ITC 1/52

Information Technology and Control
Vol. 52 / No. 1 / 2023
pp. 185-198
DOI 10.5755/j01.itc.52.1.32278

Generative Adversarial Networks for Video Summarization Based on Key-frame Selection

Received 2022/09/13

Accepted after revision 2022/12/20

https://doi.org/10.5755/j01.itc.52.1.32278

HOW TO CITE: Hu, X., Hu, X., Li, J., You, K. (2023). Generative Adversarial Networks for Video Summarization Based on Key-frame Selection. *Information Technology and Control*, 52(1), 185-198. https://doi.org/10.5755/j01.itc.52.1.32278

Generative Adversarial Networks for Video Summarization Based on Key-frame Selection

Xiayun Hu

School of Software Engineering, Jinling Institute of Technology, No. 99 Hongjing street, Nanjing, Jiangsu, China; phone: +86 18168092293; e-mail: hxy@jit.edu.cn

Xiaobin Hu

Software R&D Center (Hefei), Postal Savings Bank of China, No. 7389 Huizhou street, Hefei, Anhui, China; phone:+86 18855147417; e-mail: huxb1108@163.com

Jingxian Li, Kun You

School of Software Engineering, Jinling Institute of Technology, No. 99 Hongjing street, Nanjing, Jiangsu, China; phones: +86 18168092013, +86 18168092623; e-mails: lijingxian@jit.edu.cn, youk_07@jit.edu.cn

Corresponding author: hxy@jit.edu.cn

Video summarization based on generative adversarial networks (GANs) has been shown to easily produce more realistic results. However, most summary videos are composed of multiple key components. If the selection of some video frames changes during the training process, the information carried by these frames may not be reasonably reflected in the identification results. In this paper, we propose a video summarization method based on selecting keyframes over GANs. The novelty of the proposed method is the discriminator not only identifies the completeness of the video, but also takes into account the value judgment of the candidate keyframes, thus enabling the influence of keyframes on the result value. Given GANs are mainly designed to generate continuous real values, it is generally challenging to generate discrete symbol sequences during the summarization process directly. However, if the generated sample is based on discrete symbols, the slight guidance change of the discrimination network may be meaningless. To better use the advantages of GANs, the study also adopts the video summarization optimization method of GANs under a collaborative reinforcement learning strategy. Experimental results show the proposed method gets a significant summarization effect and character compared with the existing cutting-edge methods.

KEYWORDS: Video summarization, generative adversarial networks, reinforcement learning.

1. Introduction

With the rapid development of digital media applications, a large number of videos have been filmed and have become an important information carrier and medium in our daily work, social and entertainment activities. However, for human beings, it usually takes a long time to watch the whole video to understand the video content [1, 3]. In addition, for some applications, such as video surveillance, the length of unprocessed monitored videos are usually long [17, 27]. This can make it difficult for administrators to efficiently retrieve and process critical video information because redundant information is mixed in [17, 27]. For this reason, there is an urgent need for a technology that can automatically obtain the core content of a video and does not require the playback of the entire video. Then, video summarization happened and has been a hot topic in recent years [1, 3, 7, 8, 10-17, 19, 21, 24-25, 30-36].

The primary purpose of video summarization is to retain the maximum amount of information from the original video while streamlining the length. Generally, video summaries can be performed in two manners: keyframe selection, where the system extracts a series of video frames [16], and keyshot selection, where the system extracts a series of video segments, each of which is a group of temporally continuous video frames spanning a short time interval [11]. In this paper, the video summarization task is regarded as the keyframe extraction problem, meaning that for a given video, due to the gradual evolution of many key shot-based studies into studies of video frames [11, 31, 32].

Recently, generative adversarial networks (GANs) [5] have achieved great success in many fields, such as image processing [6, 26], abnormal event detection [18, 33], and video prediction [22, 27]. There are two basic models in GANs: a generator for forging samples and a discriminator for distinguishing between generated and actual samples. Because of the excellent generation ability of GANs, video summarization methods based on GANs were first studied in [15], followed by many studies such as [37] and [30]. The technology roadmap of these studies is to design a summarizer (selector) for extracting keyframes from original video sequences, a generator for extracting video features and a discriminator for identifying authenticity with the principles of GANs.

However, there are still many critical issues to be addressed under this framework. First, GANs shine mainly in the image domain and are eclipsed in the discrete data domain. This is because the gradient changes generated by discrete samples through GANs are subtle, which will cause the gradient influence to disappear when the gradient is returned. Therefore, generative networks may not benefit from the learning activities. Secondly, as we have mentioned above, video sequences usually have a structure like frames-shot–frames. The video summarization job is to reassemble the representative or key video frames into the new video summary without losing the original video information as much as possible. Thus, GANs should measure the completeness of the video when dealing with frame sequences, i.e., making judgments about the integrity of the whole video. Nevertheless, in previous studies, the discriminator only judges the authenticity of the summary video and thus ignores the influence of changes in the keyframes on the summary results. These issues will inevitably limit the application of GANs to video summarization.

In light of the above issues, this study propose a video summarization approach based on keyframe selection in GANs. We further consider the value judgment of the selected frames in the framework of GANs to make a complete optimization strategy for the summarization results. The proposed method select the keyframes by predicting the importance score of the frames in the original video. Meanwhile, the superior generation effect of the GAN is used to realize the correlation between the summary video and the original video to “confuse” the discriminator. In addition, to enable GANs to further play a role in video summarization work, we implement the influence feedback of the selected frames through GANs and solve the problem of difficult gradient transfer with the help of the reward mechanism borrowed from the reinforcement strategy. Moreover, the feasibility of the proposed method is verified through qualitative and quantitative comparisons of experimental results.

The remainder of this paper is organized as follows. In Section 2, we introduce works related to video summarization methods based on GANs, as well as traditional and deep learning based approaches. In Section 3, we present the overall framework and the specific implementation process of the proposed method. In Section 4, we demonstrate the experimental and analytical results of the proposed method. Finally, the conclusion of the paper is presented.

2. Related Work

Video summarization is helpful for video management and browsing tasks [14]. And with the explosive growth of video resources, it has attracted a lot of attention. We will review the related works of this paper from three aspects: i) traditional methods, ii) methods based on deep learning, iii) methods based on GANs.

2.1. Traditional Methods

Research on video summarization has been going on for a long time, and a large number of approaches have been proposed in [7-8, 34]. For example, Gygli et al. [8] adopted a linear regressor to estimate the interestingness of the video and selected the highest-scoring keyframe to form the summary video. Similarly, in [7], the authors considered the video summary task as a sum of multiple target objects and optimized the sub-module functionality through these objects. Furini et al. [4] proposed a technique, called STIMO, for generating still and dynamic plot summaries in web scenes. In [10], Hong et al. proposed an event-based solution for aggregating the content of video search results by mining the video key shots, so that users could get the main content at a glance. However, these approaches are relatively early solutions that do not use the in-depth information on time sequence, nor the advantages of deep learning technique. Avila et al. [1] proposed an algorithm called VSUMM, which extracted the color feature from video frames based on the K-means clustering algorithm. To obtain a proper video summarization, they extracted the deep features from each clip of the original video and applied the clustering-based algorithm to these features. All these works are relatively early and traditional. Although these aforementioned works have given a great impetus to the development of video summarization research, there is still room for progress in video summarization work, especially with the outstanding advantages of deep learning frameworks.

2.2. Methods Based on Deep Learning

Deep learning has been extensively studied in various areas of multimedia research topics. Similarly, the research on video summarization has also entered the deep learning era under the sweeping wave of technology. For example, Zhang et al. [35] are the first group to adopt the recurrent neural network (RNN) to implement the deep model for video summarization. They considered the video summarization task as a structured prediction problem on sequential data and used a bidirectional RNN to model the variable range correlation in the video. In [32], Zhao et al. treated the video summarization task as a video structuring problem, i.e., training with video shot as the basic unit. In this work, a two-layer RNN structure was designed, where the upper layer performed shot detection and segmentation of the video, and the lower layer performed importance detection of several shots. Finally, the most critical shots were selected from these shots to be combined into the final video. Zhou et al. [36] used the long-short term memory (LSTM) to design a depth model that considered the diversity and variability of the summarization video. They calculated the representativeness reward and the diversity reward by the DR reward function. The representativeness reward calculated the distance between each frame and its nearest selected frame. The diversity reward measured the degree of dissimilarity between the selected frames. Eventually, the summarizer was optimized by the evaluated DR reward based on reinforcement learning. Fei et al. [2] proposed an improved triplet deep ranking model. Based on an efficient entropy-based video segmentation method, the original video is sliced into several segments. The summarization result is generated by estimating the visual interest score of each segment through the use of a well-trained ranking network.

2.3. Methods Based on GANs

As one of the significant breakthrough results in the development of deep learning [3,17,28], GANs have led to greater progress in video summarization. For example, Zhu et al. [37] used the cycle-consistent adversarial network to transform images from the source domain to the target domain without paired samples. Given the video summarization work is to generate a short form of summarization video based on the original video, the primary motivation for using GANs is to consider such work as a generation process in GANs. A representative approach is a SUM-GAN model devised by Mahasseni et al. [15], where the video summarization was formulated as a sparse subset of video frames selected by a selector. In their work, the authors designed a deep summarization network for learning to minimize the distribution distance between the training video and the summarization video to the greatest extent. The model consists of two LSTMs: an autoencoder LSTM as the summarizer and the generator, and a normal LSTM as the discriminator. Thus, the summarizer is trained to acquire the ability to confuse the discriminator. In addition, Yuan et al. [30] proposed the Cycle-SUM model, which used two VAE-based generators and two discriminators to evaluate the cycle consistent loss to achieve an effective information preservation. The forward generator and the discriminator are responsible for reconstructing the original video according to the summarization result. Nevertheless the backward generator and the discriminator perform the reconstruction from the original video to the summarization result. In the analysis of these studies based on the framework of GANs, we can find that although these methods implement the frame-level video summarization, they still optimize the discrimination by the overall judgment of the video in the adversarial structure without paying much attention to the impact of the selected frames on the results. Therefore, we further implement a novel frame-level video summarization model built on the GANs framework with consideration of key-frame influence.

3. Video Summarization Model Based on the Key-frame Selection in Generative Adversarial Networks

In this section, we first introduce the overall framework of the proposed method in Section 2.1. Then, we present the implementation process and the refined training strategy of this method in Section 2.2 and Section 2.3, respectively.

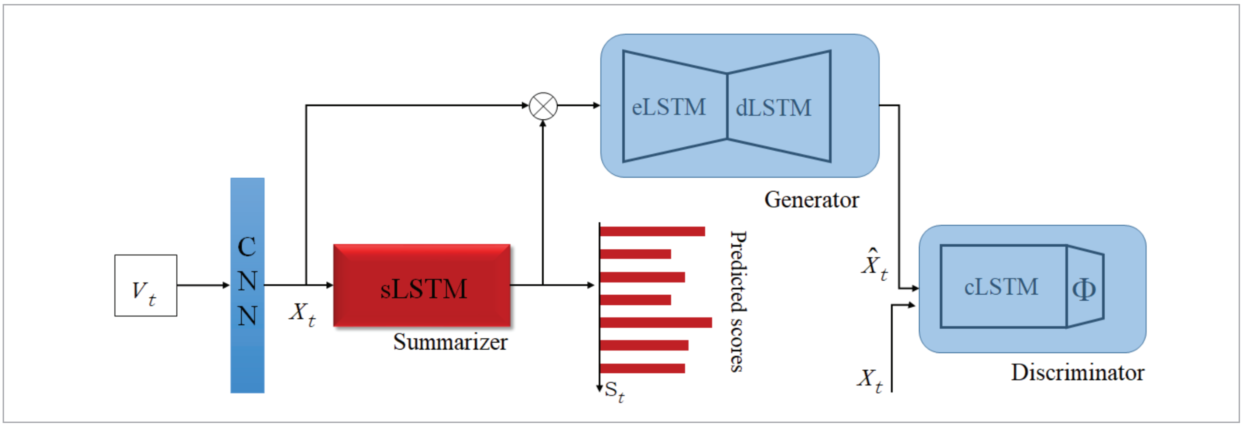
3.1. Overall Framework

We propose a video summarization method based on the key-frame selection in GANs which implements the frame-level selection and the model optimization process through the reinforcement learning strategy. The overall framework of the proposed algorithm is shown in Figure 1, where the framework mainly consists of a summarizer sLSTM, a generator, and a discriminator. The summarizer is to select keyframes in the original video, the generator is to perform the feature reconstruction and extraction of the original or summarized video, and the discriminator's task is to implement the judgment of the features of the input video.

Without loss of generality, we assume we are given a video summarization task for a veideo sequence $X=(x_1,x_2,\cdots,x_T)$, where T represents the length of the video (i.e., number of frames). Accordingly, the depth feature can be extracted from these video frames with the CNN model (we denote it as $V=(v_1,v_2,\cdots,v_T)$). For the implementation of the summarizer, we adopt the

Figure 1

Video summarization model based on the key-frame selection in generative adversarial network



classical LSTM network. We assume the summarizer, denoted as sLSTM, has selected a subset of keyframes from the original video. The output of the summarizer is the relevance score or importance score of each input frame, denoted as $S=(s_1,s_2,\cdots,s_T)$. These scores will be an important basis for the selection of keyframes. Note that the importance score reflects how meaningful the frame is in the original video. If the importance scores are normalized to {0, 1}, then they will become the key-frame selection indicators, i.e., 0 means that the frame will not be selected as an important frame, while 1 means that it will be selected definitely.

In view of the ability of the auto-encoder to reconstruct and extract features [13], the generator in the proposed model is realized with the structure of the auto-encoder. The encoder and decoder of the generator are respectively denoted as eLSTM and dLSTM. The autoencoder is a directed graph model that can define the posterior distribution of the observed data when no observed latent variable is given. Let $e\sim p_e(e)$ be the priori value of the unobserved latent variable and x be the observed data. Then, e can be considered as the encoded information on x and $q(e\mid x)$ can be defined as the observed probability distribution of the encoded information e on the given input x. In common practice, the distribution of $p_e(e)$ can be set to the standard normal distribution, and similarly, we adopt $p(x\mid e)$ to denote the conditional generating distribution of *x*. Therefore, the learning process can be accomplished by minimizing the negative log-likelihood of the data distribution, i.e.,

$$-\log\frac{p(x\mid e)\,p(e)}{q(e\mid x)}=-\log\left(p(x\mid e)\right)+D_{KL}(q(e\mid x)\|\,p(e)). \tag{1}$$

After inputting the visual feature matrix of the frame sequence *V*, and the summarizer has successfully predicts its importance score, the importance score is binarized and will be used as the behavior indicator $A=(a_1,a_2,\cdots,a_T)$. Based on these indicators, we can reconstruct the spatio-temporal features of the original and summary videos, i.e.,

$$X^{gt}=a^{gt}\cdot X, \tag{2-1}$$

$$X^{p}=a^{p}\cdot X, \tag{2-2}$$

where a^{gt} and a^{p} are the real indicator and the predictive indicator, respectively. Accordingly X^{gt} and X^{p} represent the real video summary feature and the predictive summary video feature, respectively. To produce more realistic results, the discriminator is designed to evaluate whether the generated summaries are equivalent to the original video in terms of content. Existing methods based on GANs [30, 37] distinguish the summaries at the video level. However, they ignore the subtle variations in the summaries, which may affect the final summaries. Therefore, we focus on the impact of the selected frames. For a subset of selected frames, our discriminator aims to rate these frames. We use the reconstructed video features as the input of the discriminator. The importance judgment of the frame in the summary video can be obtained through the discriminator, i.e., $C=\{c_1,c_2,\cdots,c_M\}$, where *M* represents the number of keyframes in the original video. Obviously, the value of *M* is completely random and unequal for different videos as well as for the predictive summary video and the original video.

3.2. Implementation Process

A GAN is a neural network consisting of two competing sub-networks: i) a generator that generates unknown distribution data; ii) a discriminator for distinguishing tasks. If we use mathematical language to describe the whole game process, i.e., suppose that our generator is *G*(*z*), where *z* is a random variable, then the generator *G* is to transform this random variable into a specified data type. It is assumed that the output of the generator is a picture. For any input, the task of the discriminator is to output a real number in the interval between 0 and 1, which is used to discriminate the authenticity of the input image. And the larger the real number output is, the higher the confidence level is, and vice versa. We let P_R and P_G denote the distributions of the real sample and the generated sample, respectively, then the objective function of the discriminator is

$$\max{}_D\, E_{x\sim P_R}\left[\log D(x)\right]+E_{x\sim P_G}[\log(1-D(x))]. \tag{3}$$

The main task of the discriminator is to distinguish the true from the false, and the generator's goal is to make the discriminator unable to distinguish the true sample from the generated sample correctly. Then the overall optimization objective function expression of the generative adversarial network is

$$\min_G \max_D E_{x\sim P_R}[\log D(x)] + E_{x\sim P_G}[\log(1-D(x))], \quad (4)$$

where the $\min_G \max_D$ in the above equation is exactly the game process mentioned earlier in this paper. The above equation is simplified to $\min_G \max_D V(G,D)$, where $V(G,D)$ represents the degree of difference between the real data and the generated data. First, for $\max_D V(G,D)$, the generator model is fixed during the training process, that is to say, the parameters of the generator will not be updated next, which aims to maximize the discriminator's ability to identify whether the data is from the real or generated samples. After that, the latter part is considered as a whole, i.e., $\min_G L$, where the parameters of the discriminator are fixed. Therefore, the generator can learn to confuse the behavior of the discriminator under this condition. After the above iterative process, the game between the generator and the discriminator is played continuously, finally reaching an equilibrium state. So we can get a stable generator.

Therefore, considering the characteristics of GANs, the purpose is to train the generator to the extent that it can "cheat" the discriminator. The goal of the discriminator is to effectively identify the original video feature x_t as "original video" and the reconstructed summary video feature $\hat{x}_t$ as "summary video". Based on the above discussion, we introduce the unique objective function of generative adversarial networks, i.e.,

$$\mathcal{L}_{GAN} = \log\left(cLSTM\left(x\right)\right) + \log\left(1 - cLSTM\left(\hat{x}\right)\right). \quad (5)$$

Furthermore, since the construction of the generator is derived from the autoencoder, it is necessary to include the priori loss $\mathcal{L}_{prior}$ and the reconstruction loss $\mathcal{L}_{recon}$ during the summarizer and generator training processes. Thus we can get $\mathcal{L}_{prior}$ and $\mathcal{L}_{recon}$ separately from

$$\mathcal{L}_{prior} = D_{KL}(q(e \mid x) \,||\, p(e)) \quad (6)$$

and

$$\mathcal{L}_{recon} = \mathbb{E}\left(-\log(p_e(x \mid e))\right). \quad (7)$$

3.3. Training Strategy

In general, the primary function of the discriminator in most of the current studies is to distinguish the "authenticity" of the video samples, and the process of identification is carried out as a whole unit. Thus the purpose of a discriminator is to distinguish between the generated sample and the real sample, and make the gap between the original and the summary video generated by the GAN as small as possible. However, classical GANs are deficient in dealing with serialized data. Firstly, GANs are mainly designed to generate continuous real values, it is not easy to directly generate discrete symbol sequences (such as video frames). The reason for this phenomenon is that in GANs the generator first starts with random sampling, and then the deterministic transformation is performed by the model parameters. If the generated samples are based on the discrete symbols, it will be meaningless to identify a slight guidance change from the discriminator network, because there may not be a corresponding symbol in the limited dictionary space for such slight changes. Secondly, the evaluation of classical GANs can only focus on the complete generated sequences. In contrast, for the generated incomplete sequences, it is important to unify the quality at this moment and the score of the whole sequence in the future. Therefore, there are still some issues to be solved when classical GAN is directly applied to video summarizations.

In order to solve these issues, we further improved the working process of the discriminator so that it could react to the quality of the local sequences. Meanwhile, to effectively pass the slight feedback changes from the discriminator to the generator, we give such changes backward by means of reinforcement learning. Specifically, we treat the summarizer and the generator as an agent $\pi_\theta\left(a \mid x\right)$, and let this agent continuously optimize the parameters of the generator by maximizing the expected reward, i.e.,

$$\mathcal{T}\left(\theta\right) = \mathbb{E}_{P_\theta(a_1:T)}[\log_{\pi_\theta}(a_\theta \mid \hat{x}_t)Q\left(\hat{x}_t, a_t\right). \quad (8)$$

In Equation (8), without loss of generality for the t th $(1 \le t \le T)$ frame, after binarization the importance score s_t of this frame derived by the summarizer, the behavior indicator a_t will be generated. Whether this frame is selected or not will be based on this indicator (1 means to select the frame, and 0 means to discard on the contrary). $Q\left(\hat{x}_t, a_t\right)$ is the behavior evaluation function to evaluate the feedback value generated under the selection of a_t. At the same time, to evaluate the feedback values obtained from the summary precisely, we repeat the evaluation process N rounds to get the mean value, i.e.,

$$\mathrm{Q}\left(\hat{x}_t, a_t\right) = \frac{1}{N}\sum_{t=1}^{N}\mathrm{cLSTM}\left(\hat{x}_t, a_t\right). \tag{9}$$

After the discriminator gets updated, we can train the summarizer and generator again and update they parameters. Thus, according to the above discussion, we can estimate the gradient change of the objective function $J\left(\theta\right)$ as

$$\nabla_\theta J = \sum_{t=1}^{T}\mathbb{E}_{\hat{x}_{1:T}\sim G_\theta}\left[\nabla_\theta \log_{\pi_\theta}(a_t \mid \hat{x}_t)\bullet Q\left(\hat{x}_t, a_t\right)\right]. \tag{10}$$

Equation (10) selects the frames considered "critical". The frames that are not recognized will be discarded. Given this selection operation is a random process in the training procedure, the selected results are almost impossible to be consistent with the expected goal. Therefore, there are still some important video keyframes in the discarded "unimportant" video frames. It is necessary to reuse the discarded video frames. The solution is simple, just taking the opposite selection of a_t, and we denote the objective function after this solution as $J_u\left(\theta\right)$ in the following.

According to Figure 1, it can be seen that the proposed network model contains three parts, and the training parameters can be divided into three parts correspondingly: 1) learning and updating the network parameter θ_s of the summarizer sLSM; 2) learning and updating the network parameter θ of the generator; 3) learning and updating the network parameter θ_D of the discriminator. In the training process of the summarizer, optimize its output as much as possible, so that its distribution matches the requirements for the video summarization application. An obvious requirement for the video summarization task is to reduce the length of the original video significantly, so we need to constrain the final output video length to avoid large-scale video frames being selected as keyframes. This effective constraint also effectively amplifies the importance between different video frames, which brings great convenience to the selection of the final video key frames. Therefore, for this characteristic, we design an objective function for the summarizer, i.e.,

$$\mathcal{L}_{sparsity} = \left\|\frac{1}{T}\sum_{t=1}^{T}s_t - \sigma\right\|, \tag{11}$$

where T denotes the number of frames of the whole video, and σ represents the hyper-parameter selected in the experiment whose main function is to achieve a constraint on the number of keyframes selected, i.e., constrain the weight of keyframes in the whole video frame sequence.

Therefore, combined with the above loss function, we can realize the training of the proposed model, i.e.,

$$\mathcal{L} = \mathcal{L}_{sparsity} + \lambda_1\left(\mathcal{L}_{prior} + \mathcal{L}_{recon}\right) + \lambda_2\mathcal{L}_{GAN} + \lambda_3\left(J + \mathcal{J}_u\right), \tag{12}$$

where λ_1, λ_2 and λ_3 are respectively the selected hyperparameters for the balance of the training of the proposed model.

4. Experimental Results and Analysis

In this section, we first present the experimental setup of the proposed method, including the available datasets and the rules for usage as well as the evaluation criteria. Then, we test the properties of the proposed method in the manner of an ablation experiments, and the proposed method will also be compared with the current cutting-edge methods.

4.1. Experiment Setup

Datasets and evaluation criteria. To verify the effectiveness of the proposed method, we conduct extensive experiments on publicly available standard video summary datasets, including the TVSum [8] and SumMe [24] datasets. In addition, the experiments are extended with two additional datasets (OVP and YouTube), which will extend the experiments to enrich the training of the model better. Details of these datasets are as follows.

TVSum dataset [8]: This dataset collects 50 videos from YouTube, and the categories of these videos are selected from the TRECVid Multimedia Event Detection (MED) [23] task. The entire dataset has ten categories, covering transportation, animals, sports, food, and so on. We select five videos for each category in the experiments. In addition, the duration of the videos are different, and all videos are between 2 and 5 minutes long, and each video contains at least one shot. The videos in the dataset also provide frame-level importance scores scored by 20 volunteers.

SumMe dataset [24]: this dataset collects a total of 25 videos on topics ranging from holidays, events, sports, etc. These videos are selected either raw or minimally edited, and they also have a high number of views compared to the edited videos. The duration of each video in the dataset ranges from 1 to 6 minutes, and 15 to 18 volunteers have provided frame-level ratings for these videos.

OVP dataset and YouTube dataset: these two datasets are not adopted for the metrics detection in video summary task, but are used for evaluation enhancement. OVP contains 50 videos which are usually some news and documentary records, etc.; the YouTube dataset also contains 50 videos, which are between 1 and 10 minutes long. Although these videos do not provide significant frame-level scores, they clearly indicate which frames are the keyframes.

To make an effective evaluation of the proposed model, the dataset is divided into non-enhancement and enhancement types as shown in Table 1. The non-enhancement type only uses two traditional datasets for the training of the model. 80% of the videos are selected as the training set and the remaining 20% as the test set, respectively. The enhancement type adds the other videos to the non-enhanced training set, for example, TVsum, OVP and YouTube are added to 80% of SumMe. In the partitioning of the training and test sets, we randomly partition five times and take the mean value as the final evaluation score.

Experiment environment. The experiment machine adopted in this work is configured with a GeForce GTX Titan X GPU; the operating system version is Ubuntu 14.04. And the deep learning framework Pytorch [20] is used to implement the proposed model. First, we preprocess the dataset to extract video frames at a rate of 2 FPS, and extract the depth features of each frame by the deep convolutional network ResNet152. Then we adopt these features as the input of our model. For the training of the model, we use the Adam optimization algorithm [12] and the backward gradient propagation algorithm to update the network parameters iteratively. In addition, it has been suggested that generators and discriminators have better effects with different learning rates [9]. Thus we set the learning rate l_θ of the generator to 0.001 and the learning rate l_ϕ of the discriminator to 0.002. The hyper-parameters β_1 and β_2 of the optimizer are set to 0.9 and 0.999, respectively. Moreover, the training process of the proposed model has been performed for 100 iterations during the experiment.

Evaluation criterion. In order to take a valid comparison, F-score is chosen as an objective evaluation means for the proposed model. For all datasets, the videos will generate serialized importance scores through the summarizer. In the evaluation, we use the KTS [19] algorithm to convert these frame-level scores into shot-level scores. And based on the shot-level scores, important shots are selected to form the final summary video. Assuming that ***A*** is the generated summary video and ***B*** is the real summary annotated by the user, the F-score can be calculated by

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \tag{13-1}$$

where

$$P = \frac{\textit{Duration of overlap between } \boldsymbol{A} \textit{ and } \boldsymbol{B}}{\textit{Duration of } \boldsymbol{A}} \tag{13-2}$$

and

$$R = \frac{\textit{Duration of overlap between } \boldsymbol{A} \textit{ and } \boldsymbol{B}}{\textit{Duration of } \boldsymbol{B}}. \tag{13-3}$$

4.2. Experiment Results and Analysis

Ablation analysis. In this paper, we propose the video summarization method based on the selection of

Table 1

Processing of the datasets in the experiment

	Training dataset	Test dataset
Non-enhancement	80%SumMe	20%SumMe
	80%TVSum	20%TVSum
Enhancement	80%SumMe + TVSum + OVP + YouTube	20%SumMe
	80%TVSum+ SumMe+ OVP + YouTube	20%TVSum

key video frames. The method leverages the excellent performance of GANs. To alleviate the difficulty of GANs in optimizing discrete data models, there is a gradient change strategy applied to reinforce strategy collaboration. In addition, to better and intuitively analyze the improvement of the model brought by the proposed method, the loss function in the model optimization process should be analyzed in detail.

Table 2 gives the performances of the proposed method in different cases with the datasets SumMe and TVSum. It can be seen that vsGAN5 performs the best, with 42.1% in SumMe and 58.3% in TVSum, which demonstrates the effectiveness of the proposed frame-level video summarization model learned by reinforcement strategy based on GANs. By comparing the different models, we can see the different effects of various loss functions. Comparing vsGAN1 with VSGAN2, vsGAN2 further improves the effectiveness of the model on the original basis. This is because that with the addition of the sparse constraint on the summary length, the network is forced to reconstruct more accurate information about the original video content from its subset. Then, focus on the most semantically representative video frames while less attention is paid to semantically irrelevant video frames,making it easier to distinguish between key and non-key frames. Subsequently, the priori loss and the reconstruction loss of the autoencoder are added in the experiments, which brings another small improvement effect on the proposed model. Finally, after the addition of the frame-level summary loss proposed in this paper clearly shows that the constraint has an objective improvement on the model. This phenomenon demonstrates the limited effectiveness of video summarization by GANs alone. But with the idea of feedback of the frame-level importance in the summarization results, the boundedness is removed in the proposed GANs. In addition, with the addition of J_u, the model further gains a small improvement, which also shows the effectiveness of this approach from the other side.

Table 2

The performance of the proposed model in different situations

Item	vsGAN$_1$	vsGAN$_2$	vsGAN$_3$	vsGAN$_4$	vsGAN$_5$
$\mathcal{L}_{GAN}$	√	√	√	√	√
$\mathcal{L}_{sparsity}$		√	√	√	√
$\mathcal{L}_{prior}+\mathcal{L}_{recon}$			√	√	√
J				√	√
J_u					√
SumMe	39.6	40.6	41.0	41.9	**42.1**
TVSum	54.8	55.7	55.9	56.8	**58.3**

To get an intuitive view of the actual effect of the proposed model, the comparisons between the importance score prediction of two video frame sequences through the proposed model and the real importance scores are shown in Figure 2. It can be seen that the

Figure 2

Prediction of the importance score of two video frame sequences in TVSum (using the proposed model)

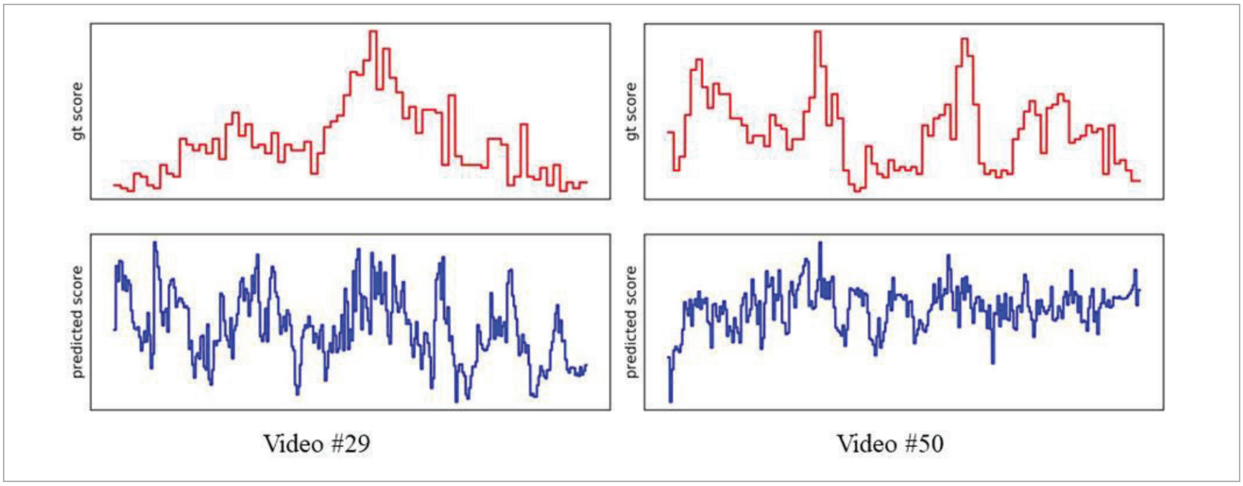
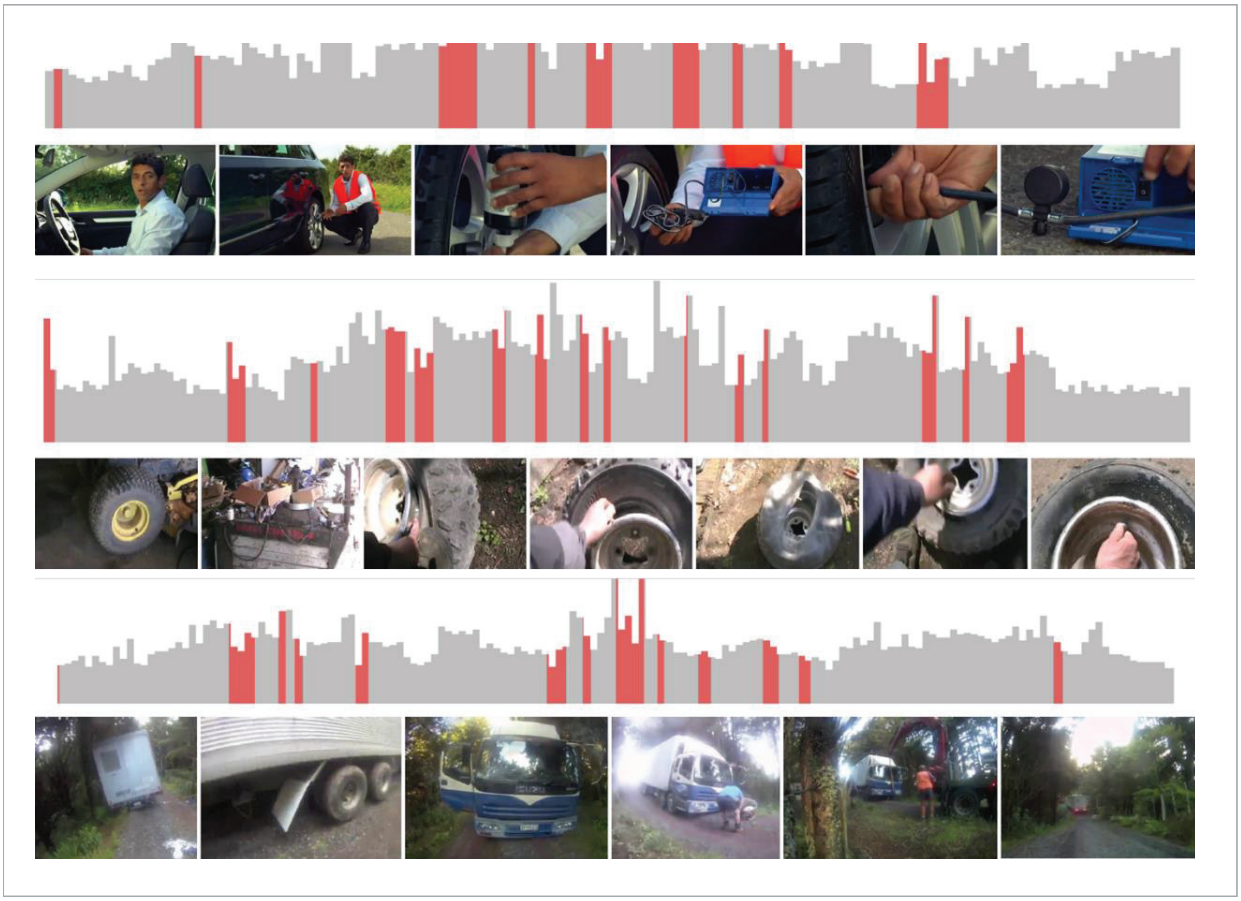


Figure 3

Video summaries of Video # 6, #26, #37 in TVSum taken by the proposed method



proposed model has a good simulation performance in predicting the importance curves of video frames and captures the critical information in the video more accurately.

Furthermore, to better illustrate the temporal selection pattern of different variations of our approach, we demonstrate the selected frames on an example video in Figure 3. It can be seen that our method can also capture the key segments in the video pretty well.

Comparison with cutting-edge algorithms. In addition to the comparison experiments in its own different situations, the proposed method has been compared with some current superior algorithms, so the strengths and weaknesses of our proposed method are demonstrated. The comparison results between the proposed algorithm and the traditional video summarization algorithms on SumMe and TVSum datasets are shown in Table 3.

Table 3

Comparison results (%) with existing video summarization algorithms

Methods/Algorithms	SumMe	TVSum
STIMO [4]	32.2	34.0
K-medoids [21]	33.4	28.8
Interestings [8]	39.4	—
Submodularity [7]	39.7	—
Summary transfer [34]	40.9	—
vsGAN	**42.1**	**58.3**

As can be seen from Table 3, the video summarization network model proposed in this paper performs significantly better than existing algorithms on both SumMe and TVSum datasets. This is because our model considers the features extracted by the deep convolution network in addition to the currently popular neural network. These features contain more information than the previous shallow features. These features also allow the proposed model to obtain a large amount of useful information, which shows the inevitable trend of using high-dimensional features for video summarization tasks.

To further demonstrate the performance of the proposed method, we compared it with the neural network-based video summarization algorithms proposed in recent years, including the vsLSTM [35] method, the dppLSTM [35] method, the DSN [36] method, the SUM-GAN [15] method, and the Cycle-SUM [30] method. The data comparison results with these methods are given in Table 4. Comparing to Table 3, we can find the video summarization models with the neural network outperform the other methods, which supports the effectiveness of applying the neural network model to the summarization task. Compared with these excellent algorithms, the proposed method can also achieve a good performance. Specifically, in the SumMe data, our method outperforms the other methods except for DSN. In the table, SUM-GAN and Cycle-SUM also adopt GANs as the main framework. In fact, SUM-GAN is the first method based on GANs. In contrast, Cycle-SUM uses the means of cyclic identification to achieve the role of confusion identification. It can be seen from the table that our method has been slightly improved compared with their research strategies. This result shows that the refinement analysis of the sequence and the back propagation through the reward mechanism can improve the adversarial model in dealing with the sequence prediction problem of video summarization.

Table 4

Comparison result (%) with video summarization algorithm based on deep learning

Methods/Algorithms	SumMe	TVSum
vsLSTM[35]	37.6	54.7
dpp-LSTM[35]	38.6	54.7
DSN[36]	**42.1**	58.1
SUM-GAN [15]	41.7	56.3
Cycle-SUM[30]	41.9	57.6
vsGAN	**42.1**	**58.3**

In addition, Zhang et al. [35] added OVP and YouTube datasets to the original dataset to enhance the results achieved by the model on SumMe and TVSum. The comparison between the proposed method and other existing algorithms in the case of the enhancement dataset is presented in Table 5. In SumMe, our method is better than the other methods. Although our method does not outperform the other methods in TVSum, it still achieves an improvement on the original data. To better demonstrate the selection effect of the proposed method, the comparison between the proposed method and two representative methods (vsLSTM [35] and SUM-GAN [15]) is shown in Figure 4. In the example, the gray background represents the user's rating of the video frames, while the red annotations represent the video clips selected by the different methods. Compared to the two methods, the proposed method selects shorter but more critical and focused shots. Compared with the results of vsLSTM and SUM-GAN, vsGAN selects more shots related to the topic or details that people care about. For example, the selected frames or shots mainly show how the doctor cleans the dog's ear. The proposed method is more capturing for the core content of the video, so that the final summary result is more representative. Thus, the effectiveness of the adopted strategy is validated.

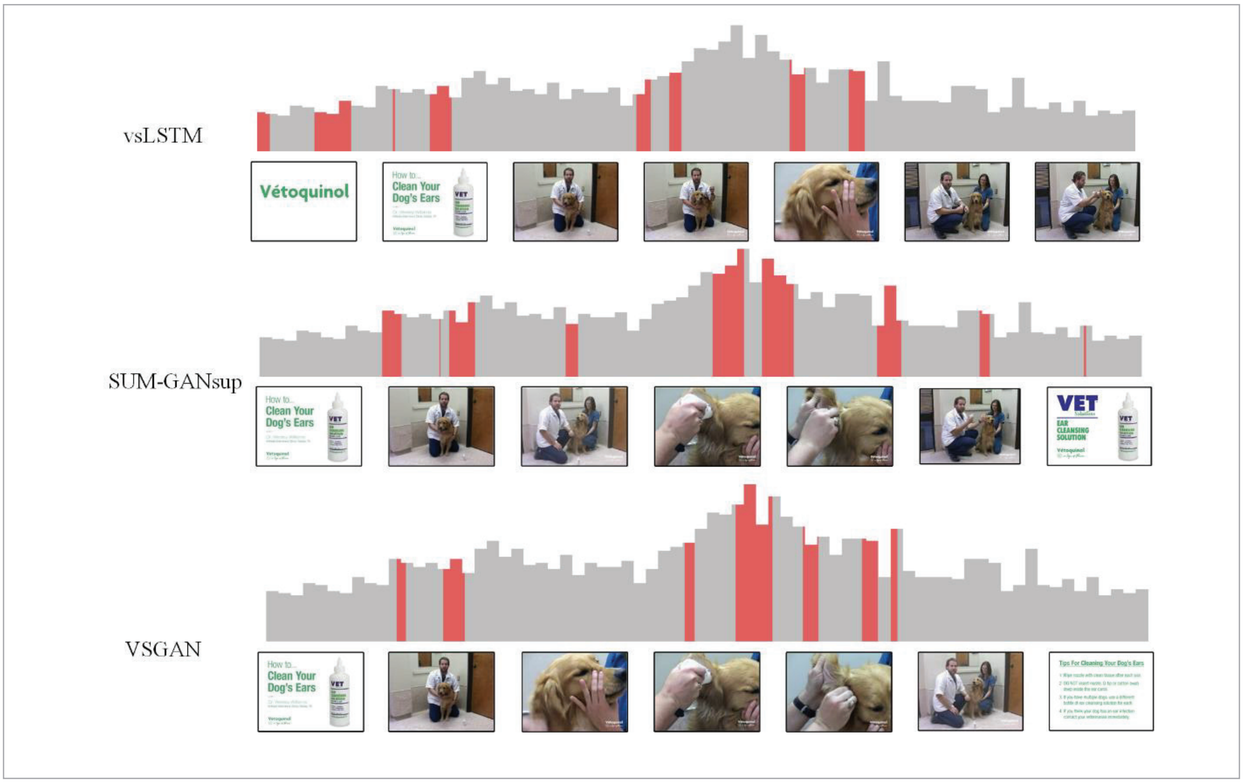
Table 5

Comparative experimental results (%) under the enhancement datasets

Methods/Algorithms	SumMe	TVSum
vsLSTM [35]	41.6	57.9
dpp-LSTM [35]	42.9	54.7
HAS-RNN [32]	44.1	59.8
SUM-GAN [15]	43.6	61.2
DSN [36]	43.9	59.8
vsGAN	**44.3**	**59.1**

Figure 4

Comparison of the effect of the proposed method with the other two methods (taking Video #16 in TVSum as an example)



5. Conclusion

A video summarization model based on the key-frame selection in GANs is proposed in this paper. Previous GANs-based summarization models aim at video-level identification and do not implement the analysis of the role of selected frames in the summarization results. Our method implements the integration of GANs in the video summarization task and further adds feedback on the importance of selected frames in the results. To better implement the feedback of keyframes in the results and avoid the problems of GANs in discrete generation tasks, we adopt a reinforcement policy reward mechanism to pass the gradient changes back to the generator, thus improving the optimization process of the summarizer and the generator. Extensive experiments on publicly available datasets have demonstrated the effectiveness of the proposed video summarization method.

Acknowledgement

This work is supported by the funding of University Quality Education and Digital Curriculum Construction of Jiangsu Universities (No.2020JDKT136).

References

1. Avila, S., Lopes, A., Luz, A. D., Araújo, A. VSUMM: A Mechanism Designed to Produce Stat-ic Video Summaries and A Novel Evaluation Method. Pattern Recognition Letters, 2011, 32(1), 56-68. https://doi.org/10.1016/j.patrec.2010.08.004
2. Fei, M., Jiang, W., Mao, W. Learning User Interest with Improved Triplet Deep Ranking and Web-Image Priors for Topic-Related Video Summariza-tion. Expert Systems with Applications, 2021, 166(1), 114036. https://doi.org/10.1016/j.eswa.2020.114036

3. Fei, M., Jiang, W., Mao, W. A Novel Compact Yet Rich Key Frame Creation Method for Compressed Video Summarization. Multimedia Tools and Ap-plications, 2018, 77(10), 11957-11977. https://doi.org/10.1007/s11042-017-4843-2

4. Furini, M., Geraci, F., Montangero, M., Pellegrini, M. Stimo: Still and Moving Video Storyboard for the Web Scenario. Multimedia Tools and Applica-tions, 2010, 46(1), 47-69. https://doi.org/10.1007/s11042-009-0307-7

5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Coirville, A.,Bengio, Y. Generative Adversarial Nets. Pro-ceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, November 18-22, 2014, 2672-2680. https://dl.acm.org/doi/10.5555/2969033.2969125

6. Gu, L., Zhang, L., Wang, Z. A One-Shot Texture-Perceiving Generative Adversarial Network for Unsupervised Surface Inspection. 2021 IEEE In-ternational Conference on Image Processing, (ICIP 2021), Anchorage, USA, September 19-22, 2021, 1519-1523. https://doi.org/10.1109/ICIP42928.2021.9506202

7. Gygli, M., Grabner, H., Gool, L. V. Video Summa-rization by Learning Submodular Mixtures of Ob-jectives. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, June 07-12, 2015, 3090-3098. https://doi.org/10.1109/CVPR.2015.7298928

8. Gygli, M., Grabner, H., Riemenschneider, H., Gool, L. V. Creating Summaries from User Videos. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision-ECCV 2014. Lecture Notes in Computer Science, 8695, Springer, Cham, 2014, 505-520. https://doi.org/10.1007/978-3-319-10584-0_33

9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. Gans Trained by a Two Timescale Update Rule Converge to a Local Nash Equilibrium, Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, December 4-9, 2017, 6629-6640. https://dl.acm.org/doi/10.5555/3295222.3295408

10. Hong, R., Tang, J., Tang, J., Tan, H. K., Ngo, C. W., Yan, S., Chua, T. S. Beyond Search: Event Driven Summarization for Web Videos. ACM Transactions on Multimedia Computing, Commu-nications, and Applications (TOMM), 2011, 7(4), 1-18. https://doi.org/10.1145/2043612.2043613

11. Ji, Z., Xiong, K., Pang, Y., Li, X. Video Summariza-tion with Attention-based Encoder-decoder Net-works. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(6), 1709-1717. https://doi.org/10.1109/TCSVT.2019.2904996

12. Kingma, D., Ba, J. Adam: A Method for Stochastic Optimization, Proceedings of the 3rd International Conference on Learning Representations, (ICLR 2015), San Diego, USA, May 7-9, 2015, 1-13. https://doi.org/10.48550/arXiv.1412.6980

13. Kingma, D. P., Welling, M. Auto-encoding Varia-tional Bayes. Proceedings of the 2nd International Conference on Learning Representations, (ICLR 2014), Banff, AB, Canada, April 14-16, 2014, 1-14. https://doi.org/10.48550/arXiv.1312.6114

14. Ma, Y. F., Lu, L., Zhang, H. J., Li, M. A User Attention Model for Video Summarization. Proceedings of the 10th ACM International Conference on Multimedia, Juan-Les-Pins, France, December 1-6, 2002, 533-542. https://doi.org/10.1145/641007.641116

15. Mahasseni B., Lam M., Todorovic S. Unsupervised Video Summarization with Adversarial LSTM Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hono-lulu, HI, USA, July 21-26, 2017, 2982-2991. https://doi.org/10.1109/CVPR.2017.318

16. Meng, J., Wang, H., Yuan, J., Tan, Y. From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection. Pro-ceedings of the IEEE Conference on Computer Vi-sion and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 27-30, 2016, 1039-1048. https://doi.org/10.1109/CVPR.2016.118

17. Muhammad, K., Hussain, T., Tanveer, M., San-nino, Giovanna., Albuquerque, V. Cost-effective Video Summarization Using Deep CNN with Hi-erarchical Weighted Fusion for IoT Surveillance Networks. IEEE Internet of Things Journal, 2019, 7(5), 4455-4463. https://doi.org/10.1109/JIOT.2019.2950469

18. Nie, L., Wu, Y., Wang, X., Guo, L., Wang, G., Gao, X., Li, S. Intrusion Detection for Secure Social Internet of Things Based on Collaborative Edge Computing: A Generative Adversarial Network-Based Approach. IEEE Transactions on Computa-tional Social Systems, 2022, 9(1), 134-145. https://doi.org/10.1109/TCSS.2021.3063538

19. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C. Category-specific Video Summarization. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), European Conference on Computer Vision, Lecture Notes in Computer Science, 8694. Spring-er, Cham, 2014, 540-555. https://doi.org/10.1007/978-3-319-10599-4_35

20. Pytorch Documentation. https://pytorch.org/doc s/stable/index.html. Accessed on September 16, 2022.

21. Rabbouch, H., Saâdaoui, F., Mraihi, R. Unsuper-vised Video Summarization Using Cluster Analy-sis for Automatic Vehicles Counting and Recog-nizing, Neurocomputing, 2017, 260(1), 157-173. https://doi.org/10.1016/j.neucom.2017.04.026

22. Shi, J., Zhu, Q., Wu, J. Unsupervised Transfer Learning For Video Prediction Based on Genera-tive Adversarial Network. Proceedings of the 27th International Conference on Mechatronics and Machine Vision in Practice, (M2VIP 2021), Shang-hai, China, November 26-28, 2021, 115-120. https://doi.org/10.1109/M2VIP49856.2021.9665045

23. Smeaton, A. F., Over, P., Kraaij, W. Evaluation Campaigns and TRECVid. Proceedings of the 8th ACM International Workshop on Multimedia In-formation Retrieval. Santa Barbara, California, USA, October 26-27, 2006, 321-330. https://doi.org/10.1145/1178677.1178722

24. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A. Tvsum: Summarizing Web Videos Using Titles. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, June 7-12, 2015, 5179-5187. https://doi.org/10.1109/ CVPR.2015.7299154

25. Yan, X., Cao, J., Sun, L., Zhou, J., Wang, S., Song, A. Accurate Analytical-Based Multi-Hop Localiza-tion with Low Energy Consumption for Irregular Networks. IEEE Transactions on Vehicular Tech-nology, 2020, 69(2), 2021-2033. https://doi.org/10.1109/TVT.2019.2957390

26. Yan, Y., Liu, C., Chen, C., Sun, X., Jin, L., Peng, X., Zhou, X., Fine-Grained Attention and Feature-sharing Generative Adversarial Networks for Sin-gle Image Super-resolution, IEEE Transactions on Multimedia, 2022, 24(1), 1473-1487. https://doi.org/10.1109/TMM.2021.3065731

27. Yang, K., Liu, D., Chen, Z., Wu, F., Li, W. Spatio-temporal Generative Adversarial Network-Based Dynamic Texture Synthesis for Surveillance Video Coding. IEEE Transactions on Circuits and Sys-tems for Video Technology. 2022, 32(1), 359-373. https://doi.org/10.1109/TCSVT.2021.3061153

28. Yoon, U. N., Hong, M. D., Jo, G. S.. Interp-SUM: Unsupervised Video Summarization with Piece-wise Linear Interpolation. Sensors, 2021, 21(13), 4562. https://doi.org/10.3390/s21134562

29. Yu, L., Zhang, W., Wang, J., Yong, Y. Seqgan: Se-quence Generative Adversarial Nets with Policy Gradient. Proceedings of the 31st AAAI Confer-ence on Artificial Intelligence. San Francisco, Cali-fornia, USA, February 4-9, 2017, 2852-2858. https://doi.org/10.1609/aaai.v31i1.10804

30. Yuan, L., Tay, F. E., Li, P., Zhou, L., Feng, J. Cycle-sum: Cycle-consistent Adversarial LSTM Net-Works for Unsupervised Video Summarization. Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, January 27-February 1, 2019, 9143-9150. https://doi.org/10.1609/aaai.v33i01.33019143

31. Zhao, B., Li, X., Lu, X. Hierarchical Recurrent Neural Network for Video Summarization. Pro-ceedings of the 25th ACM International Confer-ence on Multimedia. Mountain View, CA, USA, October 23-27, 2017, 863-871. https://doi.org/10.11 45/3123266.3123328

32. Zhao, B., Li, X., Lu, X. HSA-RNN: Hierarchical Sructure-Adaptive RNN for Video Summariza-tion. Proceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-23, 2018, 7405-7414. https://doi.org/10.1109/CVPR.2018.00773

33. Zhang, H., Hu, X., Ma, D., Wang, R., Xie, X. In-sufficient Data Generative Model for Pipeline Network Leak Detection Using Generative Adver-sarial Networks. IEEE Transactions on Cybernet-ics, 2022, 52(7), 7107-7120. https://doi.org/10.1109/TCYB.2020.3035518

34. Zhang, K., Chao, W. L., Sha, F., Grauman K. Summary Transfer: Exemplar-Based Subset Selec-tion for Video Summarization, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, June 27-30, 2016, 1059-1067. https://doi.org/10.1109/CVPR.20 16.120

35. Zhang, K., Chao, W. L., Sha, F., Grauman, K. Vid-eo Summarization with Long Short-Term Memory. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision-ECCV 2016, Lecture Notes in Computer Science, 9911, Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_47

36. Zhou, K., Qiao, Y., Xiang, T. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. Pro-ceedings of the 32nd AAAI Conference on Artifi-cial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018, 7582-7589. https://doi.org/10.1609/aaai.v32i1.12255

37. Zhu, J. Y., Park, T., Isola, P., Efros, A. A. Unpaired Image-to-image Translation Using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, October 22-29, 2017, 2223-2232. https://doi.org/10.1109/ICCV.2017.24