IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 May 2024, accepted 19 July 2024, date of publication 29 July 2024, date of current version 20 September 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3434508*

RESEARCH ARTICLE

# An Aesthetic-Driven Approach to Unsupervised Video Summarization

**HONGBEN HUANG[1], ZAIQUN WU[2], GUANGYAO PANG[3], AND JIEHANG XIE[3]**
[1]Guangxi Key Laboratory of Machine Vision and Intelligent Control, Wuzhou University, Wuzhou 543002, China
[2]Baise University, Baise 533000, China
[3]Guangxi Colleges and Universities Key Laboratory of Intelligent Industry Software, Wuzhou 543002, China

Corresponding author: Zaiqun Wu (wuzq7518@bsuc.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62262059, in part by the Natural Science Foundation of Guangxi Province under Grant 2021JJA170178, and in part by the Industry-University-Research Project of Wuzhou High-tech Zone and Wuzhou University under Grant 2020G003.

**ABSTRACT** The aim of video summarization is to condense lengthy videos into shorter versions, making them more accessible for viewing. Typically, people can identify important shots within a video by using audiovisual cues and assessing the aesthetic attributes of the frames. However, existing methods either focus only on unimodal features or neglect the aesthetic attributes of videos, resulting in the limited quality of the generated summaries. Particularly, the reliance on annotated data for training models also imposes limitations, as it not only demands significant time and resources but may not capture the diverse and subjective nature across different videos. To tackle these issues, we propose an aesthetic-driven approach to unsupervised video summarization, namely ADUVS. Specifically, ADUVS incorporates an aesthetics encoder to extract key aesthetic attributes. Additionally, we design a multimodal fusion module that assesses how different modalities of information complement each other and highlights the most relevant segments for the desired summary. Moreover, the training process for ADUVS does not require reliance on annotated data, thus reducing both time and labor costs. Extensive experiments demonstrate that our proposed method is better than various benchmark methods across commonly used evaluation metrics.

**INDEX TERMS** Video summarization, feature extraction, multimodal information.

## I. INTRODUCTION

The rapid expansion of smartphones and online video platforms simplifies video recording and sharing processes, leading to a significant increase in video content. This abundance requires viewers to spend more time searching for their preferred content and exacerbates the problem of information overload [1], [2]. In this context, providing brief and concise video summaries is crucial. Such summaries improve video browsing efficiency and reduce data storage demands. As a result, video summarization has attracted rising research interest recently.

The evolution of computing hardware has shifted the focus from manual feature extraction to deep learning-based methods in video summarization. Several supervised video summarization methods that rely on unimodal features are proposed, demonstrating notable performance [3], [4]. However, these supervised methods encounter two significant challenges. Firstly, creating ground truth labels for training involves a demanding and time-intensive process. It requires annotators to watch a large number of videos and assign importance scores to each shot within these videos. This task not only demands considerable effort but also a lot of time and human resources. Secondly, the process of labeling is subject to varying interpretations by different annotators. When different people watch the same video, they might have different opinions on what parts are important. This difference in perception leads to inconsistencies in the ground truth labels. Such variability in human judgment can introduce bias in the training data, which might affect the performance and accuracy of the supervised video summarization models. Furthermore, their reliance on unimodal features may lead to information bias in the generated summaries [5], [6]. Driven by the remarks discussed above, we worked towards the

The associate editor coordinating the review of this manuscript and approving it for publication was Xinfeng Zhang.

development of multimodal-based unsupervised framework for video summarization.

Most current video summarization methods integrate audiovisual information and human cognitive habits into their decision-making processes. Nevertheless, these methods often overlook the aesthetic attributes of video shots [7]. These attributes, which include visual appeal, composition, and emotional impact, play a crucial role in selecting high-quality shots. Their absence in the summarization process can lead to generated summaries of limited quality, lacking in visual and emotional engagement. However, it is important to note that there is no unified agreed-upon definition of aesthetics, which makes its integration into video summarization a complex task. The subjective nature of aesthetics means that what is considered visually appealing or emotionally resonant can vary greatly among individuals [8], [9]. Despite this challenge, incorporating aesthetic considerations is a valuable step towards creating more engaging and effective video summaries.

Therefore, a primary focus of our work is identifying and capturing the aesthetic attributes and effectively incorporating them into video summarization. Additionally, simply increasing the number of modalities does not guarantee improved performance in a model [10], [11]. Hence, another key aspect of our work is to develop an effective method for multimodal fusion. This involves strategically combining information from different modalities, to enhance the overall quality of the video summaries.

To overcome the above discussed problems, we propose an aesthetic-driven approach to unsupervised video summarization, namely ADUVS. Specifically, we design an aesthetics encoder, which is developed using insights from photographic guidelines and image aesthetic evaluation methods. This encoder extracts important aesthetic attributes like saturation, hue, exposure of light, contrast, and composition. Alongside the aesthetics encoder, we also develop a set of dedicated encoders for different modalities present in videos. This includes a visual encoder for capturing visual details, an audio encoder for processing sound elements, and a text encoder for incorporating textual information. Each of these encoders is crucial for capturing the rich, multimodal feature of videos. By integrating these diverse modalities, our framework provides a more complete and detailed representation of video content, enhancing the effectiveness of video summarization. Moreover, we develop a novel multimodal fusion module, which effectively captures the complementary relationships among multimodal information. In addition, we ensure that our model trains efficiently in an unsupervised manner by incorporating length considerations into the loss function. This approach enables the model to learn and adapt without the need for prelabeled data. Our contributions are the following:

- We propose an aesthetics encoder to capture a wide range of appealing aesthetic attributes from video frames, and integrates them to obtain a more enriched and comprehensive feature representation.
- We present a multimodal fusion module, which emphasizes the synergistic integration of audio, visual, and textual data, effectively capturing complementary relationships among different modalities.
- We conduct extensive quantitative and qualitative experiments, incorporating various benchmark datasets and performance metrics. The results demonstrate the effectiveness of our proposed method.

## II. RELATED WORK

Existing research on video summarization primarily focuses on two main aspects: unimodal-based and multimodal-based approaches. In this section, we initiate the literature review by discussing unimodal-based methods, followed by an exploration of multimodal methods. Besides, we present related studies on image and video aesthetics.

### A. UNIMODAL-BASED METHODS

Various approaches for video summarization have been proposed over the past few decades. Early approaches heavily relied on manually sampling features [12], which have been replaced by methods utilizing neural networks for automatic feature extraction. Following this technological path, researchers have proposed numerous unimodal-based methods. Zhang et al. [13] pioneer the utilization of long short-term memory networks for extracting visual features, modeling the temporal dependencies among frames. Rochan et al. [14] draw on semantic segmentation techniques to extract visual features by employing fully convolutional networks, which process all video frames in parallel. Particularly, since annotated data is often lacking in video summarization, researchers propose various unsupervised approaches. Apostolidis et al. [15] construct a deterministic attention auto-encoder for video summarization, enabling faster and more stable model training. Liu et al. [16] propose the 3D U-Net network to simulate sequential dependencies among frames, and show that 3D features have stronger representation abilities than 2D features in video summarization task.

### B. MULTIMODAL-BASED METHODS

More recently, some researchers have suggested including human perception of audio-visual changes within video summarization [17]. Consequently, several multimodal-based methods have been introduced. For example, Wei et al. [18] design a video descriptor to obtain textual data, map visual information into textual representations, and finally generate video summaries guided by semantics Xie et al. [7] draw inspiration from human habit of perceiving video information, and combine text, visual, and audio modalities for video summarisation. Li et al. [19] take the category information of the video as textual data, and utilize a text encoder and a video encoder to model the correspondences of multimodal data. Huang et al. [20] introduce a causal semantics extractor to effectively integrate visual and textual

information, thereby improving the quality of generated summaries.

### C. INCORPORATING IMAGE AND VIDEO AESTHETICS

Recently, several studies have shown that a desired video summary not only requires meaningful shots, but also high-quality shots of artistic expression [21]. Consequently, some methods have been proposed to select significant shots guided by image or video aesthetics. For instance, Hu et al. [22] propose a method that selects video shots based on the quality of the images. Xie et al. [5] utilize a shot selection module and a shot assembly module to generate video summaries, where the shot assembly module filters the shots according to manually predefined aesthetic rules. Despite this approach achieves remarkable results, the manually predefined rules may lead to issues of information distortion. In our approach, we take both multimodal and aesthetic features into consideration and map the aesthetic features to a vector space, jointly modelling the aesthetic and visual features to obtain informative and aesthetically appealing shots.

## III. PROPOSED APPROACH

In this section, we describe the proposed aesthetic-driven approach to unsupervised video summarization. As depicted in Figure 1, the ADUVS takes a video stream and its corresponding audio as input, and then output a coherent video summary. Specifically, the overall framework of ADUVS is composed of six parts: aesthetics encoder, visual encoder, text encoders, audio encoder, multimodal fusion module and summary predictor. Among them, the aesthetic encoder, visual encoder and text encoder take the video stream as input, the audio encoder take the audio stream as input. The aesthetic encoder is dedicated to capturing visually pleasing features to enhance the aesthetic quality of the generated summary. The visual encoder is responsible for extracting spatio-temporal information from the video frames, encompassing the dynamic aspects and temporal variations within the visual content. The text encoder focuses on the fine-grained details present in the video frames, extracting and encoding semantic information relevant to the visual content. Meanwhile, the audio encoder concentrates on capturing acoustic features within in the video, enabling ADUVS to achieve a more comprehensive understanding of the content. Subsequently, the multimodal fusion module effectively integrates the output features from these encoders, ensuring a comprehensive interaction and synergy of information from various modalities. Finally, the integrated features are fed into the summary predictor to generate a high-quality video summary. In what follows, we introduce the details of each part of our framework.

### A. AESTHETICS ENCODER

The aim of the aesthetic encoder is employ deep learning models for capturing and encoding aesthetic features in videos, improving the quality of generated summary. Specifically, the aesthetic encoder learns how humans perceive and understand aesthetics, translating these cognitive patterns into computer-interpretable feature representations. By incorporating the aesthetic encoder, the model gains the ability to extract and represent aesthetic attributes within the video. These features provide the model with richer and more valuable information, ensuring that the generated summaries align more closely with one's aesthetic preferences.

However, the aesthetics is subjective, and there are numerous aesthetic attributes that may influence a one's aesthetic experience. Given this subjectivity, we choose saturation, hue, exposure of light, image contrast and image composition as they are the most representative attributes explored for aesthetics analysis. These aesthetic attributes have been validated and accumulated over time, becoming key cues in various visual tasks. We describe how to extract these aesthetic attributes in detail below:

#### 1) SATURATION AND HUE

In both photography and psychology, saturation and hue play a crucial role and have the potential to significantly impact the aesthetics of an image, as well as the subjective perception of humans. Saturation represents the intensity or purity of color, and has the ability to evoke various emotions and convey diverse messages. Highly saturated colors usually evoke feelings of excitement and intensity, whereas desaturated colors tend to convey a more gentle or wistful atmosphere. Moreover, the hue, representing the specific color or wavelength of light in an image, also plays a significant role. Similar to saturation, different hues elicit distinct emotions. For example, warm hues such as red and orange can generate a sense of warmth and vitality, while cool hues like blue and green can instill feelings of calmness and tranquility. In this work, we compute the mean saturation and mean hue [23] to obtain these two aesthetic attributes, denoted as $f_1$ and $f_2$, respectively

#### 2) EXPOSURE OF LIGHT

The appropriate level of exposure plays a crucial role in image quality. Images with excessive exposure may appear overbright and lack details, contributing to a decline in quality. Conversely, underexposed images may suffer from insufficient light, leading to a loss of important visual information. Consequently, exposure of lights serve as a significant discriminator between high and low-quality images. In this context, we utilize the average pixel intensity [23] as a quantitative indicator of exposure, denoted as $f_3$. This metric aids in objectively assessing the overall brightness of an image and contributes to our comprehensive evaluation of aesthetic quality.

#### 3) IMAGE CONTRAST

Image contrast, a pivotal facet of visual aesthetics, refers to the extent of variation in brightness or grayscale values across different regions within an image [24]. This characteristic

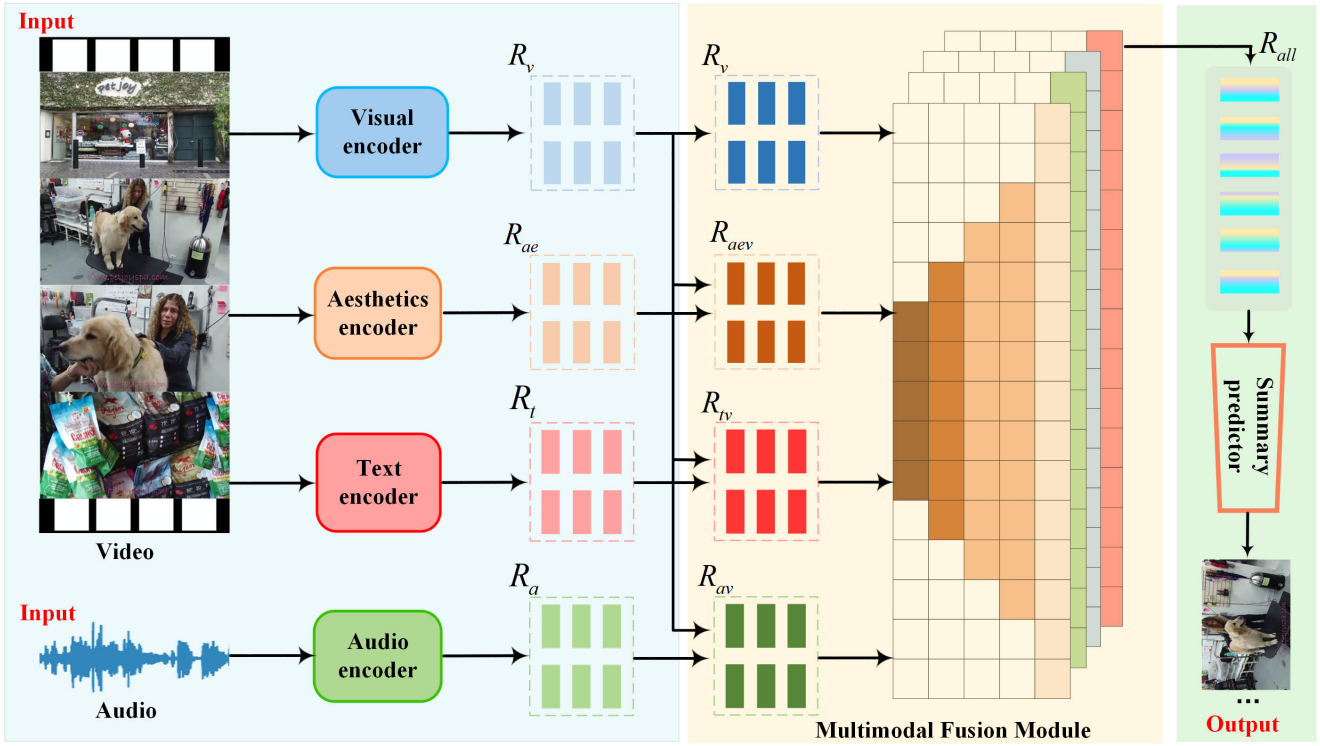


**FIGURE 1.** The framework of proposed ADUVS, best viewed in color.

holds substantial sway over the overall visual allure of an image. High-contrast images boast well-defined transitions between light and dark regions, accentuating details and augmenting the overall visual impact. Conversely, low-contrast images exhibit more subtle fluctuations in brightness, yielding a comparatively understated appearance. In our investigation, we follow to the methodology proposed in [22] for quantifying image contrast, denoted as $f_4$. This metric offers valuable insights into the distinctiveness and clarity of visual elements within an image, thereby contributing to the comprehensive assessment of aesthetic attributes.

#### 4) IMAGE COMPOSITION

Image composition, a fundamental aspect of visual aesthetics, refers to the arrangement and placement of visual elements within an image. It plays a crucial role in enhancing the aesthetics of an image, as composition guides the viewer's attention and significantly influences the overall visual impact and appeal. In our study, we leverage the pretrained SAMP-Net architecture [25] to capture the composition feature. This architecture dissects image composition from the perspectives of various composition patterns, including symmetrical and radial balance, and the rule of thirds. Particularly, we utilize the output of the previous layer of the prediction layer in the pretrained SAMP-Net as the representation of image composition, denoted as $f_5$. This compositional feature contributes valuable insights into how the arrangement of visual elements within an image impacts its overall aesthetic quality.

Based on these aesthetic attributes, we employ a two-layer fully connected network to independently map each aesthetic attribute into the vector space. This process involves transforming each attribute into a numerical representation that captures its specific features. Subsequently, these individually mapped attributes are concatenated to form the comprehensive aesthetic representation. This representation serves as a consolidated feature set, combining information from saturation, hue, exposure, contrast, and composition. The concatenated aesthetic representation is denoted as:

$$f_k' = \varsigma\left(f_k\right), \tag{1}$$

$$R_{ae} = f_1' \oplus f_2' \oplus f_3' \oplus f_4' \oplus f_5', \tag{2}$$

where $\varsigma$ denotes the two-layer fully connected network, $f_k$ is the aesthetic attribute, $R_{ae}$ means the aesthetic representation, $f'$ is the mapping of the corresponding aesthetic attribute, $\oplus$ denotes the concatenation operation.

### B. VISUAL ENCODER

Given that the interplay between spatial and temporal elements in video stream. the visual encoder is purposefully designed to adeptly capture the inherent spatial and temporal intricacies associated within a video. Specifically, we leverage S3D [26], a model that extends traditional two-dimensional convolutional networks into the three-dimensional domain. In the S3D architecture,

the two-dimensional convolutional kernels are dynamically expanded into the temporal dimension by replicating the weights across this dimension. This approach, as opposed to training the model from scratch, allows effective learning of spatiotemporal features within the RGB frames of the video. The utilization of S3D facilitates the extraction of comprehensive information encapsulating both spatial and temporal dynamics, enhancing the visual encoder's capability to represent the nuanced content of the video frames. In this work, the visual encoder effectively transforms the spatiotemporal features extracted by S3D into rich visual representation, denoted as $R_v$.

### C. AUDIO ENCODERS

Beyond the aesthetic and visual aspects, audio introduces a layer of contextual richness that significantly enhances the depth and comprehensiveness of the summarization process. In general, the audio within a video encompasses emotional tones, ambient sounds, and dialogues, offering nuanced cues that contribute to a holistic understanding of the content. For instance, background music, spoken words, or environmental sounds can convey essential information and emotions crucial to the video's narrative. Therefore, the incorporation of the audio modality in video summarization is pivotal for capturing the complete essence of a video. It enriches the summarization process with additional layers of information, contributing to a more nuanced and contextually informed representation. In this work, we construct an audio encoder based on the pretrained VGGish model [27]. Specifically, our audio encoder begins by preprocessing the input audio, applying the short-time fourier transform to create a magnitude spectrogram. mel-frequency filtering is then used to simulate the human auditory system, followed by log compression to reduce dynamic range. The processed audio waves are fed into the pre-trained VGGish model, which employs convolutional and fully connected layers to extract high-level embeddings representing semantic information from the audio. The resulting embeddings serve as the audio representations denoted as $R_a$. By incorporating the audio modality, our model acquires the capability to capture these auditory nuances, ensuring a more comprehensive and immersive representation of the video content.

### D. TEXT ENCODERS

In the context of video summarization, the incorporation of implicit text information within video frames is essential for enhancing the overall comprehension of the content. This latent information enriches the summarization content, exerting a substantial impact on the overall quality of the generated summaries. Implicit text information in videos encompasses a range of details, including contextual cues, subtle nuances, and potentially emotional or affective elements. Capturing this concealed information is crucial for capturing the intricacies of the video's narrative that may not be immediately apparent in the visual or auditory modalities. To address this need, our approach employs a dedicated text encoder [7] designed explicitly for extracting and encoding implicit textual information within video frames. This process involves acquiring both descriptive and affective text, followed by leveraging a pre-trained BERT model for encoding this textual information. The resulting text representation, denoted as $R_t$, serves as a critical component in the multimodal fusion process, ensuring that the implicit textual context is appropriately captured.

### E. MULTIMODAL FUSION

Generally, leveraging data from multiple modalities offers distinct and complementary insights compared to a single modality. However, the direct concatenation of multimodal features may lead to suboptimal results, as highlighted in previous studies [10]. To address this challenge and harness the full potential of diverse modalities, we introduce a multimodal fusion module.

Concretely, through the above encoding operations for four modalities, we obtain higherorder semantic features for aesthetics, visual, audio and text modalities. Among them, the visual modality plays a central role as a reference point for alignment. The aesthetic features are strategically aligned with the visual features to emphasize visually pleasing elements, creating a synergy between aesthetics and the inherent visual characteristics. Simultaneously, the audio modality aligns with the visual features, capturing the sound-related aspects synchronized with the visual content. Additionally, the text modality should aligned with the visual features, enhancing the overall contextual understanding by incorporating textual information related to the visual content. This comprehensive alignment ensures that aesthetics, audio, and text modalities are seamlessly integrated with the visual modality, preserving consistency and enriching the overall representation.

Formally, given the visual representation $R_v = \{v_1, v_2, \ldots v_N\}$, and another feature representation $R_u = \{u_1, u_2, \ldots u_L\}$, where $u \in \{ae, a, t\}$. We calculate the attention weight between visual representation and the representations of $u$ modality as follows:

$$ws_{j,i} = \phi\left(u_i^T v_j\right), \tag{3}$$

$$\alpha_{j,i} = \frac{exp^{ws_{j,i}}}{\sum_{l=1}^{L} exp^{ws_{j,l}}}, \tag{4}$$

$$\widetilde{u}_j = \sum_{l=1}^{L} \alpha_{j,l} u_l, \tag{5}$$

$$R_{uv} = \delta\left(\{\widetilde{u}_1, \widetilde{u}_2, \ldots, \widetilde{u}_L,\}\right), \tag{6}$$

where $R_{uv}$ means the the visual-aligned representation of modality $u$, $ws_{j,i}$ is the attention weight between $u_i$ and $v_j$, $\widetilde{u}_j$ is the weighted summation of the representation of modality $u$, $\delta$ is the average-pooling operation, $\phi$ denotes the hyperbolic tangent function, $\{\widetilde{u}_1, \widetilde{u}_2, \ldots, \widetilde{u}_L,\}$ is the visual-aligned representation.

Next, we employ a two-layer feed-forward neural network to execute the feature fusion process, acilitating the

integration of information from aesthetic, visual, audio and text modalities. This process can be formulated as follows:

$$V_m = \xi\left(W_3\phi\left(W_2R_m + b_1\right) + b_2\right), \tag{7}$$

$$\widehat{R}_m = \phi\left(W_4R_m + b_3\right), \tag{8}$$

$$R_{all} = \sum_m V_m\widehat{R}_m, \tag{9}$$

where $m \in \{aev, v, tv, av\}$, $V_m$ denotes the importance score of the corresponding modality, $\xi$ is the normalized exponential function, $b$ denotes the bias, $\widehat{R}_m$ refers to the transformed fixed-length feature vector of the corresponding modality and $R_{all}$ signifies the fused feature.

### F. SUMMARY PREDICTOR

After multimodal fusion, the fused feature serves as input for the summary predictor to generate a video summary. The summary predictor consists of a two-layer fully connected neural network, containing a dropout layer and a sigmoid function. Particularly, the ADUVS can be trained without data annotation. This feature is especially valuable in scenarios where labeled data for training is limited or unavailable. Specifically, we incorporate a length regularization loss [28]. This loss function is instrumental in penalizing the selection of frames composing the summary video based on the input video. The length regularization loss plays a pivotal role in promoting the unsupervised learning paradigm of the ADUVS module, contributing to its adaptability and efficacy in summarizing videos without the need for annotated data, which can be expressed as:

$$L_{lr} = \left\| \frac{1}{I}\sum_{i=1}^{I}\widetilde{y}_i - \varpi \right\|_2, \tag{10}$$

where $I$ denotes the total number of video frames, $\varpi$ is a tunable hyper-parameter, $\widetilde{y}$ means the predicted distribution.

## IV. EXPERIMENTS AND DISCUSSION

### A. EXPERIMENTS SETUP

#### 1) DATASETS

In our research, we conduct an extensive evaluation of the models we developed by utilizing four distinct datasets: SumMe [29], TVSum [30], OVP [31], and YouTube [31]. The SumMe and TVSum datasets are particularly notable for their diverse range of events captured from both first- and third-person perspectives. Specifically, the SumMe dataset encompasses 25 unique videos, each varying in length from a minimum of 1 minute to a maximum of 6 minutes. In contrast, the TVSum dataset is more extensive, comprising 50 videos with durations ranging between 1 and 10 minutes. This variation in video lengths and contents provides a rich ground for testing the effectiveness of our models. Additionally, our study incorporates the OVP dataset, which consists of 50 videos, and the YouTube dataset, which includes 39 videos. These additional datasets significantly enhance the diversity and complexity of our evaluation framework. By integrating these four datasets, we are able to generate a variety of data configurations. This approach allow us to conduct a more thorough and comprehensive evaluation of our developed models, ensuring their robustness and effectiveness across different types and lengths of videos.

**TABLE 1. Comparison of F-Score (%) using baselines on SumMe and TVSum under the C, A and T configurations, unsupervised methods marked with '*'. The best performance is highlighted in bold.**

| Methods | SumMe | | | TVSum | | |
|---|---|---|---|---|---|---|
| | C | A | T | C | A | T |
| SUM-FCN [14] | 47.5 | 51.1 | 44.1 | 56.8 | 59.2 | 58.2 |
| HMT [32] | 44.1 | 44.8 | — | 60.1 | 60.3 | — |
| 3DST-UNet [16] | 47.4 | 49.9 | 47.9 | 58.3 | 58.9 | 56.1 |
| SSPVS [19] | 48.7 | 50.4 | 45.8 | 60.3 | 61.8 | 57.8 |
| SUM-GAN* [28] | 39.1 | 43.4 | — | 51.7 | 59.5 | — |
| DR-DSN* [33] | 41.4 | 42.8 | 42.4 | 57.6 | 58.4 | 57.8 |
| SGA* [15] | 48.9 | — | — | 58.3 | — | — |
| RSGN* [34] | 42.3 | 43.6 | 41.2 | 58.0 | 59.1 | 59.7 |
| 3DST-UNet* [16] | 44.6 | 49.5 | 45.7 | 58.3 | 58.4 | 58.0 |
| ADUVS* | **50.4** | **52.3** | **48.7** | **60.7** | **62.3** | **59.9** |

#### 2) COMPARED METHODS

To confirm the effectiveness of our proposed method, we compared it with a variety of other methods. This includes both basic standard methods and the latest, most advanced methods. Here's a brief overview: VASNet [36], SUM-GAN [28], DR-DSN [33], SGA [15], RSGN [34] 3DST-UNet [16], VASNet [36], SUM-FCN [14], DSNet [37], HMT [32] SSPVS [19].

#### 3) EVALUATION METRICS

To thoroughly evaluate our proposed method, we used two main types of evaluation measures: the F-score [13] and rank correlation coefficients [35]. By employing these two metrics, we can evaluate both the accuracy of our method in identifying correct content and its effectiveness in maintaining the right order or rank of the content, offering a comprehensive view of its performance. Specifically, the F-score is a widely used metric that combines precision and recall into a single measure. The F-score balances these two aspects by calculating their harmonic mean, providing a single score to evaluate performance. The rank correlation coefficients measure how well the ordering of video summaries generated by our method agrees with the order of manually created summaries. We use two specific types of rank correlation coefficients: the Spearman's $\rho$ [38] and Kendall's $\tau$ [39]. Spearman's $\rho$ assesses how well the relationship between two variables can be described using a monotonic function, while Kendall's $\tau$ measures the similarity of the orderings of data when ranked by each of the quantities.

#### 4) IMPLEMENTATION DETAILS

We use the Adam optimizer with the initial learning rate of $5 \times 10^{-4}$. Dropout is applied after the fully connected layer to

**TABLE 2. Comparison with baselines on TVSum and SumMe, using the rank correlation coefficients proposed in [35]. Best results are shown in bold, while sub-optimal ones are in italics.**

| Methods | SumMe | | TVSum | |
|---|---|---|---|---|
| | Spearman's $\rho$ | Kendall's $\tau$ | Spearman's $\rho$ | Kendall's $\tau$ |
| DR-DSN [33] | 0.0501 ± 0.0470 | 0.0433 ± 0.0386 | 0.0227 ± 0.0666 | 0.0169 ± 0.0508 |
| SUM-FCN [14] | 0.0096 ± 0.0111 | 0.0080 ± 0.0091 | 0.0142 ± 0.0042 | 0.0107 ± 0.0032 |
| HMT [32] | **0.0758 ± 0.0434** | *0.0563 ± 0.0258* | **0.107 ± 0.0320** | *0.0603 ± 0.0264* |
| SGA* [15] | -0.0226 ± 0.0695 | -0.0180 ± 0.0558 | -0.0620 ± 0.0388 | -0.0472 ± 0.0299 |
| ADUVS* | *0.0632 ± 0.0095* | **0.0572 ± 0.0100** | *0.0724 ± 0.0315* | **0.0678 ± 0.0280** |

**TABLE 3. The average of subjective scores. Comparison with VASNet, DR-DSN, DSNet, HMT and human methods of video 8, 12, 16 in SumMe and 5, 13, 27 in TVSum. The human baseline is generated using leave-one-out approach.**

| Method | SumMe | | | TVSum | | |
|---|---|---|---|---|---|---|
| | 8 | 12 | 16 | 5 | 13 | 27 |
| DR-DSN [33] | 3.6 | 4.0 | 3.9 | 4.0 | 4.4 | 3.8 |
| DSNet [37] | 3.8 | 3.5 | 4.3 | 3.9 | 4.2 | 4.0 |
| VASNet [36] | 3.7 | 4.3 | 4.0 | 3.8 | 3.9 | 4.2 |
| HMT [32] | 4.0 | 3.9 | 3.8 | 4.2 | 4.0 | 4.3 |
| ADUVS | 4.2 | 4.4 | 4.5 | 4.4 | 4.8 | 4.6 |
| Human | 5.4 | 5.7 | 5.5 | 5.2 | 5.1 | 5.5 |

prevent overfitting, and its value is set to 0.4. The $\varpi$ is equal to 0.6. The experiments are conducted on Pytorch framework with an NVIDIA GeForce RTX 2080 GPU. Following [7], [37], we carry out our experiments using the dataset in three different setups: canonical (C), augmentation (A), and transfer (T). In the C setup, we focus on either the TVSum or SumMe dataset. We split the chosen dataset into two parts: 80% for training and 20% for testing. This means if we are working with the TVSum dataset, for example, 80% of its videos are used to train our model, and the remaining 20% are used to test how well our model performs. In the A setup, we expand our training data beyond the canonical configuration. We still allocate 20% of the TVSum or SumMe videos for testing, just like in the C configuration. However, for training, we not only use the remaining 80% of videos from the chosen dataset but also include all the videos from the other three datasets. This setup enriches the training data, providing a more diverse learning environment for the model. In the T setup, we utilize the entire TVSum or SumMe dataset as the test data, while designating the other three datasets as the training data. This setup is specifically designed to evaluate the model's capacity for effectively transferring its learning from multiple datasets to a new, distinct dataset. Besides, we adopt the standard 5-fold cross-validation method for all experiments, dividing the data into five parts, and the average performance is reported. This way ensures a comprehensive and reliable evaluation of the model's performance.

### B. PERFORMANCE EVALUATION AND DISCUSSION

#### 1) QUANTITATIVE RESULTS

We present the F1 scores of the ADUVS and the comparison methods in Table 1. These results show that our proposed method achieves competitive performance across three data configurations. Particularly, in the SumMe dataset, our model outperforms other unsupervised models in the C configuration, achieving an F1 score of 50.4%, which is 9.0% higher than that of DR-DSN. In the A configuration, our model achieves 62.3% on the TVsum dataset, which is 1.6% and 2.4% higher than C and T configurations, respectively. This illustrates that our model captures various video content depicting diverse events as the dataset size increases. Furthermore, our model also achieves better performance compared to some supervised models. For example, In the A configuration, our model surpassing SUM-FCN by 3.1% and multimodal based method HMT by 2.0%, respectively. Moreover, Table 2 shows the Spearman's $\rho$ and Kendall's $\tau$ correlation coefficients obtained by different methods. We find that our model achieves competitive performance compared to supervised models. For instance, ADUVS achieves the highest performance in terms of Kendall's $\tau$ correlation coefficients on the TVSum dataset, which is 0.0075 higher than the supervised multimodal-based method HMT, but obtain suboptimal results in terms of Spearman's $\rho$ correlation coefficients, which is 0.0346 lower than that of HMT. Similarly, our method also shows competitive performance on the SumMe dataset. Regarding Spearman's $\rho$ correlation coefficient, it achieves a score of 0.0632, which is 0.0131 higher than the supervised method DR-DSN, but 0.0126 lower than HMT. In terms of Kendall's $\tau$, our method achieves 0.0572, which is 0.0752 higher compared to unsupervised SGA and 0.0009 higher HMT.

#### 2) USER STUDY

To thoroughly evaluate our method, we randomly select three videos from both the TVSum and SumMe datasets. We then use various methods to create summaries for these videos. We then organized a viewing session with 50 participants. Each participant was asked to watch the generated video summaries and then rate each one on a 7-point Likert scale, where 1 represents 'poor' and 7 represents 'excellent'. This rating not only requires the participants to consider the overall

**TABLE 4.** The statistical significance (*P*-value), which is the enhanced performance of the ADUVS relative to compared methods, gauged by subjective scores via the Wilcoxon test.

| Method | SumMe | | | TVSum | | |
|---|---|---|---|---|---|---|
| | 8 | 12 | 16 | 5 | 13 | 27 |
| DR-DSN [33] | 4.32e-4 | 3.59e-3 | 4.74e-3 | 1.85e-3 | 4.98e-3 | 7.58e-5 |
| DSNet [37] | 6.25e-5 | 5.43e-4 | 2.94e-3 | 1.94e-4 | 4.29e-5 | 2.71e-4 |
| VASNet [36] | 4.67e-4 | 6.34e-3 | 9.37e-3 | 8.46e-4 | 2.25e-4 | 7.09e-3 |
| HMT [32] | 2.55e-3 | 1.24e-4 | 2.62e-4 | 1.38e-5 | 7.62e-4 | 5.25e-4 |

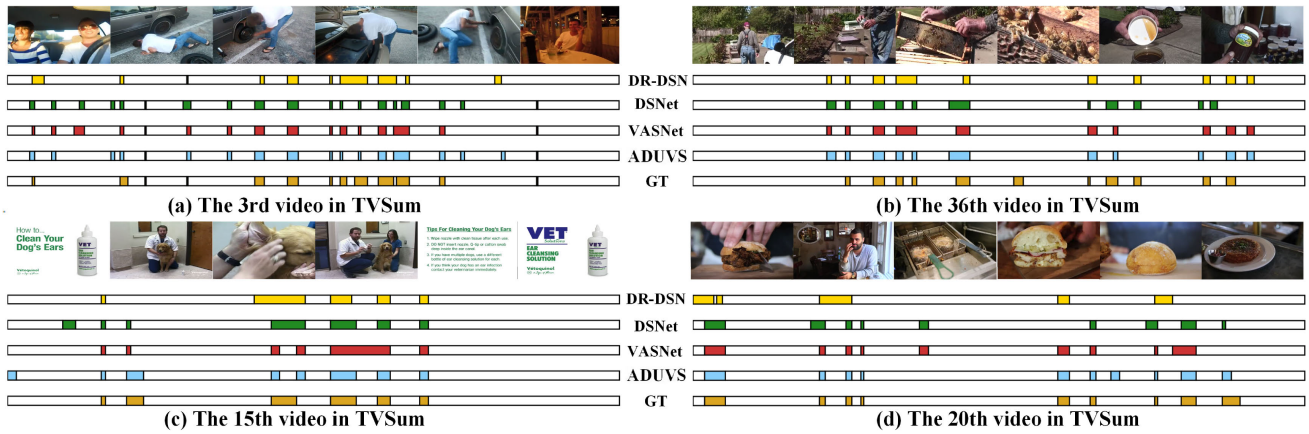


**FIGURE 2.** Selected frames from the 3rd, 15th, 20th, and 36th videos in the TVSum dataset are visualized using DR-DSN, DSNet, VASNet, ADUVS, and the ground truth. The colorful bars indicate the frames chosen by each method.

quality of the summaries but also places significant emphasis on the aesthetic experience during viewing. Participants are encouraged to assess how visually appealing, emotionally engaging, and artistically composed the summaries are, alongside their general quality. To ensure a fair assessment, every participant first watched the original, unsummarized video in full-screen mode before viewing the summary. This setup was in line with international standards for evaluating multimedia content [40]. The participants were all university students with at least one years of experience in image processing, indicating a level of expertise in the subject. About 60% of the subjects were female, and all participants were aged between 19 and 22 years. This demographic mix helped in getting diverse opinions and ratings for a more balanced evaluation of our video summaries. Table 3 displays the average scores given by the participants, and it is clear from these results that our method, ADUVS, outperforms other methods based on deep learning across all the videos tested. In addition, we also find that our results only slightly underperform the human baseline.

### 3) STATISTICAL ANALYSIS

To evaluate the statistical significance of the subjective scores related to our method, we conducted a rigorous validation using the Wilcoxon test on the user study data. Specifically calculated *P* values differentiate the ADUVS from each of the comparative methods listed in Table 3, highlighting the statistical significance. The results, summarized in Table 4, clearly indicate that our method significantly outperforms the compared methods, with statistical significance demonstrated by *P* values below the 0.05 threshold. This confirms that the superiority of our method is statistically reliable and not due to chance.

### 4) VISUALIZATION ANALYSIS

To illustrate the effectiveness of our model, we select four videos at random from the TVSum dataset and ran them through baselines, including our own. Next, we created visual representations of the video summaries that each model generated. This allowed us to directly compare these summaries with the ground truth. When looking at these visual comparisons, as shown in Figure 2, it's clear that the summaries created by our model match much more closely with the ground truth than those produced by the other methods. This close match suggests that our model is not only effective but also outperforms the basic, standard methods used for comparison. It's particularly good at identifying and highlighting the most important parts of the videos, which is a key aspect of its performance.

## V. CONCLUSION

This paper presents an aesthetic-driven approach to unsupervised video summarization. Specifically, our method includes an aesthetic encoder that effectively models aesthetic

attributes. The advantage of this encoder lies in its ability to incorporate aesthetic attributes like saturation, hue, exposure of light, contrast, and composition, enhancing the quality and attractiveness of the generated summaries. Furthermore, we design a multimodal fusion module that efficiently captures and integrates the complementary relationships among various modalities. This integration allows for a more comprehensive and representative summary of the video content. Experimental results show that our model achieves competitive performance. Despite the strengths of our approach, it does have some limitations. One major limitation is the subjective nature of aesthetics. The choices we make regarding which cinematic and photographic guidelines to use in constructing our encoders can be influenced by personal preferences. This subjectivity means that what is considered aesthetically pleasing or relevant can vary widely among different users or contexts. Furthermore, the popular guidelines we rely on may not always be suitable for certain unique artistic expressions. These guidelines might not capture the full range of aesthetic styles and choices present in diverse video content, limiting the versatility of our approach. Therefore, there is a need for further research to develop a more generalized aesthetic encoder. Such an encoder would ideally be capable of accommodating a wider range of aesthetic preferences and artistic styles, making our video summarization approach more universally applicable and robust.

## REFERENCES

[1] M. Ma, S. Mei, S. Wan, Z. Wang, X.-S. Hua, and D. D. Feng, "Graph convolutional dictionary selection with $L_{2,p}$ norm for video summarization," *IEEE Trans. Image Process.*, vol. 31, pp. 1789–1804, 2022.

[2] R. Decorte, J. De Bock, J. Taelman, M. Slembrouck, and S. Verstockt, "Fully automatic camera for personalized highlight generation in sporting events," *Sensors*, vol. 24, no. 3, p. 736, Jan. 2024.

[3] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Combining global and local attention with positional encoding for video summarization," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Nov. 2021, pp. 226–234.

[4] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107677.

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

[6] S. Lee, T. Kim, J. Shin, N. Kim, and Y. Choi, "INSANet: INtra–INter spectral attention network for effective feature fusion of multispectral pedestrian detection," *Sensors*, vol. 24, no. 4, p. 1168, Feb. 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/4/1168

[7] J. Xie, X. Chen, S.-P. Lu, and Y. Yang, "A knowledge augmented and multimodal-based framework for video summarization," in *Proc. ACM Multimedia*, 2022, pp. 740–749.

[8] 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, Jul. 2012.

[9] Y. Wang, Q. Dai, R. Feng, and Y.-G. Jiang, "Beauty is here: Evaluating aesthetics in videos using multimodal features and free training data," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 369–372.

[10] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12692–12702.

[11] Y. Watanabe, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "Text-guided image editing based on post score for gaining attention on social media," *Sensors*, vol. 24, no. 3, p. 921, Jan. 2024.

[12] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref, "Exploring video content structure for hierarchical summarization," *Multimedia Syst.*, vol. 10, no. 2, pp. 98–115, Aug. 2004.

[13] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.

[14] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 347–363.

[15] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *Proc. Int. Conf. MultiMedia Model.*, 2020, pp. 492–504.

[16] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, "Video summarization through reinforcement learning with a 3D spatio-temporal U-Net," *IEEE Trans. Image Process.*, vol. 31, pp. 1573–1586, 2022.

[17] A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, and M. Hasanuzzaman, "A survey on multi-modal summarization," *ACM Comput. Surv.*, vol. 55, no. 13, pp. 1–36, Dec. 2023.

[18] H. Wei, B. Ni, Y. Yan, H. Yu, and X. Yang, "Video summarization via semantic attended networks," in *Proc. Nat. Conf. Artif. Intell.*, 2018, pp. 216–223.

[19] H. Li, Q. Ke, M. Gong, and T. Drummond, "Progressive video summarization via multimodal self-supervised learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5573–5582.

[20] J.-H. Huang, C.-H.-H. Yang, P.-Y. Chen, M.-H. Chen, and M. Worring, "Causalainer: Causal explainer for automatic video summarization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2629–2635.

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

[22] T. Hu, Z. Li, W. Su, X. Mu, and J. Tang, "Unsupervised video summaries using multiple features and image quality," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data (BigMM)*, Apr. 2017, pp. 117–120.

[23] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 288–301.

[24] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 6, pp. 460–473, Jun. 1978.

[25] B. Zhang, L. Niu, and L. Zhang, "Image composition assessment with saliency-augmented multi-pattern pooling," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 1–14.

[26] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 318–335.

[27] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[28] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2982–2991.

[29] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 505–520.

[30] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5179–5187.

[31] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, and A. D. A. Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 56–68, Jan. 2011.

[32] B. Zhao, M. Gong, and X. Li, "Hierarchical multimodal transformer to summarize videos," *Neurocomputing*, vol. 468, pp. 360–369, Jan. 2022.

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

[34] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2793–2801, May 2022.

[35] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkilä, "Rethinking the evaluation of video summaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7588–7596.
[36] 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.
[37] W. Zhu, J. Lu, J. Li, and J. Zhou, "DSNet: A flexible detect-to-summarize network for video summarization," *IEEE Trans. Image Process.*, vol. 30, pp. 948–962, 2021.
[38] D. Zwillinger and S. Kokoska, *CRC Standard Probability and Statistics Tables and Formulae*. Boca Raton, FL, USA: CRC Press, 1999.
[39] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, Nov. 1945.
[40] *Methodology for the Subjective Assessment of Video Quality in Multimedia Applications*, document BT.1788, ITU-R, 2007, pp. 1–13.

**HONGBEN HUANG** received the B.S. and M.S. degrees from Guilin University of Electronic Technology. He is currently an Associate Professor with Guangxi Key Laboratory of Machine Vision and Intelligent Control, School of Data Science and Software Engineering, Wuzhou University, China. His research interests include deep learning, recommendation systems, image processing, and data mining.

**ZAIQUN WU** received the M.S. degree from Guilin University of Electronic Technology. He is currently an Associate Professor with the College of Information Engineering, Baise University, China, the Director of Guangxi Artificial Intelligence Association, the Deputy Director of the Artificial Intelligence Application Engineering Research Center, Guangxi, and a High-level Talent, Baise. His research interests include signal detection and control technology, artificial intelligence, and big data.

**GUANGYAO PANG** received the M.S. degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2013, and the Ph.D. degree in computer software and theory from Shaanxi Normal University, China, in 2021. He is currently an Associate Research Fellow with Guangxi Key Laboratory of Machine Vision and Intelligent Control, School of Data Science and Software Engineering, Wuzhou University, China. His research interests include deep learning, recommendation systems, and multimodal data processing.

**JIEHANG XIE**, photograph and biography not available at the time of publication.