






Unsupervised Video Summarization Based on the Diffusion Model of Feature Fusion

Qinghao Yu , Hui Yu , Senior Member, IEEE, Ying Sun ,
Derui Ding , Senior Member, IEEE, and Muwei Jian 

Abstract—Video summarization (VS) technologies can automatically extract key frames with effective information and thus can help to quickly identify the events or speed up the decision-making process, especially for accidents. With the fast development of deep learning technologies, many generative adversarial network (GAN)- and reinforcement learning (RL)-based unsupervised VS methods have been developed in recent years. However, these methods could suffer from the problems of unstable training and difficulty of reward function formulation, respectively. To this end, we present an unsupervised VS method called diffusion model of feature fusion (DMFF) in this article, which consists of a diffusion module (DM), a feature extraction and compression module (FECM), and a coarse-fine frame selector (CFFS). DM is designed to avoid the training instability problem caused by GAN's alternate training generator and discriminator. FECM is used to extract and compress video features. CFFS is designed to capture both low-level and high-level features between frames to handle complex and diverse accident videos. Then, high-level local and global features are fused to generate a multigrained final frame score. Experiments on two widely used benchmark datasets, SumMe and TVSum, demonstrate the effectiveness and superiority of the proposed network to the state-of-the-art methods, and the training is more stable.

Index Terms—Coarse-fine frame selector (CFSS), diffusion model, feature fusion, multigrained, unsupervised video summarization.

I. INTRODUCTION

VIDEOS can provide effective information and evidence for events, accidents, and even natural disasters [1], [2], [3], [4], [5] and have been widely used in recent years. However, it is impossible to manually browse and identify key information quickly from the large volume of video data. Thus,

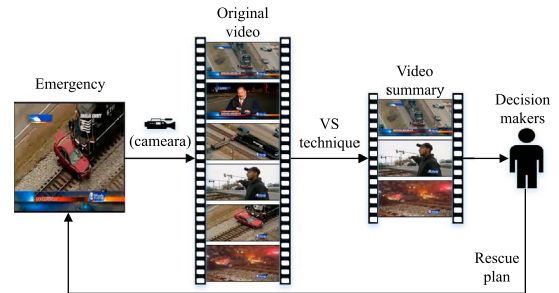


Fig. 1. Process of dealing with an accident from occurrence to decision makers based on the VS.

video summarization (VS) techniques have attracted increasing attention. VS can automatically extract key frames of information and improve the efficiency of information transmission, which can significantly reduce viewer's viewing time. In addition, the VS technology can help decision makers quickly identify incidents and thus to speed up the response process, reducing casualties and losses in accidents. Fig. 1 shows a process of identifying emergency accidents and decision-making based on the VS technology. Therefore, there is an urgent need for novel video summary methods to help viewers reduce video browsing time and improve rescue efficiency.

In recent years, deep learning technologies have been widely used with impressive performance, such as in big data [6], [7], image clustering [8], traffic flow prediction [9], intelligent vehicles [10], [11], and linguistic intelligence [12]. A large body of works on VS based on deep learning have been also carried out in recent years. In the supervised domain, to improve the ability for capturing the delicate features of videos. Lin et al. [13] designed an attention-based hierarchical long short-term memory (LSTM) module. In the unsupervised field, Pang et al. [14] directly quantified the importance of the frame level using contrastive loss in the representation learning literature and proposed three key frame indicators of local difference, global consistency, and uniqueness. The multigranularity encoder (MGE) proposed by Zhang et al. [15] integrated self-attention and temporal convolution to model coarse-grained and fine-grained contextual information. However, most of these works only consider videos taken from a single angle with a simple scenario. In many cases, videos are taken from arbitrary angles with complex details. Thus, it needs a more fine-grained network to capture global and local video features to select

Manuscript received 12 November 2023; revised 4 February 2024 and 14 March 2024; accepted 31 March 2024. Date of publication 6 May 2024; date of current version 2 October 2024. (Corresponding author: Hui Yu.)

Qinghao Yu and Derui Ding are with the School of Control Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: qinghao_yu@hotmail.com; deruiding2010@usst.edu.cn).

Hui Yu is with the School of Control Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China, and also with the School of Creative Technologies, University of Portsmouth, PO1 2DJ Portsmouth, U.K. (e-mail: hui_yu@ieee.org).

Ying Sun is with the Business School, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: yingsun1991@163.com).

Muwei Jian is with the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China, and also with the School of Information Science and Technology, Linyi University, Linyi 276000, China (e-mail: jianmuwei@163.com).

Digital Object Identifier 10.1109/TCSS.2024.3384627

key frames. Furthermore, studies in [16], [17], [18], [19], [20] considered the score of importance of each frame as the final global frame importance score. Most existing methods cannot fully explore the correlation between the predicted frame importance scores and ground-truth. It is necessary to refine the modeling of the frame importance scores in order to obtain an interesting degree score that is more consistent with the annotated frames in the label. We propose a method that uses the frame importance score obtained at the first instance as the input of another fine network for secondary modeling to obtain a more accurate frame importance score. Since the frame importance score is the most important factor in video summary generation algorithms, such as 0-1 knapsack, and for videos with accidents, the local frame importance score will have a great impact on the global frame importance score due to the paying attention to details. Thus, we design a local feature extraction algorithm based on attention windows to extract local information of video frames.

Two types of unsupervised VS techniques have received attention including generative adversarial network (GAN) [21] and reinforcement learning (RL) [22]. However, the GAN- and RL-based methods in VS face some challenges. Specifically, GAN performs alternating adversarial training through minimizing the deep feature distance between the summary generated by the generator and the original video [23], so it might encounter training instability [24]. RL-based summarization methods rely on a hand-crafted reward function to guide the agent's learning and adjust its behavior according to the reward signal for keyframe selection. It is a tough task to formulate an appropriate reward function that allows the agent to learn the desired behavior. The diffusion model (DM) [25], similar to VAE and GAN, is also a type of generative model. DM offers fine-grained control over the generation process, providing more diversity of the generated data with much more stable training process. We thus introduce the DM into the VS application for the first time. Recently, the DM has shown the huge potential in the generation task. For the GAN model, DM adopts a different training method. That is, it is trained through a probability diffusion process. The probability diffusion process gradually converts real data and generated data by gradually increasing the level of noises until the final generated data and real data approach the same distribution, which avoids the problem that GAN encounters when trying to reach a suitable balance point due to alternate training of the generator and discriminator [26], [27]. For RL, DM uses a probability-based walking process to model the agent's behavior. The DM's reward signals are computed from the agent's current state and the probability distribution of the actions taken, which does not require an explicit reward function.

Inspired by this observation, we propose a diffusion model of feature fusion (DMFF) network for unsupervised VS in this study. Compared with GAN, the DM-based framework is more stable during training. To handle complex and diverse accident videos, we propose the concept of frame score features and design a coarse-fine frame selector (CFFS) to capture low-level and high-level features between frames. Both the coarse and fine frame selectors contain LSTM networks. The purpose is

to enhance the ability of extracting the temporal information dependence between frame sequence features. We design a model frame importance scores with a hierarchical structure, rather than outputting the final frame importance scores all at once, which can gradually fit the actual frame importance scores more precisely. The proposed method can learn variable range dependencies between frames at different levels of fine-grained while maintaining the stability of the framework. Inspired by the idea of the Stacking ensemble algorithm in machine learning, we regard the frame score as a feature with a single dimension, which can handle complex and diverse accident videos more delicately. Specifically, fine frame selector (FFS) includes a global frame score feature extractor (GFSFE) and a local frame score feature extractor (LFSFE), which integrates a novel local feature extraction module (LFEM), aiming to improve the recognition degree of local feature details. Considering the different importance of local and global features, we linearly fuse these two outputs of FFS to generate the final secondary importance score. The main contributions of the method can be summarized as follows.

- 1) We propose a DM-based network for VS. It can learn variable range dependencies between frames at different levels of fine-grained while maintaining the stability of the framework, which overcomes the technical shortcomings of GAN and RL based summarization methods.
- 2) We design a CFFS module to model the frame score features, which can learn contextual information between frames more delicately. We propose a hierarchical approach to gradually extract frame score features to predict the final frame importance score, rather than obtaining it all at once. Furthermore, the output of GFSFE and LFSFE of FFS is also linearly fused to generate the final frame score.
- 3) We propose a LFEM, which can effectively extract the intrinsic dependencies between local frames in the window and improve the classification accuracy of key frames.

The rest of this article is organized as follows. Section II reviews related work. In Section III, we describe the proposed method in detail. In Section IV, we present the experimental results, comparison, and analysis. In the last section, we draw conclusion and future works.

II. RELATED WORK

In this section, we first review closely related literature on unsupervised VS based on GAN and RL. Then, we briefly introduce DM's principle. Finally we discuss related techniques for global and local dependencies related to the video.

A. Unsupervised VS

Supervised classification models need to be trained with a large volume of labeled data [28]. In this regard, unsupervised models show the advantages. Unsupervised VS technologies based on GAN and RL have attracted increasing attention. Many GAN-based VS methods have been proposed. Mahasseni et al. [18] explored the use of GAN to evaluate the distance between the reconstructed summaries and the original video for

adversarial training to generate summaries. Apostolidis et al. [17] based on [18], applied an incremental approach to train different components of the architecture proposing a stepwise, label-based learning process to improve the training efficiency of the adversarial part of the model. Apostolidis et al. [16] integrated an attention mechanism in the middle of the autoencoder layer based on the work in [17]. For the VS method of RL, [19], [29], [30] considered the diversity and representativeness of summaries, and guided the agent to generate summaries by designing a specified reward function. However, although the above-mentioned summarization methods based on GAN and RL have achieved impressive progress, the effect of VS is limited due to the intrinsic issues of the networks. The DM-based unsupervised VS method we proposed has the following advantages: 1) It overcomes the problems of unstable training of GAN and difficulty in formulating the reward function of RL. 2) Higher quality summary results can be generated.

B. DM

DM is a generative model. Its forward and reverse processes are respectively progressive noise addition and denoising procedures, which are used to asymptotically learn the potential connections between images at adjacent time steps. Before DM was developed, one of the most popular image generation frameworks was GAN. However, DM is found to have a better performance for generating higher quality images than GAN. At present, DM has been applied to many fields, but not the VS task. The specific principles of the DM are as follows: In the process of the forward diffusion stage for DM, assuming that the diffusion process satisfies the Markov chain, by continuously adding T times Gaussian noise to the original data x_0 , the mean value of the noise after parameter renormalization is determined by the fixed value $\beta_t \in (0, 1)$ and the current time t (the variance is determined by a fixed value β_t), so that the data change from the original distribution to the desired distribution. Finally, when t tends toward infinity, x_t becomes independent Gaussian distribution. Its diffusion process is shown in the following equation:

$$q(x_t | x_{t-1}) = N(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I). \quad (1)$$

Among them, x_t is the sample at time t . N represents a Gaussian distribution, β_t is a fixed value at time t , and I is an identity matrix.

According to the technique of parameter renormalization and the property of superposition of Gaussian distribution, x_t can be calculated by x_0 and β_t ($\beta_t = 1 - \alpha_t$) without iteration, and its equation is as follows:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}z \quad (2)$$

$$\bar{\alpha}_t = \prod_{i=1}^T \alpha_i \quad (3)$$

where x_0 is the target distribution, and $z \sim N(0, I)$ is Gaussian noise.

Then, $q(x_t | x_0)$ can be obtained by the following equation:

$$q(x_t | x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I). \quad (4)$$

In the process of reverse diffusion, the process is to restore the original data x_0 from Gaussian noise. When β_t is small enough, $p(x_{t-1} | x_t)$ still conforms to the Gaussian distribution [25] but cannot gradually fit the distribution due to requiring the entire dataset. Therefore, it is necessary to build a parameter distribution model p_θ to estimate that the inverse diffusion process is still a Markov chain process

$$p_\theta(x_{t-1} | x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (5)$$

The mean μ_θ and variance Σ_θ are both related to x_t and t .

The posterior diffusion conditional probability $q(x_{t-1} | x_t, x_0)$ can be expressed by the following equation:

$$q(x_{t-1} | x_t, x_0) = N(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t I) \quad (6)$$

where $\tilde{\mu}_t = (1/\sqrt{\alpha_t})(x_t - (\beta_t/\sqrt{1 - \tilde{\alpha}_t})z_t)$, $\tilde{\beta}_t = (1 - \bar{\alpha}_{t-1}/1 - \bar{\alpha}_t)$, $x_0 = (1/\sqrt{\bar{\alpha}_t})(x_t - \sqrt{1 - \bar{\alpha}_t}z_t)$, z_t is the standard Gaussian noise at time t .

The purpose of training the DM is to obtain the probability distribution of the Markov chain transformation in the inverse diffusion process through maximum likelihood estimation. The following equation to maximize the log likelihood of the model prediction distribution and to minimize the loss from the perspective of negative log-likelihood:

$$L_{\text{dif}} = \mathbb{E}_q[-\log p_\theta(x_0)]. \quad (7)$$

The final loss function of the DM is shown in the following equation:

$$L_{\text{dif}} := \mathbb{E}_{t, x_0, \varepsilon} \left[\left\| \varepsilon - \varepsilon_\theta \left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, t \right) \right\|^2 \right] \quad (8)$$

where ε is the Gaussian noise, and ε_θ is the parameter neural network to be estimated about x_t and t .

It is worth noting that [25] adopted the method of predicting random variables, by constructing a model to predict Gaussian noise instead of predicting the real data distribution.

C. Global and Local Dependencies of Videos

Modeling of local and global dependencies of videos is crucial. Early methods were mainly based on clustering and singular value decomposition to manually extract the local and global features of the video, but this cannot fully reflect the content of the video. Deep learning methods can automatically learn more comprehensive and deep features through the deep neural network [31]. For example, Andra and Usagawa [32] summarized videos using RNN by integrating the attention mechanism. To overcome the problem of RNN's insufficient dependence on long-term time, Wang et al. [33] tried to stack multiple layers of RNN. However, the consequence of stacking multilayer RNNs may lead to the problem of gradient vanishing or explosion. LSTM provides a solution to alleviate this problem to a certain extent. Zhang et al. [34] first proposed to use LSTMs to model variable sequence-wide dependencies between frames. Naturally, LSTM is still relatively inefficient in dealing with long videos and parallel processing, so the transformer framework [35] was proposed with a better solution.

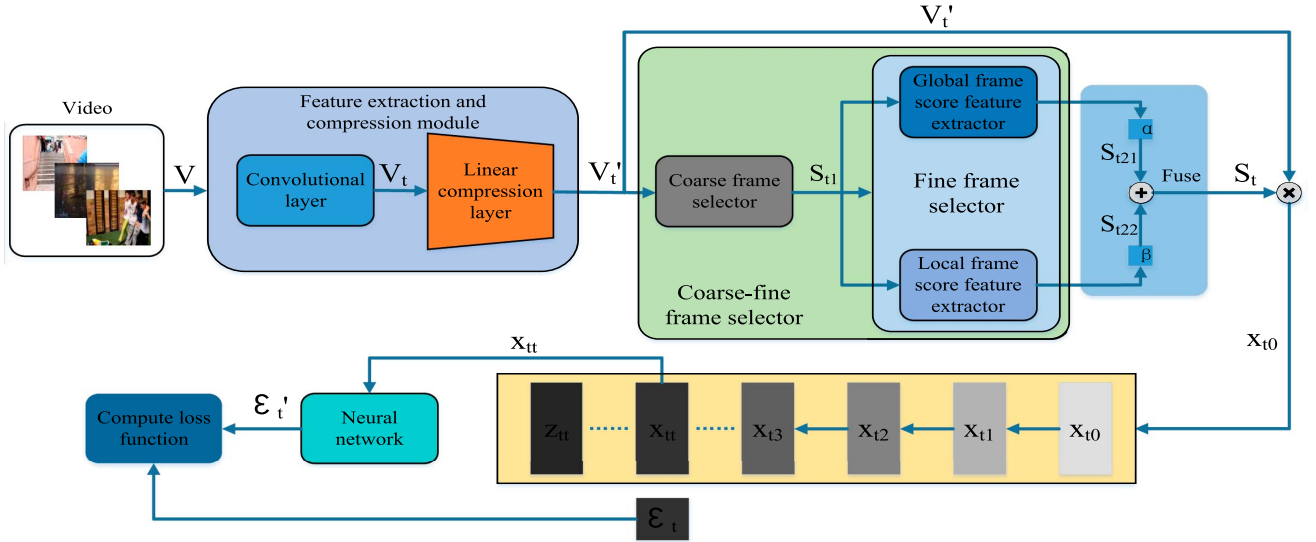


Fig. 2. Diagram of the DMFF-based network structure. The original video V passes through FECM to obtain the compressed feature V'_t of the original video. Input V'_t is then processed by CFFS to calculate the importance S_t of each frame. Then, it calculates the hadamard product of S_t and V'_t to obtain the weighted feature vector X_{t0} . The diffusion of X_{t0} through t steps finally results in X_{tt} . We use X_{tt} to predict the sampling noises ϵ'_t in time step t through a neural network and then calculate the loss with the actual noises ϵ_t to update the parameters.

Furthermore, Apostolidis et al. [36] combined global and local multihead attention mechanisms to discover different levels of frame dependencies to model. Zhao et al. [37] exploited the hierarchical structure of videos using LSTM to capture the correlation between frames within a shot, and then GNN was used to model the global dependencies between shots and shots. Our method differs from the above work in that we design CFFS for the complex diversity of accident videos, where CFS and FFS are used to capture low-level and high-level features between video frames, respectively. Furthermore, GFSFE and LFSFE in FFS capture the global and local feature dependencies of high-level features [38], respectively. Finally, considering the differences in the decision makers' attention to the global and local features of different accident videos, the outputs of GFSFE and LFSFE are linearly fused with different weights to find a pair of more robust hyperparameters when it can also maintain stable performance for complex and diverse accident video.

III. THE PROPOSED METHOD

In this section, we elaborate the network structure of DMFF. Fig. 2 illustrates the structure of DMFF. DMFF consists of a feature extraction and compression module (FECM), designed for feature extraction and compression of input video, a coarse-fine frame selector CFFS, designed to learn contextual information at different granularity levels from video sequences and diffusion module.

A. CFFS

Although traditional VS methods take into account the global and local dependencies of videos, one issue is often ignored. That is, viewers usually spend different time in understanding the general content of the video (global features) and the details of a certain event (local features). The details actually need

more attention, so that the viewer can make a more accurate rescue plan. Furthermore, we treat the frame score as a feature and focus on improving the accuracy of the predicted frame score. In this work, CFFS is designed to fully mine the potential connection between low-level and high-level features based on frame score features and global and local features.

1) *Coarse Frame Selector*: The CFS module is designed to capture the low-level feature dependencies of video sequences. CFS contains two layers of bidirectional LSTM including a fully connected layer and a Sigmoid function activation layer. These two layers of bidirectional LSTM are used to capture the intrinsic relationship between forward and reverse low-level features between frame sequences. The sigmoid function is used to keep frame scores in the (0,1) range. Input M frames of video V into the FECM to obtain feature $V'_t \in \mathbb{R}^{M \times 1 \times d_c}$. Then, V'_t is used as the input of CFS to obtain the first-level frame importance score $S_{t1} \in \mathbb{R}^{M \times 1}$. The CFS process can be expressed using the following equation:

$$S_{t1} = \text{CFS}(V'_t). \quad (9)$$

2) *Fine Frame Selector*: High-level semantic features of videos such as actions, scenes, and emotions are crucial for people to understand videos, which are much more complex and abstract than low-level features. Previous work regards the output of CFS directly as the final frame score. However, we design FFS to extract the corresponding high-level feature dependency information to fully explore the relationship between frame score and high-level features.

GFSFE is used to capture the high-level global information of the video. The process is shown in Fig. 3. It consists of global linear layer 1 (the number of hidden layer units is assumed to be f_h), global LSTM (single-layer LSTM), and global linear layer 2. The global linear layer 1 maps the low-dimensional frame score features of the hidden space to a higher dimension,

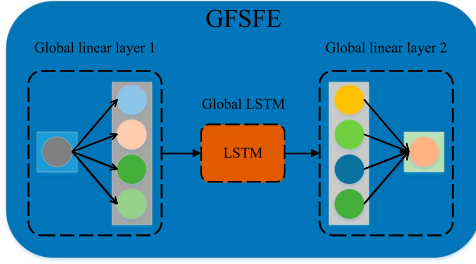


Fig. 3. Structure of GFSFE.

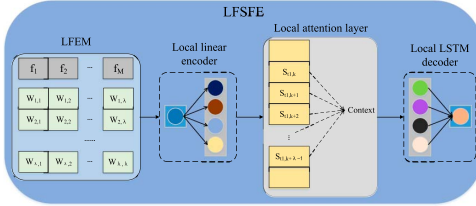


Fig. 4. Structure of LFSFE.

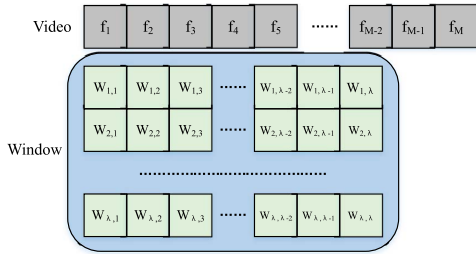


Fig. 5. Schematic diagram of local frame score feature extraction module.

which is convenient for the global LSTM to capture richer global feature information dependencies. Finally, the output of the global LSTM is linearly mapped to the global linear layer 2. We can then obtain the frame score feature that the global advanced feature of the video depends on.

Since the local features of video are more susceptible to noise interference, it is more difficult to obtain high-quality local features compared with global features. Traditional methods use motion estimation or optical flow estimation techniques, which rely on hand-designed features. The proposed LFSFE can automatically extract the high-level local feature information dependence of the video. The structure of the LFSFE is shown in Fig. 4.

LFSFE consists of a LFEM, a local linear encoder, a local attention mechanism [39], and a local LSTM decoder. Zhou et al. [19] proposed that when the distance between two frames of the video is greater than the λ frame, the correlation between the two frames will become very small. To prevent out-of-frame influences during local feature extraction, we thus set up a learnable attention window $W \in \mathbb{R}^{\lambda \times \lambda}$, which is used to learn high-level local correlations between frames inside the window. The schematic diagram of the algorithm is shown in Fig. 5, and the process is as follows. 1) The left half (from f_1 to $f_{M-\lambda}$)

Algorithm 1 Pseudocode of LFSFE algorithm

Input: the output S_{t1} of the coarse frame selector.

Output: local frame scores on S_{t1} .

```

1: for  $i$  each  $S_{t1} \in [1, 2, \dots, M]$  do
2:   if  $i \leq M - \lambda$  then
3:     for  $l$  in  $S_{t1\_left}$  do
4:        $S_l \leftarrow AttWin(l)$ 
5:     end for
6:   else
7:     for  $r$  in  $S_{t1\_right}$  do
8:        $S_r \leftarrow AttWin(r)$ 
9:     end for
10:  end if
11: end for
12:  $S_{lr} \leftarrow S_l + S_r$ 
13: return  $S_{lr}$ 

```

frame score feature extraction. We slide the window from left to right on the video frame, align the leftmost column of the window with $f_j, j \in (1, 2, \dots, M - \lambda - 1, M - \lambda)$, then the rightmost column of the window will be aligned with $f_{j+\lambda-1}$. That is, let the frames between f_j and $f_{j+\lambda-1}$ pass through the learnable window matrix to get $S_j \in \mathbb{R}^{\lambda \times I}$, which only considers the frame score of the connection between the frames in the current window. When the window continues to slide to the right to align with f_{j+1} , the calculation of S_{j+1} needs to consider the influence of its previous $\lambda - 1$ frame, that is, $f_i, i \in (t - \lambda + 2, t - \lambda + 3, \dots, t - I, t)$ until the end of $j = M - \lambda$. 2) Right half (from $f_{M-\lambda+1}$ to f_M) frame score feature extraction. It is worth noting that when extracting this part of features, we extract frames from f_M to $f_{M-\lambda+1}$. The process is similar to that of (1). The pseudo code of the specific algorithm is shown in Algorithm 1. After S_{t1} passes through the local frame feature extraction module, a relative frame score that only considers the dependencies around the frame can be obtained. Then, it is passed through the local linear encoder. Its function is similar to that of the global linear layer 1 in GFSFE. The subsequent local attention mechanism is to integrate the values in the attention window range to obtain local high-level semantic frame scores. It is finally decoded into a frame score with the same shape as S_{t1} through the local LSTM decoding layer. The frame score only considers the relative score between itself and the surrounding window size.

After obtaining different score features of the fine-grained advanced global and local frames, we perform linear feature fusion of the features. Specifically, we set a pair of hyperparameters α, β ($\beta = 1 - \alpha$), where α is called the fusion factor. We set the proportion of global features as α and the proportion of local features as β . Finally, the two weighted features are added together to obtain the final frame scores of features with different importance degrees. This process can be expressed by the following equation:

$$S_t = \alpha \cdot GFSFE(S_{t1}) + \beta \cdot LFSFE(S_{t1}), \quad 0 < \alpha < 1. \quad (10)$$

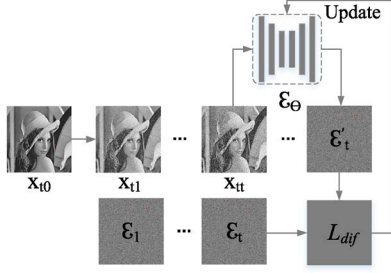


Fig. 6. Diffusion process structure with “Lena” as an example.

B. Diffusion-Based Process VS

Unsupervised VS tasks usually use the GAN or RL technology [17], [18], [29], [40]. However, the quality of unsupervised VS is hard to be further improved due to the intrinsic characteristic of the network models. To this end, we propose a DM-based model, as shown in Fig. 6.

Specifically, suppose we diffuse a total of T time steps, applying Gaussian noise at each time step. We input the weighted fusion feature X_{t0} , that is, the data to be fitted, into the DM. For step $t (1 \leq t \leq T)$, we can use the conditional probability equation (1) of the forward diffusion process to obtain the characteristic distribution of the sampling value at time t as follows:

$$Q(X_{tt}|X_{t(t-1)}) = N(X_{tt}; \sqrt{1 - \beta_t}X_{t(t-1)}, \beta_t I) \quad (11)$$

$$X_{tt} = \sqrt{\alpha_t}X_{t0} + \sqrt{1 - \alpha_t}\varepsilon_t \quad (12)$$

where ε_t is the random Gaussian noise generated under t time step.

The characteristics at time t can be directly calculated from X_{t0} and β_t without iteration.

It is worth noting that in the process from $X_{t(t-1)}$ to X_{tt} , we do not directly apply Gaussian noise on $X_{t(t-1)}$ but use parameter renormalization techniques to prevent the network from failing to perform gradients during the reverse process. And in order to improve the generalization performance of the model, instead of obtaining X_{tt} sequentially, we randomly generate a t (t is a learnable embedding that encodes the position encoding information of its current time step) and obtain the sampled value X_{tt} at the time step.

It is worth noting that we only need to train the learnable parameters in ε_θ and CFFS because our input is the original video to generate a summary without the need of generating new data through the network. The purpose of our network construction is only to output a frame score close to the ground truth. Once X_{tt} is obtained, we input it into the neural network ε_θ (3 linear layer stacks). The output ε'_t is used to fit the Gaussian noise ε_t under t time step, which is then used to calculate the loss L_{dif} for backpropagation to update the learnable parameters in ε_θ and CFFS. The loss function of the diffusion process is shown in the following equation:

$$L_{dif} = \frac{1}{M} \sum_{i=1}^M \|\varepsilon_t - \varepsilon'_t\|_2^2. \quad (13)$$

Algorithm 2 DMFF training process pseudocode

Repeat
 $V_t \sim q(V_t)$
 $t \sim \text{Uniform}(\{1, 2, \dots, T\})$
 $\varepsilon \leftarrow (0, I)$
 $V_t \leftarrow \text{CNN}(V)$
 $X_{t0} \leftarrow \text{CFFS}(V'_t) \sim q'(X_{t0})$
 Take gradient descent step on
 $\nabla_\theta (\|\varepsilon - \varepsilon_\theta(\sqrt{\alpha_t}X_{t0} + \sqrt{1 - \alpha_t}\varepsilon, t)\|^2 + \|V_t - X_{t0}\|^2 + \|\theta\|^2)$
 until converged

C. Optimization and Inference

In our work, ε_θ and CFFS are optimized in an end-to-end manner. Specifically, we use the three loss functions of L_{dif} , L_{fus} , and L_{weg} to form the total loss function as follows:

$$L = L_{dif} + L_{fus} + L_{weg}. \quad (14)$$

For each video, the mean square error (MSE) loss between the original video feature V_t and the weighted feature X_{t0} is used to calculate the fusion loss L_{fus} as follows:

$$L_{fus} = \frac{1}{N} \sum_{t=1}^N \|V_t - X_{t0}\|_2^2 \quad (15)$$

where N represents the number of all videos in the training set. The dimensions of V_t and X_{t0} are not the same, so we use a singular value decomposition (SVD) for V_t to preserve its maximum amount of information. In addition, optimizing L_{fus} is equivalent to making X_{t0} closer to V_t , without directly inputting V_t into the DM. The purpose is that we need to model the feature dependencies of different granularity levels on the frame score features to generate a model that is consistent with human subject's option.

In addition, in order to prevent the model from overfitting, we apply an L_2 regularization term on the model weight parameter θ

$$L_{weg} = \sum_{i,j} \theta_{i,j}^2. \quad (16)$$

We summarize the entire training process of DMFF in Algorithm 2. To generate summaries, keyframes are chosen to maximize the total score while ensuring that the length of the summaries is less than a predefined limit, which is set to be 15% of the number of frames in the original video, as shown in [19] and [34]. The maximum problem is modeled as a 0-1 knapsack problem and solved with dynamic programming. After the training process of DMFF, we obtain CFFS and ε_θ that can learn the distribution of key information in the video.

IV. EXPERIMENTS

In this section, we describe the designed experiments to verify the effectiveness of the proposed method. First, we elaborate the experimental setup, including datasets, preprocessing and implementation details, and evaluation metrics. To demonstrate

the superiority of the proposed model, we compare DMFF with other methods. We then perform ablation studies to illustrate the impact of key components in the model. Furthermore, we visualize the generated summaries. Finally, a sensitivity analysis is performed on the hyperparameters of the model. We use a computer with an NVIDIA GeForce 3090 GPU and Intel(R) Xeon(R) W-2265 CPU to implement our proposed framework on the Pytorch platform.

A. Experiment Settings

1) *Datasets*: This study uses two widely used public benchmark datasets for VS: SumMe [41] and TVSum [42] to evaluate the proposed method. SumMe consists of 25 user videos, each with a duration of 1–6 min, and each video is annotated by 15–18 users in the form of key clips. This dataset covers a lot of topics such as sports and holidays. TVSum consists of 50 user videos. And each video lasts 2–10 min, and each video is annotated by 20 users, including news, documentaries, and other topics.

2) *Preprocessing and Implementation Details*: Following [34], each video is down-sampled to two frames per second to remove redundancy. The Adam optimizer is used with a learning rate of 10^{-5} . The gradient clipping threshold is 5, and the default number of hidden layer units for ε_θ is 128. The deep convolutional neural network uses the output feature $V_t = \{f_j\}_1^M$, where t denotes time, j indicates the j th frame of the video, $f_j \in \mathbb{R}^{d_f}$, $d_f = 1024$. Then, V_t is passed to a linear compression layer (the number of nodes in the hidden layer is $h = 512$) to obtain the compressed feature $V'_t \in \{f'_j\}_1^M$.

B. Evaluation Metrics

For a fair comparison with existing methods, F-score is used as a similarity measure between automatic summarization and ground truth summarization. The automatically generated summaries are denoted by A , and the ground truth summaries are denoted by G . We calculate the precision rate P and recall rate R of each pair of A and G as a measure of their overlapping frames, and F-Score is the comprehensive performance of the two

$$P = \frac{|A \cap G|}{|A|} \times 100\% \quad (17)$$

$$R = \frac{|A \cap G|}{|G|} \times 100\% \quad (18)$$

$$F = 2 \times \frac{P \times R}{P + R} \times 100\%. \quad (19)$$

A higher F-score indicates that the automatic summarization has a large overlap with the ground truth summaries while containing less redundancy.

C. Analysis and Discussion

1) F-Score Performance and Loss:

a) *F-Score performance*: In the dataset used, the training set accounts for 80% of the total, and the remaining 20% is used as the test set. We have iteratively trained our model 100 times [16], [17]. Finally, we perform five-fold cross-validation

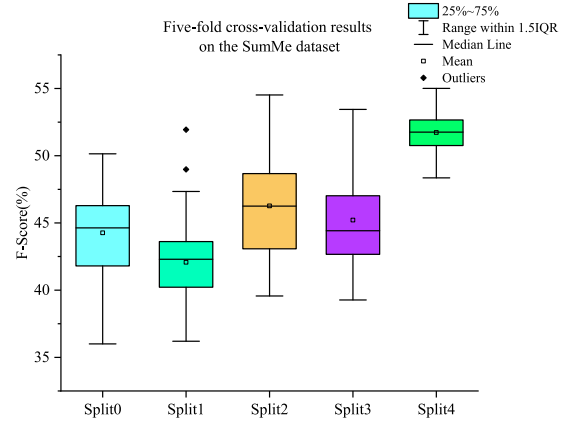


Fig. 7. F-score data distribution on SumMe.

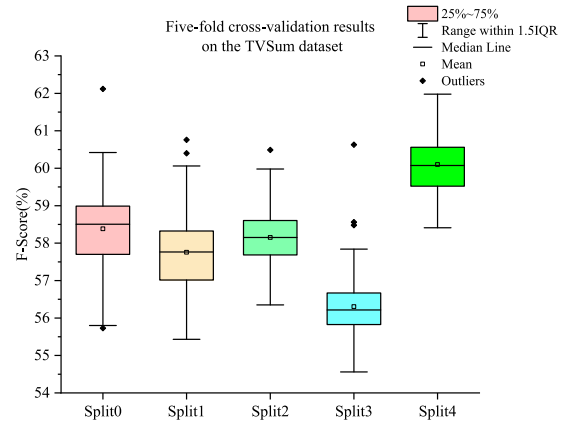


Fig. 8. F-score data distribution on TVSum.

and report the average performance of these splits. The F-score of DMFF on the SumMe and TVSum datasets is shown in Figs. 7 and 8, respectively.

From Figs. 7 and 8, we can draw the following conclusions. 1) The performance on TVSum is far better than SumMe. This is because the TVSum dataset contains more videos, so the model can learn more diversified features. 2) There are more abnormal points on TVSum than on SumMe. The reason behind this may be that there are more noises in the former video during shooting.

b) *Loss*: GAN is prone to training instability. For example, the loss function curve in the VS method [16] using GAN appears this phenomenon after iterating for many generations. This indirectly causes the performance of VS to suffer. However, DMFF overcomes this shortcoming, and its five-fold validation's average loss function curves on SumMe and TVSum are shown in Figs. 9 and 10.

It shows that the loss curve of DMFF can reach the convergence slightly faster (around the 30th epoch), and the training is easier than the GAN-based unsupervised video summary method [16].

2) *Comparisons*: For a fairer comparison, we use two evaluation methods, namely multiple user-annotated ground truth summaries and single ground truth summaries. The latter is

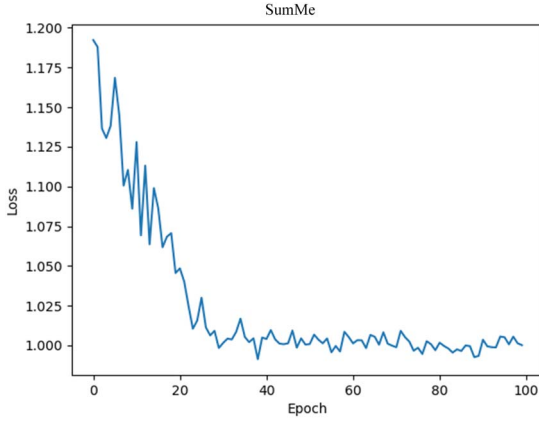


Fig. 9. F-score data distribution on SumMe.

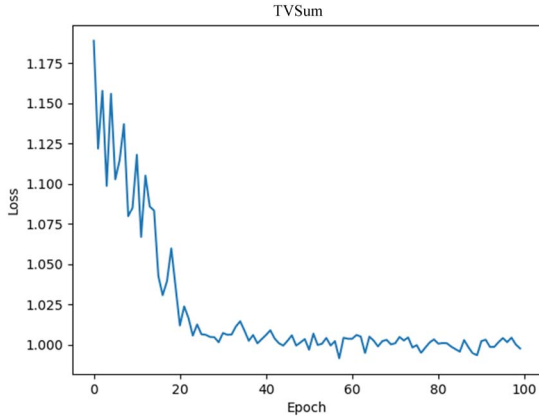


Fig. 10. F-score data distribution on TVSum.

TABLE I
F-SCORE (IN PERCENT) COMPARISON USING MULTIUSER
SUMMARY PROTOCOL ANNOTATION WITH OTHER
UNSUPERVISED METHODS

Methods	SumMe	TVSum	Average
*SUM-GDA [45]	50.0 (−)	59.6 (−)	54.8
*SUM-GAN-sl [17]	47.3 (−)	58.0 (−)	52.65
*SUM-GAN-AAE [16]	48.9 (−)	58.3 (−)	53.6
~DR-DSN [19]	41.4 (−)	57.6 (−)	49.5
~DSR-RL-LSTM [46]	43.8 (−)	61.4 (+)	52.6
~DSR-RL-GRU [46]	50.3 (−)	60.2 (−)	55.25
~AC-SUM-GAN [40]	50.8 (−)	60.6 (−)	55.7
~PRLVS [47]	46.3 (−)	63.0 (+)	54.65
RS-SUM [48]	52.0 (−)	61.1 (−)	56.55
SSPVS [49]	48.7 (−)	60.3 (−)	54.5
AMFM [15]	51.8 (−)	61.0 (−)	56.4
DMFF (Ours)	53.01	61.2	57.12

Note: The symbol “+” means better, and “−” means worse than our method. The subscript “uns” means unsupervised. The asterisk “*” means GAN-based methods. The symbol “~” means RL-based method. The bold entries indicate the best performance.

mainly for comparison with some current specific and classical methods [18], [43], [44] for comparison. Under the evaluation using the former method, we compare the performance of DMFF with other state-of-the-art unsupervised methods on SumMe and TVSum, as shown in Table I.

TABLE II
SUMMARY PROTOCOL USING A SINGLE GROUND
GROUND-TRUTH SUMMARY F-SCORE (%)
COMPARED TO OTHER METHODS

Methods	SumMe	TVSum
SUM-GAN [18]	38.7 (−)	50.8 (−)
SUM-GAN _{dpp} [18]	39.1 (−)	51.7 (−)
SUM-GAN _{sup} [18]	41.7 (−)	56.3 (−)
Cycle-SUM [50]	41.9 (−)	57.6 (−)
A-AVS [44]	43.9 (−)	59.4 (−)
M-AVS [44]	44.4 (−)	61.0 (−)
AALVS [43]	46.2 (−)	63.6 (−)
SUM-GAN-sl [17]	46.8 (−)	65.3 (−)
SUM-GAN-AAE [16]	56.9 (−)	63.9 (−)
AC-SUM-GAN [40]	60.7 (−)	64.8 (−)
DMFF (Ours)	62.65	65.33

Note: “±” means better or worse than our method. The bold entries indicate the best performance.

TABLE III
RESULTS OF THE ABLATION EXPERIMENT

FSFF	LFEM	SumMe			TVSum		
		F	P	R	F	P	R
	Baseline	48.58	46.22	52.13	58.86	58.80	58.94
✓	×	51.42	50.14	53.18	60.25	60.13	60.38
×	✓	48.99	47.53	51.55	60.73	60.76	60.72
✓	✓	53.01	51.44	55.38	61.20	61.27	61.14

Note: F (in Percent) is the F-score, P (in Percent) is the precision rate, R (in Percent) is the recall rate, and baseline is only the diffusion probability module. The bold entries indicate the best performance.

Our method achieves the best performance in general. The performance of the proposed DMFF on SumMe outperforms other methods. Our method outperforms the second-best method RS-SUM by 1.01%. On TVSum, deep self-attention recurrent summarization network with reinforcement learning (DSR-RL-LSTM), restorative score video summarizer (RS-SUM) comparable and AMFM to our method. Finally, in terms of average performance, our method achieves the best performance, outperforming the second place RS-SUM by 0.57%. GAN-based and RL-based VS methods have similar performance, but they face challenges due to the inherent characteristics of these two types of network architectures. For the RS-SUM method, which has the second highest average score, our method has comparable performance on TVSum, but the F-score on SumMe is about 1% higher than RS-SUM. Judging from the performance of F-score, our method is more competitive than the SOTA method. This is mainly attributed to our proposed DM-based VS framework, quadratic modeling of frame importance scores, novel local feature extraction algorithms, and the fusion of global and local features.

We also evaluate our model performance using a single ground truth summary per video, as shown in Table II.

Table II shows that our method also achieves the best performance. On SumMe, our method outperforms the second place AC-SUM-GAN by 1.95%. And on TVSum, our method also achieves the best performance. It is worth noting that SUM-GAN-sl [17] achieves the second best performance because it adopts a novel label-based incremental training method. However, the performance of SUM-GAN-sl on SumMe is much worse than ours.

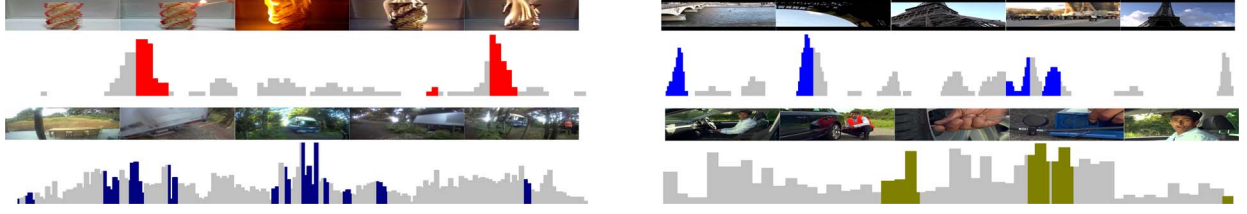


Fig. 11. Partial summary visualization results generated by DMFF. Gray is ground truth, color is the summary result of model prediction, and the height of the column indicates the importance of the frame. The above two video summaries are from videos 11 and 9 of SumMe, respectively, and the two below are videos 6 and 45 of TVSum

In general, as shown in Tables I and II, our method achieves the best performance regardless of summarization protocols demonstrating the superiority of the proposed method. This demonstrates that our method can learn key information from videos for generating high-quality summaries without supervision.

3) *Ablation Study*: In this section, we explore the impact of frame scores feature fusion (FSFF) and LFEM. It should be noted that we do not take the DM as the object of exploration because the diffusion probability module is the premise of our optimization of the objective function.

The results of the ablation experiments in Table III show that our baseline model is already comparable to those unsupervised VS methods based on GAN and RL in Table I, which demonstrates the effectiveness of the diffusion probability model. Through the fusion of frame score features, the three indicators on the two datasets have been significantly improved overall. This means that the global features and local features of the video are critical for video representation and the important relationship between them. On the one hand, we can see that the effect of LFEM is very clear on TVSum. The reason may be that TVSum's annotations between frames are more fine-grained, so it can better guide the model to learn the dependencies between global and local frames. On the other hand, LFEM appears to sacrifice the recall rate for accuracy on small datasets, because LFEM may be more inclined to capture some obvious local features, but miss some less apparent but equally important local features. Furthermore, as the number of samples increases, this phenomenon can be gradually eased. It is worth noting that the overall recall rate on the two datasets is greater than the precision rate. This means that our method can capture important information in the video, which is in line with actual needs.

We visualize some summaries generated by DMFF, as shown in Fig. 11. The results show that most of the keyframes predicted by DMFF overlap with the peaks of the ground truth indicating that our method can capture important information in the video and generate summaries similar to the subjective feelings of humans.

4) *Hyperparameter Sensitivity Analysis*: In this subsection, we perform a hyperparameter sensitivity analysis. Specifically, we explore the influence of the fusion factor α , the number of hidden layer units h of ε_θ , and the number of hidden layer units f_h of the linear layer in FFS.

a) *Fusion factor α* : The fusion factor α aims to determine a pair of more robust hyperparameters to balance the

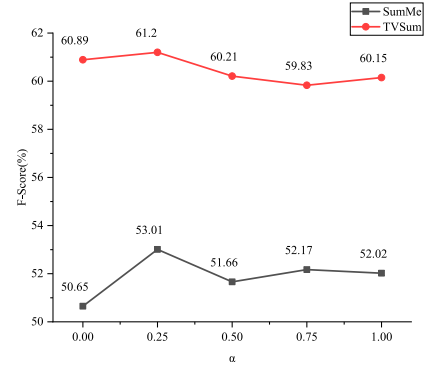


Fig. 12. Impact of fusion factor α on the model.

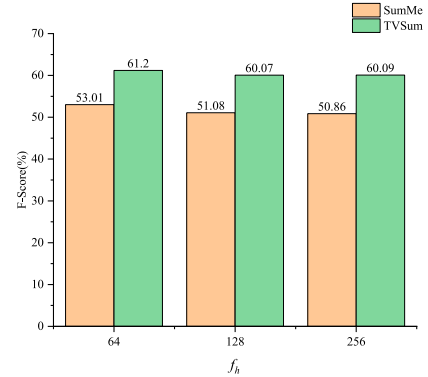


Fig. 13. Impact of f_h on the model.

importance of global frame score features and local frame score features to locate the various accident video. As shown in Fig. 12, the performance on both datasets first increases and then decreases with α . When α increases to 0.25, the model performance reaches the best. This shows that it is effective to model the frame importance scores in a more fine-grained manner and assign different weights to local and global features. Furthermore, as the video scene becomes more complex, local features play a more critical role in the performance of the final summary.

b) *The number of hidden layer units f_h of the frame selector*: f_h affects the complexity of the frame selector and the size of the intermediate output feature dimension. Since we design linear layers in both GFSFE and LFSFE, we would explore the effect of the number of its hidden layer units on the model. By comparing Fig. 13, we find that the difference

TABLE IV
EFFECT OF h ON MODEL
PERFORMANCE

h	SumMe	TVSum
64	51.16	59.83
128	53.01	61.2
256	52.45	59.8

of the f_h values has different effect on the performance of the model. When $f_h = 64$, the performance on the two datasets is generally better than 128 and 256. The reason is that $f_h = 64$ is more in line with the capacity of the model. When f_h is too large, it will cause overfitting.

c) *The number of hidden layer units h of ε_θ :* ε_θ is the core parameter of our trained network. Only when ε_θ fits the noise distribution at the current time step, we can obtain higher quality summary results. The influence of h on the model is shown in Table IV ($f_h = 64, \alpha = 0.25$). When h increases to 128, the model achieves the best performance. When h continues to increase, the performance of the model decreases instead. The reason is that the increase of the number of hidden layer neurons may lead to increasing redundant information during the learning process, making it more difficult for the network to focus on effective features for the video summary task.

V. CONCLUSION AND FUTURE WORK

In this article, we present a novel DMFF-based unsupervised VS method. This method can overcome the shortcomings of methods that are based on GAN and RL. Furthermore, to be able to handle complex and diverse accident videos, we design CFFS to capture both low-level and high-level features in videos. And the learned window of LFEM can effectively capture the local dependencies of frames in the window and improve the detail recognition of key frames. The results of ablation experiments illustrate the effectiveness of our proposed modules, and the comparison results with the state-of-the-art methods demonstrate the superiority of our method. In future work, we will consider fusing information from multiple modalities. For example, sound signals in the video can be used as an additional input to the model to further improve accuracy of VS.

REFERENCES

- [1] S. Ghosh, S. Maji, and M. S. Desarkar, "Unsupervised domain adaptation with global and local graph neural networks under limited supervision and its application to disaster response," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 2, pp. 551–562, Apr. 2023.
- [2] F. K. Sufi and I. Khalil, "Automated disaster monitoring from social media posts using AI-based location intelligence and sentiment analysis," *IEEE Trans. Comput. Social Syst.*, early access, 2022.
- [3] A. Srinivasan, A. Srikanth, H. Indrajit, and V. Narasimhan, "A novel approach for road accident detection using DETR algorithm," in *Proc. Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA)*, Valencia, Spain, 2020, pp. 75–80.
- [4] E. Pinto, E. Nepomuceno, and A. Campanharo, "Individual-based modelling of animal brucellosis spread with the use of complex networks," *Int. J. Netw. Dyn. Intell.*, vol. 1, no. 1, pp. 120–129, 2022.
- [5] F. Reggina and E. Indriani, "Psychological education in overcoming trauma due to natural disasters," *Socio-Econ. Humanistic Aspects Township Ind.*, vol. 1, no. 2, pp. 160–165, 2023.
- [6] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: A survey," *J. Internet Technol.*, vol. 21, no. 2, pp. 393–405, 2020.
- [7] J. Dou and S. Yan, "An improved generative adversarial network with feature filtering for imbalanced data," *Int. J. Netw. Dyn. Intell.*, vol. 2, no. 4, 2023, Art. no. 100017.
- [8] X. Yan, Y. Mao, M. Li, Y. Ye, and H. Yu, "Multitask image clustering via deep information bottleneck," *IEEE Trans. Cybern.*, vol. 54, no. 3, pp. 1868–1881, Mar. 2024.
- [9] C. Chen, K. Li, S. G. Teo, X. Zou, K. Li, and Z. Zeng, "Citywide traffic flow prediction based on multiple gated spatio-temporal convolutional neural networks," *ACM Trans. Knowl. Discovery Data (TKDD)*, vol. 14, no. 4, pp. 1–23, 2020.
- [10] H. Yu, Y. Wang, Y. Tian, H. Zhang, W. Zheng, and F.-Y. Wang, "Social vision for intelligent vehicles: From computer vision to foundation vision," *IEEE Trans. Intell. Veh.*, vol. 8, no. 11, pp. 4474–4476, Nov. 2023.
- [11] S. Teng et al., "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Trans. Intell. Veh.*, vol. 8, no. 6, pp. 3692–3711, Mar. 2023.
- [12] F.-Y. Wang, Q. Miao, X. Li, X. Wang, and Y. Lin, "What does ChatGPT say: The DAO from algorithmic intelligence to linguistic intelligence," *IEEE/CAA J. Automatica Sinica*, vol. 10, no. 3, pp. 575–579, Mar. 2023.
- [13] J. Lin, S.-h. Zhong, and A. Fares, "Deep hierarchical LSTM networks with attention for video summarization," *Comput. Elect. Eng.*, vol. 97, 2022, Art. no. 107618.
- [14] Z. Pang, Y. Nakashima, M. Otani, and H. Nagahara, "Contrastive losses are natural criteria for unsupervised video summarization," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, 2023, pp. 2010–2019.
- [15] Y. Zhang, Y. Liu, and C. Wu, "Attention-guided multi-granularity fusion model for video summarization," *Expert Syst. Appl.*, vol. 249, 2024, Art. no. 123568.
- [16] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *Proc. 26th Int. Conf. MultiMedia Model.*, Daejeon, South Korea, Springer, 2020, pp. 492–504.
- [17] E. Apostolidis, A. I. Metsai, E. Adamantidou, V. Mezaris, and I. Patras, "A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization," in *Proc. 1st Int. Workshop AI Smart TV Content Prod., Access Del.*, 2019, pp. 17–25.
- [18] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2982–2991.
- [19] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, doi: 10.1609/aaai.v32i1.12255.
- [20] M. Shim, T. Kim, J. Kim, and D. Wee, "Masked autoencoder for unsupervised video summarization," 2023, *arXiv:2306.01395*.
- [21] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [22] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Reinforcement Learn.*, vol. 8, no. 3–4, pp. 229–256, May 1992.
- [23] Z. Yuan, H. Li, J. Liu, and J. Luo, "Multiview scene image inpainting based on conditional generative adversarial networks," *IEEE Trans. Intell. Veh.*, vol. 5, no. 2, pp. 314–323, Jun. 2020.
- [24] Q. Yu, H. Yu, Y. Wang, and T. D. Pham, "SUM-GAN-GEA: Video summarization using GAN with Gaussian distribution and external attention," *Electronics*, vol. 11, no. 21, 2022, Art. no. 3523.
- [25] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2015, pp. 2256–2265.
- [26] F. J. Tey, T.-Y. Wu, Y. Wu, and J.-L. Chen, "Generative adversarial network for simulation of load balancing optimization in mobile networks," *J. Internet Technol.*, vol. 23, no. 2, pp. 297–304, 2022.
- [27] L. Zhao, Y. Zhang, and Y. Cui, "A multi-scale u-shaped attention network-based gan method for single image dehazing," *Human-Centric Comput. Inf. Sciences*, vol. 11, Oct. 2021.
- [28] J. Wang, C. Zhao, S. He, Y. Gu, O. Alfarrarj, and A. Abugabah, "Logquad: Log unsupervised anomaly detection based on Word2Vec," *Comput. Syst. Sci. Eng.*, vol. 41, no. 3, 2022, Art. no. 1207.
- [29] Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, "Deep reinforcement learning for query-conditioned video summarization," *Appl. Sci.*, vol. 9, no. 4, 2019, Art. no. 750.

- [30] S.-S. Zang, H. Yu, Y. Song, and R. Zeng, "Unsupervised video summarization using deep non-local video summarization networks," *Neurocomputing*, vol. 519, pp. 26–35, Jan. 2023.
- [31] G. Zhao, Y. Li, and Q. Xu, "From emotion AI to cognitive AI," *Int. J. Netw. Dyn. Intell.*, vol. 1, no. 1, pp. 65–72, 2022.
- [32] M. B. Andra and T. Usagawa, "Automatic lecture video content summarization with attention-based recurrent neural network," in *Proc. Int. Conf. Artif. Intell. Inf. Technol. (ICAIIIT)*, Yogyakarta, Indonesia, 2019, pp. 54–59.
- [33] J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan, "Stacked memory network for video summarization," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 836–844.
- [34] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, Springer, Oct. 2016, pp. 766–782.
- [35] S. Feng, Y. Xie, Y. Wei, J. Yan, and Q. Wang, "Transformer-based video summarization with spatial-temporal representation," in *Proc. 8th Int. Conf. Big Data Inf. Anal. (BigDIA)*, Guiyang, China, 2022, pp. 428–433.
- [36] 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)*, Naples, Italy, 2021, pp. 226–234.
- [37] 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.
- [38] J. Qian, M. Pan, W. Tong, R. Law, and E. Q. Wu, "URRNet: A unified relational reasoning network for vehicle re-identification," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11156–11168, Sep. 2023.
- [39] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.
- [40] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3278–3292, Aug. 2021.
- [41] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Springer, Sep. 2014, pp. 505–520.
- [42] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TvSum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Boston, MA, USA, 2015, pp. 5179–5187.
- [43] T.-J. Fu, S.-H. Tai, and H.-T. Chen, "Attentive and adversarial learning for video summarization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2019, pp. 1579–1587.
- [44] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, Jun. 2020.
- [45] 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.
- [46] A. Phaphuangwittayakul, Y. Guo, F. Ying, W. Xu, and Z. Zheng, "Self-attention recurrent summarization network with reinforcement learning for video summarization task," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Shenzhen, China, 2021, pp. 1–6.
- [47] G. Wang, X. Wu, and J. Yan, "Progressive reinforcement learning for video summarization," *Inf. Sci.*, vol. 655, Jan. 2024, Art. no. 119888.
- [48] M. Abbasi and P. Saeedi, "Adopting self-supervised learning into unsupervised video summarization through restorative score," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Kuala Lumpur, Malaysia, 2023, pp. 425–429.
- [49] 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.*, Waikoloa, HI, USA, 2023, pp. 5584–5593.
- [50] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-Sum: Cycle-consistent adversarial LSTM networks for unsupervised video summarization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 9143–9150.



Qinghao Yu is currently working toward the master's degree in control engineering with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China.

His research interests include deep learning and video summarization.



Hui Yu (Senior Member, IEEE) received the Ph.D. degree in computing from the Brunel University, London, U.K., in 2009.

He worked with the University of Glasgow and Queen's University, Belfast, U.K. Since 2012, he has been a Professor of Visual Computing with the University of Portsmouth, Portsmouth, U.K. His research interests include visual computing, cognition and machine learning, particularly 3-D/4-D facial expression reconstruction and perception, and image analysis for human-machine/

social interaction and intelligent vehicles.

Dr. Yu serves as an Associate Editor for IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, and IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS.



Ying Sun received the B.Sc. degree in physics from Harbin Normal University, Harbin, China, in 2013, and the Ph.D. degree in control theory and control engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2021.

She is currently a Postdoc with the Business School, University of Shanghai for Science and Technology. She is an Active Reviewer for many international journals. Her research interests include nonlinear stochastic control and filtering, as well as networked control systems, stochastic control and

filtering as well as H_∞ control and ℓ_2 - ℓ_∞ filtering. She has published around 10 papers in refereed international journals.



Derui Ding (Senior Member, IEEE) received the B.Sc. degree in industry engineering and the M.Sc. degree in detection technology and automation equipment from Anhui Polytechnic University, Wuhu, China, in 2004 and 2007, respectively, and the Ph.D. degree in control theory and control engineering from Donghua University, Shanghai, China, in 2014.

He is currently a Senior Research Fellow with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, VIC, Australia. From July 2007 to December 2014, he was a Teaching Assistant and then a Lecturer with the Department of Mathematics, Anhui Polytechnic University, Wuhu, China. From June 2012 to September 2012, he was a Research Assistant with the Department of Mechanical Engineering, the University of Hong Kong, Hong Kong. From March 2013 to March 2014, he was a Visiting Scholar with the Department of Information Systems and Computing, Brunel University London, Uxbridge, U.K. From June 2015 to August 2015, he was a Research Assistant with the Department of Mathematics, City University of Hong Kong, Hong Kong. He has published over 100 papers in refereed international journals. His research interests include nonlinear stochastic control and filtering, as well as multiagent systems, and sensor networks.

Dr. Ding received the 2021 Nobert Wiener Review Award from the IEEE/CAA *Journal of Automatica Sinica*, the 2020 Andrew P. Sage Best Transactions Paper Award from the IEEE Systems, Man, and Cybernetics Society, and the 2018 IET Premium Award. He is the Standing Director of the IEEE PES Intelligent Grid and Emerging Technologies Satellite Committee, China. He is serving as an Associate Editor for *Neurocomputing* and *IET Control Theory and Applications*, a member of Early Career Advisory Board for *IEEE/CAA Journal of Automatica Sinica*, and also served as a Guest Editor for several issues, including the *International Journal of Systems Science*, *International Journal of General Systems*, and *Kybernetika*.



Muwei Jian received the Ph.D. degree in computer vision from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, in 2014.

Currently, he is a Distinguished Professor and the Ph.D. Supervisor with the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China. His current research interests include human face recognition, image and video processing, machine learning and computer vision.

Prof. Jian was awarded the Royal Society—K. C. Wong International Fellowship under Newton International Fellowship (NIF). In particular, he was supported by the “Taishan Young Scholars Program” of Shandong Province. He holds 11 granted national patents and has published over 80 papers in refereed international leading journals/conferences such as *IEEE TRANSACTIONS ON CYBERNETICS*, *IEEE TRANSACTIONS ON MULTIMEDIA*, *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, *Pattern Recognition*, *Information Sciences*, *Signal Processing*, *ISCAS*, *ICME*, and *ICIP*. He was the recipient of “The Most Cited Articles from Journal of Visual Communication and Image Representation published since 2016, 2017, and 2018” and “High Impact Research Articles of MTAP.” He was actively involved in professional activities. He has been a member of the Program Committee and Special Session Chair of several international conferences, such as *IJCAI*, *AAAI*, *ACM MM*, *ICME*, *VCIP*, and *ICTAI*. Currently, he serves as an Associate Editor of *IET Computers & Digital Techniques* and the *Journal of Image and Graphics (JIG)*.