values of f_i. The shot stability F_{ss} is computed as the negative average of local shake values over time, and N_s^f is the number of sampled frames in shot s, and the sampling step is 8.

Opposite Movements. Adjacent shots with opposite camera movements may result in a terrible viewing experience for the audience [43]. We avoid this situation by calculating the two-dimensional motion of the shots:

$$F_{om}\left(f_l, \widetilde{f}_f\right) = \sum_{i=1}^{4} \frac{\rho(f_l,i)\rho\left(\widetilde{f}_f,i\right)}{|\rho(f_l,i)|*|\rho\left(\widetilde{f}_f,i\right)|}, \tag{10}$$

$$\rho(f_l, i) = \varpi(i) - H\left(f_l, \widetilde{f}_f\right)\varpi(i), \tag{11}$$

where f_l and $\widetilde{f}_f$ respectively represent the last and the first frame of two adjacent shots s and $\widetilde{s}$. F_{om} denotes the cosine distance of the frame corner movement vectors of f_l and $\widetilde{f}_f$. $\rho(f_l, i)$ is the function estimating the two-dimensional movement of the i-th corner $\varpi(i)$ of f_l. $H\left(f_l, \widetilde{f}_f\right)$ is the homography transformation matrix between f_l and $\widetilde{f}_f$.

Shots length. Some long shots without interesting content may easily let the viewer lose attention, and oppositely, some too short shots may affect visual smoothness. In order to avoid these extreme shots, we set the duration of individual shots as 3 to 8 seconds inspired by [43].

B-roll selection. In professional narrative videos, those A-roll and B-roll shots are reasonably mixed [42]. For telling stories, most shots are usually A-roll, and B-roll is usually placed at the speaker's natural pause to support A-roll in visual experience. In this work, we can take the shots with and without subtitles as A-roll and B-roll, respectively. It is noticeable that the frequent insertion of B-rolls would easily distract the audience, while too few b-rolls may make the story much less interesting. Therefore, we set the interval between two B-rolls as 9 seconds motivated from [42].

Color continuity. The video color continuity is often the most representative feature in identifying professional videos [41]. Therefore, preserving the continuity of saturation and brightness in a shot is crucial to improving the viewing experience. In our work, the color continuity between two adjacent frames is measured by the histogram difference of the saturation and brightness:

$$F_{cc}(e, b) = \frac{1}{2}\left(\Psi\left(\eta_S(e), \eta_S(b)\right) + \Psi\left(\eta_L(e), \eta_L(b)\right)\right), \tag{12}$$

where $F_{cc}(.)$ denotes the tonal difference between two shots, e and b respectively represent the last and the first frame of adjacent shots, $\eta_S(.)$ and $\eta_L(.)$ represent the S-channel and L-channel histogram of this frame, both quantified to 256 bins. Ψ means the chi-square measure.

B. Shots Assembly and Post-processing

Based on aforementioned classical cinematographic aesthetic guidelines, it becomes practical for us to appropriately place the beginning and ending points of the selected shots, remove some repeated shots and extend incomplete ones. Specifically, our method automatically checks whether the selected shots meet the stability and opposite movements constraints, and then preserves the shots which pass the checks. Furthermore, our method checks whether each preserved shot is long enough to contain the duration of a complete voiceover based on the time-aligned subtitles. We summarize three types, adding up to eight possible situations and corresponding strategies of shot complement as shown in Fig. 5, to make the preserved shots contain complete narrative content while satisfying aesthetic constraints such as shots length and color continuity. Now we provide more details of such strategies.

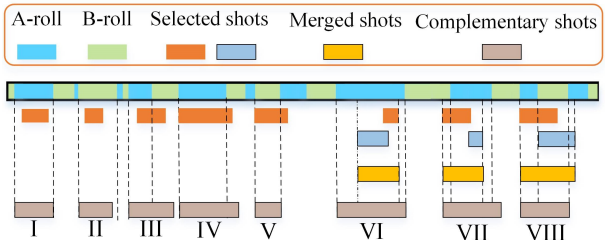

Fig. 5. Eight possible shot situations and corresponding shot complement strategies. The long bar outlined by black is the timeline of video.

(1) Incomplete and non-overlapping shot. In this type, the selected shot is within a either A-roll or B-roll. Moreover, for selected shots their timelines do not overlap with each other. Therefore, there would be two situations. (I) Incomplete A-roll (ξ_A): it should be lengthened such that its timeline is equal to that of a complete A-roll (δ_A), or we directly remove it once the timeline ratio of ξ_A to δ_A is smaller than a constant threshold (e.g. 0.5). (II) Incomplete B-roll (ξ_B): we extend this kind of shot according to the color continuity. For each iteration, if the difference of color continuity between the beginning frame and its previous frame is not great, the beginning time point is updated to the previous frame. Similarly, the end time point is processed by comparing the end frame and its next frame. The extreme case is extending ξ_B to a complete B-roll (δ_B).

(2) Across boundary and non-overlapping shots. For each beginning and ending time points pair of such shots, they are located within two complete shots and do not overlap with each other. This type can be simply divided into three cases: (III) $\xi_A \cup \xi_B$, (IV) $\delta_A \cup \xi_B$, and (V) $\xi_A \cup \delta_B$. However, selected shots can be decomposed into separate sub shots which exactly follow (I) or (II) in all these cases.

(3) Overlapping shots. This type contains several cases below: (VI) the overlapping parts are in a complete shot, (VII) at least one overlapping shot is across the boundary of complete shots, and (VIII) at least an overlapping shot contains a complete shot. For these shots, we first remove the repeated parts, and then merge the consecutive shots into a single one. This merged shot then conforms to the above processing method of type (1) or (2).

Similarly, the beginning and ending time points pair that is located within more than two complete shots can firstly be separated by the complete shot boundary, and then be

complemented by leveraging corresponding strategies. Finally, we select the qualified shots according to the aforementioned aesthetic constraints, and the resulting shots are assembled to get the video summary. In particular, if the color continuity between two shots is quite different, we employ fade-in and fade-out effects to ensure visual smoothness.

Post processing: we further apply a series of post processing effects, it is worth noting that the operations 2) and 3) are optional: 1) we leverage the Spleeter [55] to extract the original BGM and only keep the human voice. 2) We automatically select a coherent audio clip from the extracted BGM that matches the style of the generated summary. We also allow users to select some externally BGM for the summary. 3) We allow users to manually design a text of voiceover and select the start and end time that they desire to insert. We leverage the text-to-speech method, i.e. espnet [56], to implement it.

VI. Experiments

In this section, we describe the datasets, evaluation metrics, baselines, experimental designs and manage to evaluate the effectiveness of our approach by answering the following research questions (RQs):

RQ1: Compared to traditional video summarization methods, can our method provide a more high-quality summary?

RQ2: Does every shot selection component or aesthetic constraint have positive influence to our final result?

RQ3: In contrast to professional manual summarization, what are the advantages and disadvantages of MANVS?

A. Datasets

We employ several existing and collected datasets to conduct extensive experiments. These datasets are shown below.

(1) *TVsum* [57]. It is a widely used and manually annotated video summarization dataset, which contains 50 multimodal videos from public websites. The topics of videos include how to change tires for off road vehicles, paper wasp removal etc., with durations varying from 2 minutes to 10 minutes. In our experiments, Aliyun interface[1] is used to automatically generate subtitle documents for these videos.

(2) *MPII movie description (MPII)* [58]. It is a popular video description dataset in language video localization field, containing 94 movies and 68375 manually written description sentences of movie plot.

(3) *Documentary description*. We collected 72 documentaries with different themes such as earth, travel, adventure, food, history and science from public websites, and annotated 21,643 temporal descriptions of plot like [58].

(4) *Movie-documentary (M-D)*. This is a video repository that consisted of the movies in MPII and the documentaries in documentary description dataset. We further collect their corresponding subtitles from public websites. Unlike TVsum, M-D does not provide frame-level importance scores.

[1]https://www.aliyun.com/

B. Baselines

We compared the performance of our model with several advanced video summarization methods. *1) Random*: we labeled scores of importance for each frame randomly to generate summaries which are independent with the video content. *2) DR* [18]: it proposes an encoder-decoder framework based on reinforcement learning to predict probabilities for every video frames. For training the framework, a diversity and a representativeness reward function is designed to generate summaries. *3) DR-Sup* [13]: it is a ablation version of DR, which has the same model backbone as DR, but only utilizes a representativeness reward function in the training process. *4) VAS* [7]: it proposes a self-attention based network, which performs the entire sequence to sequence transformation in a single feed forward pass and single backward pass during training. *5) DSN-AB* [13]: it firstly samples a series of temporal interest proposals with different scales of intervals. After that, long-range temporal features are extracted for both predicting important frames and selecting the relatively representative ones. *6) DSN-AF* [13]: it is a variant of DSN-AB. The feature extraction and key shot selection steps of this method are the same as those of DSN-AB, but the importance scores at frame level are converted into shot level scores. *7) HSA* [8]: it proposes a framework with two layers. The first layer is to locate the shot boundaries in the video and generate the visual features for them. The second layer is utilized to predict which shots are most representative to the video content. *8) FCN* [59]: it adapts semantic segmentation models based on fully convolutional sequence networks for video summarization. *9) HMT* [60]: it proposes a hierarchical transformer model based on audio and visual information, which can capture the long-range dependency information among frames and shots. *10) VSN* [61]: it proposes a deep learning framework to learn video summarization from unpaired data.

C. Experimental Designs

In order to ensure the comparison as fair as possible, we set our method as the ones with and without manual operation, which are denoted by MANVS and MANVS-auto, respectively. Specifically, the manual operation includes visual semantic matching component, user-designed voiceover and BGM in the post processing component. Next, the following experiments are designed to answer aforementioned RQs:

Comparison to traditional video summarization methods (RQ1). We compared MANVS-auto with baselines in the following two experimental designs. R1a: we conduct the comparison experiments on the TVsum dataset, by quantitatively evaluating the performance of the proposed method and the comparison methods. R1b: we conduct user study on the comparison experiments for the TVsum and M-D dataset, and further explore the statistical significance of the user data.

Ablation study (RQ2). We conduct two types of ablation studies. R2a: we conduct a quantitative experiment on MPII and documentary description dataset to evaluate the performance of our visual semantic matching component objectively. In this experiment, we compare the employed method LG4 and an alternative method VSL [33] for visual-semantic localizing.

Besides, we compare them against those without either text semantic similarity (TSS) component, without visual semantic localizing component, or both. R2b: we evaluate the effect of every single component and aesthetic constraint to the quality of our generated video summary on the TVsum and M-D datasets. For the former dataset, the experiments includes quantitative evaluations and user study, while for the latter, due to the lack of annotations, only user study is conducted. Specifically, for TVsum, we achieve which MANVS is applied without a specific individual component or aesthetic constraint: without subtitle summarization (w/o SS), without key shot selection (w/o KSS), without highlight extraction (w/o HE), without shots length (w/o SL), without B-roll selection (w/o BS), without color continuity (w/o CC), without shot stability (w/o SSA), without opposite movements (w/o OM) and without post processing (w/o PP). Since the videos in TVsum are unlabeled with temporal descriptions, we do not use VSM component in this experiment on TVsum. Similarily, we achieve which MANVS is applied, keep the same setting on M-D dataset.

Comparison to professional manual editing (RQ3). R3: we compare the editing time and quality of generated summaries among MANVS, MANVS-auto and an experienced video producer who creates a video summary manually. Here the manual editing tool is the commonly used Adobe Premiere©, and the producer is asked to make a video summary for test videos. The summary should also contain a coherent BGM, and some simple splicing and fading effects could be used to make the summary visually appealing. To make a fair comparison, we only counted the human active time during producing.

Implementation details. In the user study of R1b, R2b and R3, we randomly select 10 movies, 10 documentaries and 10 videos from M-D and TVsum datasets to generate summaries. In order to keep the consistency of the evaluations, we refer to [43], randomly pick a movie (*Big Fish*), a documentary (*Planet Earth Season I Episode 2*) and a video from TVsum (*Poor Man's Meals: Spicy Sausage Sandwich*) for demonstrating results. In the quantitative experiments, the TVsum, MPII and documentary description datasets are randomly divided into training and test dataset with the ratio of 8:2. For the extractive text summarization method [47] leveraged in the text summarization component, we adapted the model pretrained on the CNN/DailyMail [62]. The official code and pretrained model are available online[2]. This model employs a pretrained uncased base model of BERT as transformer encoder, and utilizes a pointer network with attention mechanism as decoder. The number of transformer blocks and self-attention heads are both 12, the hidden layer size is 768, the maximum input sequence length is 512, the batch size is 32 and the vocabulary size is 30000. For the temporal language grounding method [50] leveraged in the visual semantic localizing component, the official code is available online[3]. We follow the official training setup and train the model on the Charades dataset [63], which is composed of 12408 and 3720 time interval and text query pairs in training and test set, respectively. This model utilizes I3D [64] to extract segment features for training data, while fixing their parameters during a training step. The feature dimension is set to 512. This method uniformly samples 128 segments from each video and uses the Adam optimizer to learn models with a mini-batch of 100 video-query pairs and a fixed learning rate of 0.0004. Then, we fine-tune the pretrained model on the MPII movie and Documentary Description dataset. Parameters of the pretrained model are fixed during training. For the detect-summarize network in the key shot selection component, we use the pretrained anchor-base model provided by [13]. This model includes a multi-head self-attention layer with 8 heads, a layer normalization, a fully-connected layer with a dropout layer and a tanh activation function, followed by two output fully-connected layers. In our setting, all the hyperparameters of implemented methods are kept the same with the official ones. We evaluate the performance of the implemented method with the same evaluation criteria as the official ones, and the performance of these methods is also consistent with official ones. The supplementary material provides partially generated summaries, and we encourage readers to watch these videos.

[2]https://github.com/maszhongming/Effective_Extractive_Summarization
[3]https://github.com/JonghwanMun/LGI4temporalgrounding

D. Evaluation Metrics

In order to comprehensively evaluate the performance of our framework, we adopt different evaluation metrics for different experimental designs. A detailed version can be seen in supplementary material.

1) In the quantitative evaluations of R1a and R2b for TVsum, commonly used [9] *Precision*, *Recall* and *F-score* are utilized as the evaluation metrics to evaluate the quality of generated summaries.

2) In experiments R2a, we adopt *R@n, IoU* = μ and *mIoU* as the evaluation metrics, which is commonly used in the field of language video localization [33, 50]. *R@n, IoU* = μ represents the percentage of testing samples which have at least one of the top n results with *IoU* larger than μ. *IoU* means intersection of time between visual semantic matching and ground truth on union. *mIoU* means the average *IoU* over all testing samples. In this paper, following [28, 29], when reporting *R@n*, IoU = μ, we set n as 1 and $\mu \in \{0.3, 0.5\}$, and when reporting mIoU, we set $\mu \in \{0.1, 0.3, 0.5, 0.7\}$.

3) In the user study of R1b and R3, 100 participants were invited to watch the involved video summaries, and they were asked to rate every video by using a 7-point Likert scale (1 = poor, 7 = excellent), taking *visual attraction* (VA) and *narrative completeness* (NC) into account. The NC reflects viewers' subjective perceptions of the narrative coherence and content integrity of the generated summaries, while the VA reflects viewers' viewing experience. Before rating, we explain the whole procedure of our model and present some example videos which are with incomplete shots and do not meet aesthetic guidelines to participants. When displaying these videos, we explain the definition of NC and VA to each participant for a more precise understanding. Videos which they require to rate are not included in these examples, and concepts are emphasized when they start rating videos.

TABLE I
THE STATISTICS FOR SUMMARY ATTRIBUTES OF SUMMARIES GENERATED BY BASELINES AND MANVS-AUTO. THE IC, CON, TNS AND TAL RESPECTIVELY REPRESENT INFORMATION COVERAGE, CONSISTENCY, THE NUMBER OF SHOTS IN THE SUMMARY AND THE AVERAGE LENGTH OF SHOTS.

Method	TVsum				Movie				Documentary			
	IC	CON	TNS	TAL	IC	CON	TNS	TAL	IC	CON	TNS	TAL
Random	6/10	5/14 (35.71%)	16	3.7	25/32	17/212 (8.02%)	623	1.8	17/18	3/68 (4.41%)	234	1.9
DR [18]	6/10	5/15 (33.33%)	17	3.5	23/32	26/232 (11.21%)	681	1.6	17/18	5/97 (5.15%)	237	1.9
DR-Sup [18]	6/10	6/15 (40.00%)	16	3.7	23/32	24/208 (11.54%)	643	1.6	17/18	2/64 (3.13%)	229	2.0
VAS [7]	5/10	6/17 (35.29%)	14	4.3	24/32	29/218 (13.24%)	612	1.8	17/18	6/74 (8.11%)	202	2.2
DSN-AB [13]	5/10	6/18 (33.33%)	16	3.7	23/32	31/210 (14.76%)	580	1.9	16/18	6/82 (7.32%)	226	2.2
DSN-AF [13]	5/10	4/14 (28.57%)	17	3.5	22/32	28/223 (12.56%)	623	1.8	16/18	4/72 (5.56%)	236	1.9
HSA [8]	3/10	6/9 (66.67%)	7	9.4	16/32	77/114 (67.54%)	115	9.3	8/18	30/43 (69.77%)	30	17.1
FCN [59]	5/10	6/15 (40.00%)	16	3.8	24/32	35/206 (16.99%)	621	1.8	16/18	4/70 (5.71%)	232	1.9
HMT [60]	6/10	7/13 (53.85%)	15	4.0	18/32	72/165 (43.64%)	628	1.8	17/18	4/74 (5.41%)	224	1.9
VSN [61]	6/10	7/16 (43.75%)	16	3.8	22/32	21/211 (9.95%)	553	2.1	15/18	1/64 (1.56%)	214	2.1
MANVS-auto	8/10	12/12 (100%)	15	4.1	25/32	188/188 (100.00%)	256	4.2	17/18	56/56 (100.00%)	72	6.2

TABLE II
THE USER STUDY OF VIDEO SUMMARIES IN THE QUALITY OF VISUAL ATTRACTION (VA) AND NARRATIVE COMPLETENESS (NC) GENERATED BY STATE-OF-ART METHODS AND MANVS-AUTO.

Method	TVsum		Movie		Documentary	
	VA	NC	VA	NC	VA	NC
Random	3.4	3.6	1.9	1.7	2.1	1.9
DR [18]	3.3	3.5	1.8	1.9	1.7	2.0
DR-Sup [18]	3.4	3.3	2.1	1.8	1.9	2.1
VAS [7]	3.7	3.6	2.0	1.9	2.2	1.8
DSN-AB [13]	3.5	3.4	1.7	2.1	1.7	1.8
DSN-AF [13]	3.3	3.6	1.8	1.8	2.0	2.1
HSA [8]	4.3	4.4	3.5	3.6	3.7	3.7
FCN [59]	3.3	3.6	2.3	1.8	2.0	1.8
HMT [60]	3.7	3.7	3.8	3.6	1.9	1.9
VSN [61]	3.3	3.8	2.0	2.1	1.6	1.7
MANVS-auto	**4.7**	**4.6**	**5.0**	**5.2**	**5.1**	**4.7**

TABLE III
THE STATISTICAL SIGNIFICANCE (P-VALUE) WHICH IS THE IMPROVEMENT OF MANVS-AUTO OVER OTHER METHODS IN TERMS OF VA AND NC (WILCOXON TEST). NUMBERS FROM 1 TO 10 CORRESPOND FROM THE RANDOM TO VSN METHOD IN TABLE II RESPECTIVELY.

	TVsum		Movie		Documentary	
	VA	NC	VA	NC	VA	NC
1	1.9e-18	2.1e-18	1.3e-18	9.3e-19	1.5e-18	1.9e-18
2	1.9e-18	2.0e-18	2.0e-18	2.0e-18	1.3e-18	1.9e-18
3	2.3e-18	9.5e-19	2.0e-18	2.3e-18	1.8e-18	2.2e-18
4	6.4e-18	1.8e-18	6.4e-18	1.9e-18	2.3e-18	1.9e-18
5	2.9e-18	1.5e-18	2.9e-18	1.9e-18	1.5e-18	1.7e-18
6	2.0e-18	1.7e-18	6.4e-18	1.8e-18	2.0e-18	2.1e-18
7	4.1e-13	9.2e-7	5.3e-16	1.6e-15	2.8e-18	1.9e-18
8	3.2e-18	1.9e-18	3.2e-18	2.1e-18	1.4e-18	1.8e-18
9	4.4e-18	1.9e-18	1.1e-16	9.1e-16	1.7e-18	1.7e-18
10	1.8e-18	1.3e-17	6.1e-18	5.3e-18	9.3e-19	2.0e-18

Videos were displayed in full-screen mode on calibrated 27-inch LED monitors (Dell P2717H). Viewing conditions are in accordance with the guidelines of international standard procedures for multimedia subjective testing [65]. The subjects are all university undergraduate or graduate students with at least two years experience in image processing, and they claimed to browse videos frequently. The percentage of female subjects is about 40%. All the subjects are aging from 20 to 27 years old. Before giving the final rating, we allowed participants to watch each video for multiple times. Besides, we have taken the time needed to make video summaries by different methods into account in R3. Similarly, we asked 100 participants to rate each video according to the given requirements in the user study of R2b.

4) For further investigating the quality of generated summaries, we analyse some desirable attributes of them in experiments R1b following [17]. Specifically, our study ask 100 participants to first read a text summary for the input video. The text summaries of the TVsum videos are made manually, while those of the documentaries and movies are taken from Wikipedia. Subsequently, they were asked to watch the generated video summaries and determine whether there was a particular plot or entity in the video that the text summaries described. Participants answered with "Yes" if they were certain it was present in the video, "No" if the event was absent. Afterwards, we compute the *information coverage* (IC), i.e. the ratio of the selected entities or plots to the total number of representative ones. This measures how much representative information is retained in the summary from the input video. Furthermore, we record the ratio of complete sentences and emerging sentences read by voiceover in the summary, so as to acquire *consistency* (CON). Intuitively, videos in high consistency can provide viewers with a enjoyable viewing experience. Finally, we counted *the number of shots in the summary* (TNS) and *the average length of shots* (TAL). This measures shots switching frequency and content completeness, videos owning too many or too few shots switching degrade the viewer experience and satisfaction.

E. Results and Analysis

Fig. 6 and Table I-VII show our results. Among them, Table I to Table III show the comparison in objective and subjective evaluations between MANVS-auto and several baseline methods. Table V, Table VI and Fig. 6 show the results of different kinds of ablation studies. Table VII shows the comparison of the spent time and quality evaluation of summaries between manually editing method and ours.

RQ1. Table I tabulates the video attributes of generated summaries. We clearly observe that our method achieves the best performance. Specifically, our method performs the best in the attribute of CON and IC. Especially with the support of the shot complement strategies, the consistency attribute far

TABLE IV
THE QUANTITATIVE EVALUATIONS OF MANVS-AUTO WITH BASELINE METHODS ON TVSUM DATASET.

Method	Precision	Recall	F-score
Random	58.77	59.00	58.89
DR [18]	57.76	57.40	57.60
DR-Sup [18]	58.11	58.09	58.10
VAS [7]	61.51	61.35	61.42
DSN-AB [13]	62.07	62.03	62.05
DSN-AF [13]	61.86	61.86	61.86
HSA [8]	61.55	58.15	59.80
FCN [59]	56.14	57.43	56.78
HMT [60]	60.53	59.75	60.14
VSN [61]	56.52	54.74	55.62
MANVS-auto	**62.58**	**63.61**	**63.09**

TABLE V
ABLATION STUDY OF DIFFERENT COMPONENTS IN VSM ON DOCUMENTARY AND MPII MOVIE DESCRIPTION DATASETS.

Method	Documentary Description			MPII Movie		
	R@0.3	R@0.5	mIoU	R@0.3	R@0.5	mIoU
RD	8.96	1.49	7.68	3.44	0.95	5.44
VSL [33]	16.04	10.07	12.99	14.96	9.62	12.79
LG4 [50]	15.30	9.70	10.17	12.11	4.87	7.60
RD + TSS	17.70	3.83	11.65	15.19	5.06	11.81
VSL + TSS	27.27	15.31	20.17	40.51	24.05	33.05
LG4 + TSS	**70.65**	**56.18**	**48.56**	**72.15**	**41.77**	**41.55**

superior to other methods and can reach 100%. In addition, the TAL of the generated summary by our method conforms to the cinematographic aesthetic guidelines.

From the user study in Table II, we observe that our method acquires the highest score on both visual attraction and narrative completeness, which demonstrates that by leveraging multimodal information and aesthetic guidance, our method can produce high-quality summaries. In the visual attraction, the appropriate length of shots helps improve the professional of the entire video summary, neither boring the audience nor affecting visual continuity, while those in other methods are either too long or too short. In the narrative completeness, incomplete voiceover and low information coverage can greatly reduce for the audience's understanding of the video content.

In order to validate the performance of our method on both VA and NC from a statistical perspective, a Wilcoxon test was implemented on the data of user study. The P-values between MANVS-auto and every comparison method were calculated to study the statistical significance by a Wilcoxon test. The results shown in Table III suggest that our method is significantly better than other methods under the P-value threshold of 0.01 in terms of both VA and NC.

TABLE VI
ABLATION STUDY OF DIFFERENT COMPONENTS IN MANVS ON THE TVSUM DATASET.

Method	Precision	Recall	F-score
w/o SS	25.79	37.61	30.60
w/o KSS	33.55	36.49	39.40
w/o HE	43.32	43.38	43.35
w/o SL	32.41	**69.18**	44.14
w/o BS	59.32	62.38	60.81
w/o CC	58.41	58.18	58.14
w/o SSA	58.79	58.61	58.70
w/o OM	61.79	62.70	62.25
w/o PP	**62.58**	63.61	**63.09**
MANVS	**62.58**	63.61	**63.09**

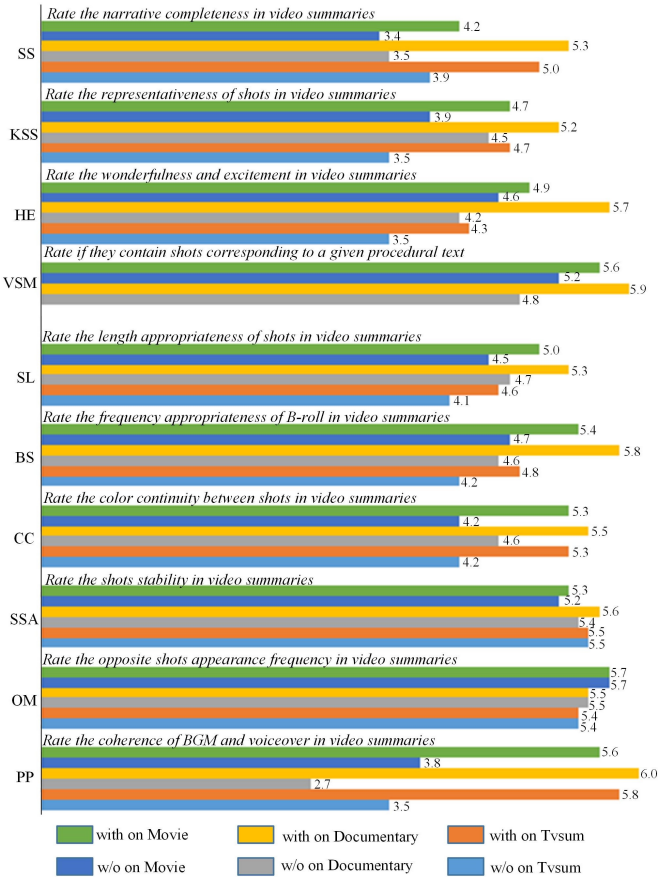

Fig. 6. User study questions and results for individual component and aesthetic constraint.

According to the quantitative evaluation for TVsum in Table IV, our method clearly outperforms baselines. Comparing F-score of HSA and DR, we find that our method improves the performance significantly. Compared with DR-Sup, the F-score of MANVS-auto have 5% gain. Besides, Figure 7 illustrates some summarization examples generated on TVsum, we compare the durations of selected key frames of MANVS-auto with DR-Sup, VAS, DSN-AB and ground truth. The results show that compared with baselines, the shots selected by our method is closer to ground truth.

It is worth noting that the quantitative results in Table IV are inconsistent with the user study results in Table II. For example, the F-score of HSA in Table IV is lower than that of DSN-AF and VAS, but its visual attraction and narrative completeness in Table II are higher than both. The reason may be that the annotations only record the importance of each frame, and the objective evaluation metrics in Table IV calculate the scores through the selected frame, but they can not evaluate the overall quality of video summary [17]. Combining the results of Table I, we notice the average shot length of the former six methods is too short, which results in frequent changes of scenes, and makes the video summaries

like selecting several frames in some consecutive frames. Though they select most of the representative entities or plots, they appear intermittently or just for a few seconds in these video summaries. Therefore, the summary can not tell a fluent story and causes low ratio of consistency, which means that for many emerging sentences in the summary, only a few words are read by the voiceover and can dramatically reduce the narrative completeness and visual attraction of the generated summary, especially for these strong narrative videos such as documentary and movie. In the opposite, the performance of HSA in Table II is relatively better compared with the former six baselines possibly because of its longer shots length, higher ratio of consistency. However, compared to our method, it also suffers from low ratio of complete sentences and less ratio of information coverage, overlong average length of shots, which is tedious for watching. These results prove that our method can achieve the best results in both subjective and objective evaluation.

RQ2. Table V shows experimental results of ablation study R2a. We see that the performance significantly increases by leveraging methods with VL and TSS. In particular, our method achieves 48.56%, 41.55% in mIoU on Documentary Description and MPII datasets respectively, which performs the best in all approaches. Besides, we can find that the performance of LG4 without using TSS is slightly less than that of alternative VSL. However, the performance of LG4 + TSS outperforms a lot than the alternative VSL + TSS. The reason may be that after utilizing TSS, the generated sub-video is much shorter than original one. VSL performs better than LG4 in dealing with long videos, while LG4 is better at localizing language in short videos. Figure 8 shows some shot localization results of our component and alternatives, which shows that our VSM component can acquire the shots that users are interested in. The above results demonstrates that the performance of our VSM component is better than that of alternatives and each cascaded components plays a vital role.

Fig. 6 and Table VI present the subjective and objective results of ablation study R2b, respectively. For the shot selection components, we find that w/o SS can cause serious impacts on the narrative completeness of video summaries compared with the original one. For instance, the performance of w/o SS on the documentary with the score of only 3.5 and the score is lower than our result by 1.8 in the user study. Besides, the F-score of w/o SS on Tvsum also suffer from low accuracy, with the value of only 30.60%. These results show that narrative information is an important factor in affecting the quality of the generated summary, and the SS component is able to capture narrative details. In addition, w/o VSM gets the close performance to our result in Fig. 6. This is in line with our expectation since the only difference in this group is the shots matching the procedural text. Besides, the subjective evaluation results indicate that KSS and HE are plays a positive role in its corresponding aspect. For the aesthetic constraints and post processing component, it can cause some score discrepency between w/o SL, w/o BS and w/o CC and our original method in user study, which are lower than out result by 0.5, 0.7 and 1.1 respectively for the summary of movie. This is possibly because of they increase the viewing experience by providing shots in an appropriate length, supplementing visual content and preserving color smoothness respectively. However, w/o SSA and w/o OM perform nearly the same compared with the original one. This is reasonable since most of our videos are professional videos and they hardly suffer from the issues of shot stability and opposite movement. Apart from that, w/o PP performs the worst with only 2.7 and the score is lower than our result by 3.3 for the summary of documentary. This is because w/o PP significantly reduces the attraction by directly assembling the incoherent BGM clips together. For the objective evaluation, without individual aesthetic constraint can cause the score discrepancy for a little bit. It keep the same for w/o PP because the function of PP lies in producing a coherent BGM and user-designed voiceover not selecting frames. These results verifies the effectiveness of our multimodal-based shot selection and aesthetic-guided shot assembly module.

TABLE VII
COMPARISON BETWEEN PROFESSIONAL MANUAL EDITING AND OURS.

Datasets	Meric	Method		
		Manual	MANVS	MANVS-auto
TVsum	Time	00:16:24s	00:03:18s	00:00:00s
	VA	5.3	4.9	4.7
	NC	6.3	5.7	4.6
Documentary	Time	2:28:12s	0:12:34s	00:00:00s
	VA	5.6	5.2	5.0
	NC	6.2	5.5	5.2
Movie	Time	4:52:37s	0:23:42s	00:00:00s
	VA	5.8	5.2	5.1
	NC	6.0	5.4	4.7

RQ3. As shown in Table VII, our results are close to the quality of video summary generated by experienced video producers. For instance, MANVS is only lower than the manually editing result on documentary by 0.4 and 0.7 in visual attraction and narrative completeness respectively. More importantly, manually editing a documentary summary costs 2 hours 28 minutes 12 seconds and MANVS only costs 12 minutes 34 seconds. Furthermore, the MANVS-auto does not cost any human producing time, such as writing program text and user-designed voiceover and background music, because it is completely automatic. Therefore, the results of MANVS-auto are slightly inferior to the MANVS in narrative completeness and visual attraction as a whole. After finishing the task, we invited the producer for commenting. Through these comments, we can draw the conclusions below: Even if the fast forward function is leveraged to browse video quickly, a lot of time has been spent in watching and cutting it when dealing with an unfamiliar video. Besides, these videos involved different places and seasons and the visual content changed frequently, so it cost lots of time in assembling the appropriate video shots and adding the necessary fade-in and fade-out effects. In contrast, users only need to spend little time writing procedural text in our method or even do not need to spend any time by choosing MANVS and MANVS-auto, respectively. This illustrates that our method can rapidly produce a high quality video summary.

Fig. 7. Summarization examples (on TVsum). The five bars below each video represent the results generated by DR-Sup, VAS, DSN-AB, MANVS-auto and gound truth, respectively. The long gray bar and the short colored bar are the time stream of video and the selected key frame, respectively.

Fig. 8. Some visual semantic matching examples on documentary and MPII Movie datasets. Procedural texts are listed top the thumbnails. Our results are outlined in green.

VII. Discussion and Future Work

We have proposed MANVS, a novel method for generating a narrative video summary with aesthetic appealing. By inputting a narrative video and the corresponding text of subtitles, MANVS automatically selects representative video shots, and assembles them into a visual appealing video summary based on cinematographic aesthetics guidelines. Our method also allows users to 1) design procedural text to find the desired video shots, and 2) provide customized voiceover to create a personalized video summary. Both experiments and user study show that MANVS can significantly save editing time and help users generate satisfactory narrative video summaries. However, our current MANVS still has some limitations, which also points out the direction of future work.

Multimodal Feature Fusion. To the best of our knowledge, there are currently no multimodal datasets available for the video summarization task. Learning-based multimodal feature fusion heavily depends on a great number of well labelled audio, subtitles, visual data. In the future, constructing multimodal datasets and training models suitable for video summarization may motivate more extensive applications.

Consistency for Subtitle Summarization. Our subtitle summarization component relies on the extractive text summarization method. However, if the text of subtitles contains too many pronouns, directly connecting multiple sentences into a summary may produce inconsistency of subjects, which has a negative impact on understanding the video. Although other components in the shot selection module may make up for this defect in some sense, semi-automatic generation of subtitle summary combined with script information and user interaction may bring better performance.

Detailed Expression of Film Art. Our aesthetic constraints are set according to the general film rules, but sometimes in order to reflect a special artistic style, the photography conventions may be deliberately broken. Our shot assembly component may face challenges when dealing with some videos with special narrative methods such as flashback. In the future, we will combine more interactions with users to provide various artistic modes of fine shot switching.

References

[1] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Video storytelling: Textual summaries for events," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 554–565, 2020.

[2] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 996–1009, 2019.

[3] S. Y, C. D, and H. Y, "Multiple pairwise ranking networks for personalized video summarization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1–10.

[4] X. Li, H. Li, and Y. Dong, "Meta learning for task-driven video summarization," *IEEE Trans.Ind. Electron.*, vol. 67, no. 7, pp. 5778–5786, 2020.

[5] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *IEEE Trans. on Multimedia*, vol. 19, no. 9, pp. 2010–2021, 2017.

[6] A. Truong, F. Berthouzoz, W. Li, and M. Agrawala, "Quickcut: An interactive tool for editing narrated video," in *Pro. ACM Symp. User Interface Softw. Technol.*, 2016, pp. 497–507.

[7] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 39–54.

[8] B. Zhao, X. Li, and X. Lu, "Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2018, pp. 7405–7414.

[9] M. Basavarajaiah and P. Sharma, "Survey of compressed domain video summarization techniques," *ACM Comput. Surv.*, vol. 52, no. 6, pp. 1–29, 2019.

[10] S. A. Ahmed, D. P. Dogra, S. Kar, R. Patnaik, S.-C. Lee, H. Choi, G. P. Nam, and I.-J. Kim, "Query-based video synopsis for intelligent traffic monitoring applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3457–3468, 2020.

[11] J.-H. Choi and J.-S. Lee, "Automated video editing for aesthetic quality improvement," in *Proc. ACM Multimedia*, 2015, pp. 1003–1006.

[12] P. Varini, G. Serra, and R. Cucchiara, "Personalized egocentric video summarization of cultural tour on user preferences input," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2832–2845, 2017.

[13] W. Zhu, J. Lu, J. Li, and J. Zhou, "Dsnet: A flexible detect-to-summarize network for video summarization," *IEEE Trans. Image Process*, pp. 948–962, 2021.

[14] M. Sun, A. Farhadi, B. Taskar, and S. Seitz, "Summarizing unconstrained videos using salient montages," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2256–2269, 2017.

[15] S.-P. Lu, S.-H. Zhang, J. Wei, S.-M. Hu, and R. R. Martin, "Timeline editing of objects in video," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 7, pp. 1218–1227, 2012.

[16] B. Zhao, X. Li, and X. Lu, "Property-constrained dual learning for video summarization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3989–4000, 2020.

[17] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2019, pp. 7596–7604.

[18] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI Conf. Artif. Intell*, 2018, pp. 7582–7589.

[19] L. Yuan, F. E. H. Tay, P. Li, and J. Feng, "Unsupervised video summarization with cycle-consistent adversarial lstm networks," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2711–2722, 2020.

[20] K. Zhang, K. Grauman, and F. Sha, "Retrospective encoders for video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 383–399.

[21] Z. Li, J. Tang, X. Wang, J. Liu, and H. Lu, "Multimedia news summarization in search," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–20, 2016.

[22] T. Makino, T. Iwakura, H. Takamura, and M. Okumura, "Global optimization under length constraint for neural text summarization," in *Proc. Assoc. Comput. Linguist.*, 2019, pp. 1039–1048.

[23] S. Xu, H. Li, P. Yuan, Y. Wu, X. He, and B. Zhou, "Self-attention guided copy mechanism for abstractive summarization," in *Proc. Assoc. Comput. Linguist.*, 2020, pp. 1355–1362.

[24] H. Jin, T. Wang, and X. Wan, "Multi-granularity interaction network for extractive and abstractive multi-document summarization," in *Proc. Assoc. Comput. Linguist.*, 2020, pp. 6244–6254.

[25] J. Xu, Z. Gan, Y. Cheng, and J. Liu, "Discourse-aware neural extractive text summarization," in *Proc. Assoc. Comput. Linguist.*, 2020, pp. 5021–5031.

[26] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala, "Video digests: a browsable, skimmable format for informational lecture videos." in *Pro. ACM Symp. User Interface Softw. Technol.*, 2014, pp. 1–10.

[27] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "Coot: Cooperative hierarchical transformer for video-text representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–14.

[28] X. Qu, P. Tang, Z. Zou, Y. Cheng, J. Dong, P. Zhou, and Z. Xu, "Fine-grained iterative attention network for temporal language localization in videos," in *Proc. ACM Multimedia*, 2020, pp. 4280–4288.

[29] D. Cao, Y. Zeng, X. Wei, L. Nie, R. Hong, and Z. Qin, "Adversarial video moment retrieval by jointly modeling ranking and localization," in *Proc. ACM Multimedia*, 2020, pp. 898–906.

[30] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proc. AAAI Conf. Artif. Intell*, 2019, pp. 8199–8206.

[31] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proc. AAAI Conf. Artif. Intell*, 2019, pp. 9062–9069.

[32] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2070–2083, 2019.

[33] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proc. Assoc. Comput. Linguist.*, 2020, pp. 6543–6554.

[34] M. Merler, D. Joshi, K.-N. C. Mac, Q.-B. Nguyen, S. Hammer, J. Kent, J. Xiong, M. N. Do, J. R. Smith, and R. S. Feris, "The excitement of sports: Automatic highlights using audio/visual cues," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, June 2018, pp. 2520–2523.

[35] Z. Guo, Z. Zhao, W. Jin, W. Dazhou, L. Ruitao, and J. Yu, "Tao-highlight: Commodity-aware multi-modal video highlight detection in e-commerce," *IEEE Trans. on Multimedia*, pp. 1–12, 2021.

[36] M. Merler, K. C. Mac, D. Joshi, Q. Nguyen, S. Hammer, J. Kent, J. Xiong, M. N. Do, J. R. Smith, and R. S. Feris, "Automatic curation of sports highlights using multimodal excitement features," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1147–1160, 2019.

[37] T. Decroos, V. Dzyuba, J. Van Haaren, and J. Davis, "Predicting soccer highlights from spatio-temporal match event streams," in *Proc. AAAI Conf. Artif. Intell*, 2017, pp. 1302–1308.

[38] L. Hu, W. He, L. Zhang, T. Xu, H. Xiong, and E. Chen, "Detecting highlighted video clips through emotion-enhanced audio-visual cues," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2021, pp. 1–6.

[39] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[40] Y. Zhang, L. Zhang, and R. Zimmermann, "Aesthetics-guided summarization from multiple user generated videos," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 11, no. 2, pp. 1–23, 2015.

[41] Y. Niu and F. Liu, "What makes a professional video? a computational aesthetics approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1037–1049, 2012.

[42] B. Huber, H. V. Shin, B. Russell, O. Wang, and G. J. Mysore, "B-script: Transcript-based b-roll video editing with recommendations," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–11.

[43] M. Wang, G.-W. Yang, S.-M. Hu, S.-T. Yau, and A. Shamir, "Write-a-video: computational video montage from themed text." *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, 2019.

[44] T. Hu, Z. Li, W. Su, X. Mu, and J. Tang, "Unsupervised video summaries using multiple features and image quality," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 2017, pp. 117–120.

[45] B. Shi, L. Ji, Z. Niu, N. Duan, M. Zhou, and X. Chen, "Learning semantic concepts and temporal alignment for narrated video procedural captioning," in *Proc. ACM Multimedia*, 2020, pp. 4355–4363.

[46] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

[47] M. Zhong, P. Liu, D. Wang, X. Qiu, and X. Huang, "Searching for effective neural extractive summarization: What works and what's next," in *Proc. Assoc. Comput. Linguist.*, 2019, pp. 1049–1058.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[49] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," in *Proc. Assoc. Comput. Linguist.*, 2018, pp. 675–686.

[50] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2020, pp. 10 810–10 819.

[51] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–20.

[52] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.

[53] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2018, pp. 7794–7803.

[54] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis*, 2006, pp. 404–417.

[55] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a

fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw.*, vol. 5, no. 50, pp. 1–4, 2020.
[56] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *In Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7654–7658.
[57] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2015, pp. 5179–5187.
[58] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2015, pp. 3202–3212.
[59] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 347–363.
[60] B. Zhao, M. Gong, and X. Li, "Hierarchical multimodal transformer to summarize videos," *Neurocomputing*, vol. 468, pp. 360–369, 2022.
[61] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *CVPR*, 2019, pp. 7902–7911.
[62] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *SIGNLL*, 2016, pp. 280–290.
[63] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5267–5275.
[64] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.
[65] document Rec. ITU-R, "Methodology for the subjective assessment of video quality in multimedia applications," *BT.1788*, pp. 1–13, 2007.

Jiehang Xie received the master degree in computer software and theory in 2020 from Shaanxi Normal University, China. He is currently working toward the Ph.D. degree in the College of Computer Science at Nankai University, Tianjin, China. His research interests include multimodal multimedia analysis and affective computing.

Xuanbai Chen received the B.E. degree in computer science and technology in 2021 from Nankai University, Tianjin, China, where he is currently working toward the M.S. degree in computer vision with the Robotics Institute, School of Computer Science, Carnegie Mellon University. His research interests include domain adaptation and video summarization in computer vision.

Tianyi Zhang is currently working toward the PhD degree in the faculty of Electrical Engineering, Mathematics & Computer Science (EEMCS) in Delft University of Technology. He is associated with the Distributed & Interactive Systems (DIS) group at Centrum Wiskunde & Informatica (CWI), the national research institute for mathematics and computer science in the Netherlands. His research interests lie in human-computer interaction and machine learning based affective computing.

Yixuan Zhang is doing his undergraduate studies at Nankai University. His research interests include movie summary and the sentiment analysis.

Shao-Ping Lu is currently an associate professor at Nankai University, China. Prior to that he was a postdoc and senior researcher at Vrije Universiteit Brussels (VUB) in Belgium. He received the Ph.D. degree in Computer Science from Tsinghua University, China. His research interests lie primarily in the intersection of visual computing, with particular focus on computational photography, 3D image and video representation, visual scene analysis and machine learning.

Pablo Cesar (Senior Member, IEEE) currently leads the Distributed & Interactive Systems Group, Centrum Wiskunde & Informatica (CWI) and is Professor with TU Delft, The Netherlands. His research combines HCI and multimedia systems, and focuses on modelling and controlling complex collections of media objects distributed in time and space. He is IEEE Senior and ACM Distinguished member, and part of the editorial board of *IEEE Multimedia*, *ACM Transactions on Multimedia* and *IEEE Transactions on Multimedia*. He received the prestigious Netherlands Prize for ICT Research in 2020 because of his work on human-centered multimedia systems. He is the principal investigator from CWI in a number of National and European projects, and acted as an Invited Expert at the European Commission's Future Media Internet Architecture Think Tank.

Yulu Yang received the BE degree from Beijing Agriculture Engineering University, China in 1984. He received the ME and PhD degrees from Keio University, Japan in 1993 and 1996, respectively. He is currently a full professor in the Department of Computer Science, Nankai University, China. His research interests include parallel processing and intelligence computing.