International Journal of Network Dynamics and Intelligence Scilight

Article

Video Summarization Using U-shaped Non-local Network

Shasha Zang [1], Haodong Jin [1,*], Qinghao Yu [1], Sunjie Zhang [1], and Hui Yu [2]

[1] School of Control Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2] School of Creative Technologies, University of Portsmouth, Portsmouth, United Kingdom
[*] Correspondence: hui.yu@port.ac.uk, 231260086@st.usst.edu.cn

Received: 22 November 2023
Accepted: 5 March 2024
Published: 26 June 2024

Abstract: Video summarization (VS) refers to extraction of key clips with important information from long videos to compose the short videos. The video summaries are derived by capturing a variable range of time dependencies between video frames. A large body of works on VS have been proposed in recent years, but how to effectively select the key frames is still a changing issue. To this end, this paper presents a novel U-shaped non-local network for evaluating the probability of each frame selected as a summary from the original video. We exploit a reinforcement learning framework to enable unsupervised summarization of videos. Frames with high probability scores are included into a generated summary. Furthermore, a reward function is defined that encourages the network to select more representative and diverse video frames. Experiments conducted on two benchmark datasets with standard, enhanced and transmission settings demonstrate that the proposed approach outperforms the state-of-the-art unsupervised methods.

Keywords: non-local neural network; video summarization; U-shaped network; reinforcement learning.

1. Introduction

With the widespread application of cameras in various scenarios, video data has exploded extensively in recent years. To make effective use and analysis of the video data, video summarization techniques are attracting increasing attention with the development of computer vision and deep learning technologies [1−5]. Video summarization, which applies a network model to extract important information segments from long videos, can significantly improve the efficiency of processing large amounts of video data. The processed video data can be used as the final presentation or combined with other technical means, such as inputs for subsequent tasks' networks, like natural language processing (NLP) for video captioning [6−8]. Furthermore, video summarization can be used to generate highlight reels as previews of the event video [9, 10]. Traditional local convolutional networks apply local convolution operations iteratively on images or videos, and this is computationally inefficient process due to its step-by-step nature. [11, 12]. Meanwhile, the propagation process must be carried out step by step, making it computationally inefficient. Therefore, the advantages of non-local operations, which calculate relationships such as correlations between the responses at various locations, are used to address the deficiencies of deep convolutional networks in this aspect [13]. Zhang et al. [9] proposed a video summary method based on LSTM for the first time. In addition, the method of combining LSTM with GAN [3, 14], and a subsequent improved algorithm [15] all have inherent limitations, that is, for LSTM, although long-term dependencies on video can be captured, they cannot be parallelized. GAN has the problem of unstable training, which indirectly affects the performance of methods. Non-local networks are also capable of capturing long-term dependencies in data, allowing them to model relationships and interactions between distant elements in a sequence. Moreover, the non-local network propagates information directly across the entire input, eliminating the need for step-by-step processing [16]. This efficient propagation mechanism helps improve the computational efficiency of the network. Overall, non-local networks provide a powerful tool for capturing long-term dependencies, modeling global relationships, and improving computational efficiency compared to local networks [17].

In this paper, we present a novel U-shaped neural network [18] model developed on top of a non-local network

[13] and this model combines a custom reward function for video summarization. Compared with structure-based methods, we propose a utility-based method, which defines predefined functions as well as network structures and is easy to implement. The proposed method can handle a wider range of objects. The policy-based reinforcement learning approach is employed to maximize the accumulated reward value, which is achieved by setting a custom reward function to filter the range of valid frames. Subsequently, the expected value of the accumulated reward is applied as a loss function to update the network parameters with stochastic gradient descent. The contributions can be summary as follows:

1) We introduce a novel U-shaped neural network model based on a non-local network structure. This model enhances the capability of capturing long-term dependencies and global relationships in video data, significantly improving the process of video summarization

2) An affective loss function is designed to integrate a custom reward function within a policy-based reinforcement learning framework. This approach focuses on maximizing the accumulated reward value, leading to more efficient selection of relevant frames in video summarization.

3) Experiments demonstrate that the proposed model outperforms other mainstream unsupervised methods.

2. Related work

The conventional methods for video summarization technologies typically begin by employing convolutional networks to extract features from each frame of the video [19]. Subsequently, the extracted video frame feature information is processed through a pre-built network model. According to the needs of the summary, the scoring criteria are established to judge the probability score of each frame of the video as a valid frame. While in processing the extracted feature information, the existing literature employs methods divided into utility-based and structure-based approaches [20]. Utility-based approaches rely on predefined functions and networks to filter the video frames, and are generally applied to identify salient objects and scenes in the video [21, 22]. This is accomplished by creating a visual attention model incorporating both motion and static attention. The motion attention model is responsible for capturing object motion information in the video, whereas static attention focuses on filtering background region information. Alternatively, some utility-based methods segment the video into different categories based on environmental information such as brightness and sound [23]. Subsequently, the starting and ending scenes are selected according to the movie grammar. These selections often need to be made artificially, and the resulting information is used to generate a summary of movie trailers.

Thanks to the advancement in machine learning, deep learning and other techniques, utility-based approaches are no longer limited to the use of attention mechanism networks [24, 25]. Instead, more suitable networks and techniques have been applied in this field. For example, the long and short-term memory networks (LSTM) in natural language processing, which are commonly used in speech recognition and structured prediction problems in video subtitle information [26]. Zhang et al. [9] treated the video summarization task as a structured prediction problem by using LSTM with aim to model the correlation between video frames. Rochan et al. [10] modified a full convolutional network originally used for semantic segmentation due to the fact that its input and output satisfy the conditions of video summarization. In addition to improve network structures, existing means are combined with other learning methods for unsupervised video summarization based on existing networks. For instance, unsupervised learning has achieved impressive performance by using adversarial learning combining attention mechanisms [15].

Another method, known as the structure-based approach, requires the prior knowledge of the hierarchical structure of the video being processed [27–29]. Since the typical movie layering structure is with a corresponding story structure, the structure-based method is generally based on scenes, actions and events, etc. as references for extracting information. Yeung et al. [30] divided the clustered shots and other similar shots into a class of scenes, and generated a scene transition map by using the scene transfer graph (STG) to represent the movie trailer. Li et al. [23] combined the visual images with audio information to achieve the classification of scenes, i.e., shots with the same background audio are also in the same class of scenes. At present, due to the dependency of understanding of video's hierarchical structure, this technique is more suitable for generating summaries of movies, TV series, event videos, etc., which have fixed editing rules.

The medical image processing faces the challenge of database shortage compared to general image processing [31]. The use of standard deep convolutional neural networks is prone to low training efficiency and overfitting. However, the emergence of the U-shaped network has addressed the issue of medical image segmentation [32]. This architecture maintains the original accuracy while using less training data. The U-shaped network consists of similar structures on both the left and right sides [33]. In the process of acquiring image information, the contraction path on the left side utilizes the down-sampling coding operation to gain the low-frequency detailed features of the image. On the other hand, the expansion path on the right side is responsible for integrating the information acquired at each

stage of the down-sampling phase and achieving the final output through the up-sampling decoding operation [34]. The key advantage of this architecture lies in its ability to enable the network to capture richer and more comprehensive information about the image features, which can then be combined and analyzed in the subsequent processing stage. Furthermore, the effectiveness extends to applications with samples from smaller datasets.

Reinforcement learning aims at maximizing the reward through the interaction between an intelligent agent and its environment [35, 36, 39]. Custom reward functions can be used to filter the desired video frames using the principles of reinforcement learning [37]. The reward mechanism plays a pivotal role in continuously trying to interact with the environment until the optimal state-behavior correspondence is obtained. Unlike supervised learning, reinforcement learning does not require the labeling of samples during the training process.

3. Methods

The proposed model employs a U-shaped network structure to capture the detail feature information from different dimensions, which enables the model to extract the feature information more effectively. And the non-local network is used to capture the temporal dependencies within the model. We first employ the GoogLeNet as a CNN model following [38] to extract features from the video data. Each frame of the video is treated as image data, and convolutional neural networks are used to extract features from the images to obtain feature vectors for each video frame. These feature vectors serve as the input to our designed network model based on a special convolutional network known as the non-local network [13]. The network structure is illustrated in Figure 1.

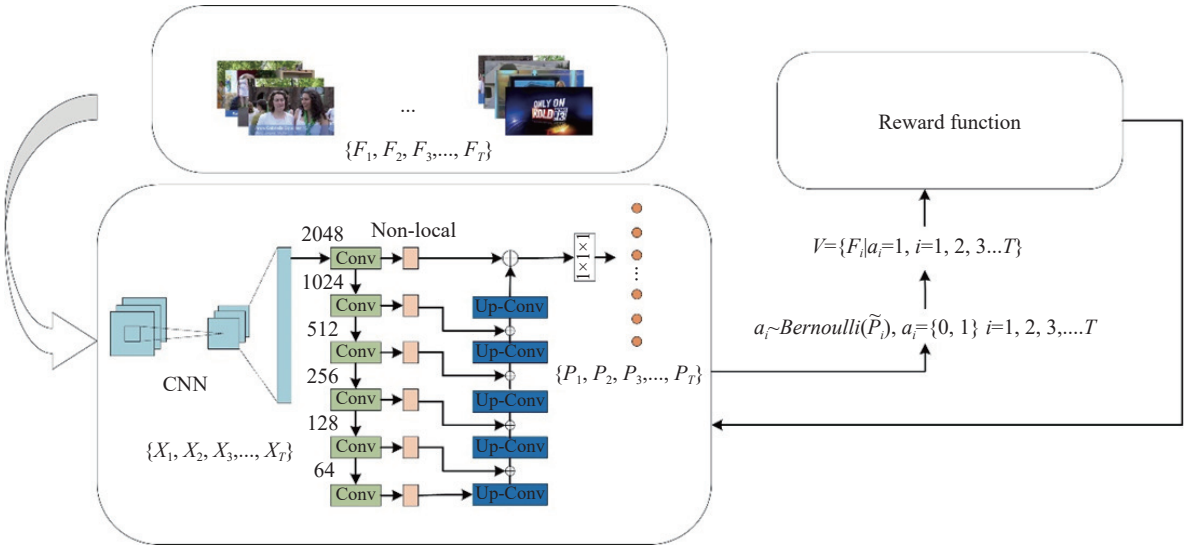

Figure 1. Basic architectural flow of a U-shaped structured non-local network applied to video summarization.

As shown in Figure 1, $\{F_1, F_2, F_3, \cdots, F_T\}$ represents a video containing T-frames. $\{x_1, x_2, x_3, \cdots, x_T\}$ refers to the video frame features after CNN processing. The importance probability scores $\{p_1, p_2, p_3, \cdots, p_T\}$ of each frame are generated by a U-shaped nonlocal network. These probabilities are transformed into 0/1 probabilities by the Bernoulli operation. Finally, the frames with a score of 1 are selected to form the V-set. In this paper, we take advantage of the U-shaped network structure, by capturing the contraction path of context relations and supporting the symmetric expansion path of precise non-localization. The use of data-dependent enhancement makes the sample training more effective. Figure 2 demonstrates the U-shaped network framework.

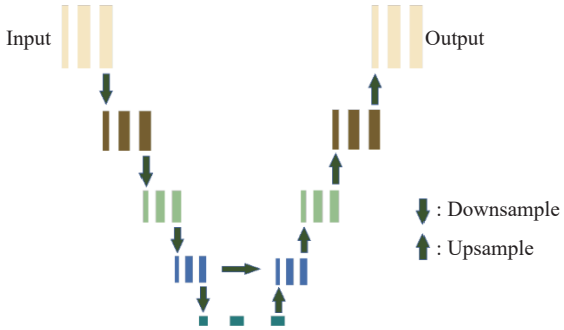

Figure 2. U-shaped network framework structure.

The U-shaped network consists of a shrinking path on the left and an expanding path on the right. The left path is similar to the traditional convolutional network model in that it is a unit consisting of convolutional operation and rectification linear units and pooling layers. The right path includes the up-sampling operation of the feature.

Video summaries usually require a summary module to evaluate the importance score of each frame, and then guide the training of the summary module by evaluating the generated summary. The proposed model is inspired by the U-shaped network structure in image segmentation, which can capture long-term time dependencies in deep neural networks through the U-shaped network summary module. In the proposed model, it is used to simulate the dependencies between video frames to generate an importance score for each frame. In the training stage, the stochastic gradient strategy method based on reinforcement learning principle is used to train the model. Finally, the video frame level importance score is used to calculate the shot level score, and the optimal segment is selected to generate the summary video.

The structure of the U-type network summary module in the model is mainly composed of the convolutional network and non-local module. A unit takes the visual features of the video frame $\{x_i\}_{i=1}^{T}$ as the input. These features are processed by the convolution layer, and the output is recorded as $\{C_i\}_{i=1}^{T}$. Then, the output is fed as the input of a non-local module, which captures the dependency between the features of the video frame. After the whole U-type network summary module is processed, the prediction probability $\{p_i\}_{i=1}^{T}$ for each video frame is generated.

For subsequent probability distribution operations, $\{p_i\}_{i=1}^{T}$ needs to map between 0 and the final probability for each video frame, and can be performed using the following formula:

$$\tilde{p}_i = sigmoid(Wp_i + b) \tag{1}$$

After that, the Bernoulli operation is performed to filter the frame composition of the summary video (V) with the prediction label 1. Finally, the summary video consisting of keyframes is formulated below:

$$\alpha_i \sim Bernoulli(\tilde{p}_i), \alpha_i = \{0,1\}, i = 1,2,3,\cdots T \tag{2}$$

$$V = \{F_i | \alpha_i = 1, i = 1,2,3\cdots\} \tag{3}$$

In the training stage, the self-defined reward function is used to train the network using the strategy gradient method of reinforcement learning. In this paper, the conditions for limiting video summaries are mainly included: selecting the most representative clips in different categories, using a clustering algorithm to classify the set of video frames to ensure that each frame in the selected clips has low similarity, and applying cosine distance to calculate the similarity between different video frames, which is used to filter out the clips with excessive similarity. Considering these conditions, the model calculates a probability score for each frame to be included in the summary. Afterward, the whole video is divided into segments by the KTS algorithm [50]. The probability score of each segment becoming a summary is then calculated based on the known probability scores of the video frames contained in each segment. The specific formula is as follows.

The K-median method is applied to classify the video frames initially, and the set of video frames with less similarity in each category is selected as the representative frames of this category.

$$R_{rep} = exp\left[-\frac{1}{T}\left(\sum_{i=1}^{T} \underset{l \epsilon V}{min} \|x_i - x_l\|_2\right)\right] \tag{4}$$

After clarifying the video frames in each category, they are compared in terms of similarity, and the video frames with low similarity are selected. Considering that a long interval is a high probability but not a decrease in the effectiveness of appearing the same scene, the similarity between two frames is chosen to be ignored when they are too far apart.

$$R_{div} = \begin{cases} \dfrac{1}{|V|(|V|-1)} \displaystyle\sum_{i \epsilon V} \sum_{\substack{i \epsilon V \\ l \neq i}} \left(1 - \dfrac{|x_l - \bar{x}_l||x_i - \bar{x}_i|^T}{\|x_i - \bar{x}_i\|_2 \|x_l - \bar{x}_l\|_2}\right), & |i-l| \leqslant \lambda \\ \dfrac{1}{|V|(|V|-1)}, & |i-l| > \lambda \end{cases} \tag{5}$$

where l and i represent the subscripts of key frames within the video sequence V. The parameter λ denotes the temporal distance between the frames indexed by l and i. The ultimate reward function formulated in the proposed method is expressed as R, which is the sum of R_{rep} and R_{div}.

The reward function is used as the expectation function, and the policy gradient algorithm in reinforcement

learning is used to maximize the expectation function. (6) The equation $\alpha_{1,T}$ represents the action sequence, and $p\left(\alpha_{1,T};\theta\right)$ is the probability distribution of the possible action sequences $\alpha_{1,T}$ under the parameter θ.

$$R(\theta) = \mathrm{E}_{p(\alpha_{1,T};\theta)}[R] \tag{6}$$

$$\nabla_\theta R(\theta) \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{T}(R_n - b)\nabla_\theta \log \pi_\theta(\alpha_i|P_i) \tag{7}$$

where α_i in is the action taken by the policy at the i moment, P_i is the output of the nonlocal network at the i moment, and N is the number of episodes. The parameter b is the baseline to reduce the increase in the variance. The value of b is the average of all rewards so far.

In addition, to control the percentage of frames selected by the model while improving performance, a regularization term with parameters is designed to limit the percentage of frames to be selected as shown in the following equation.

$$\theta = \theta - \alpha\nabla_\theta\left[\eta_1\left|\left|\frac{1}{T}\sum_{i=1}^{T}p_i - \varepsilon\right|\right|_2 + \eta_2\sum_{i,j}\theta_{i,j}^2 - R(\theta)\right] \tag{8}$$

The regularization term R(.) is added to avoid overfitting in Eq. (8) where the learning rate α is set to be le-05, η_1 and η_2 and is set to ber 0.01 and le-05, respectively.

Finally, after network processing, each video frame will get a probability score. The KTS algorithm [50] is used to take the points with drastic changes in the video as truncation points, and the video is divided into J segments, where the length of each segment is ℓ. The video frame probability score $g_{j,t}$ is used to calculate the probability score of each segment G_j. Finally, the video segment with the highest comprehensive probability score is selected to form the video summary as follows:

$$G_j = \frac{1}{\ell_j}\left\{\gamma_1\sum_{t=1}^{\ell_j}g_{j,t} + \gamma_2\sum_{t=1}^{\ell_j}[g_{j,t} - average\left(g_{j,t}\right)]^2\right\} \tag{9}$$

$$\begin{cases} max\sum_{j=1}^{K}\nu_j G_j \\ \sum_{j=1}^{K}\nu_j\ell_j \leqslant L, \nu_j \in 0,1 \end{cases} \tag{10}$$

4. Experiment

4.1. Experimental dataset and Settings

Experiments in this paper are mainly conducted based on four data sets, where most of the videos are sourced from various online websites. These four datasets are TVSum [40], SumMe [41], YouTube [42] and OVP [43]. It should be noted that the experimental results obtained were cross-validated five times in order to reduce randomness and noise in the data.

As shown in Figure 3, TVSum contains 50 different categories of videos collected from the YouTube video website, including news, documentaries, event logs, personal video logs, etc., with each video lasting 2-10 minutes. Each video in the TVSum dataset is manually rated and annotated for importance by 20 users.

Figure 3. TVSum Dataset.

According to Figure 4, SumMe is composed of 25 videos covering various topics ranging in duration from 1 minute to 6 minutes. Each frame of each video is rated by 15 to 18 people based on its importance, with a frame level importance rating provided. Finally, a summary video with a duration not exceeding 15% of the original video is generated based on the rating.

Figure 4. SumMe Dataset.

Figure 5 and Figure 6 illustrate the YouTube and OVP datasets, respectively, which are used as supplementary datasets in the enhancement experiment and migration experiment: 39 videos are included in YouTube and 50 videos are mostly documentaries in OVP.

To comprehensively evaluate the network performance, three sets of experiments are formed by combining the existing four basic datasets: the basic experiment, enhancement experiment, and transfer experiment.

Figure 5. YouTube Dataset.

Figure 6. OVP Dataset.

Canonical experiment: The network performance is tested on two basic datasets, TVSum and SumMe, and compared with the latest experimental data. As shown in Figure 7, the training set, validation set, and testing set are all derived from unrelated data from the same dataset. We randomly select 80% of the data from the TVSum dataset for training and validation, and the remaining 20% for testing.

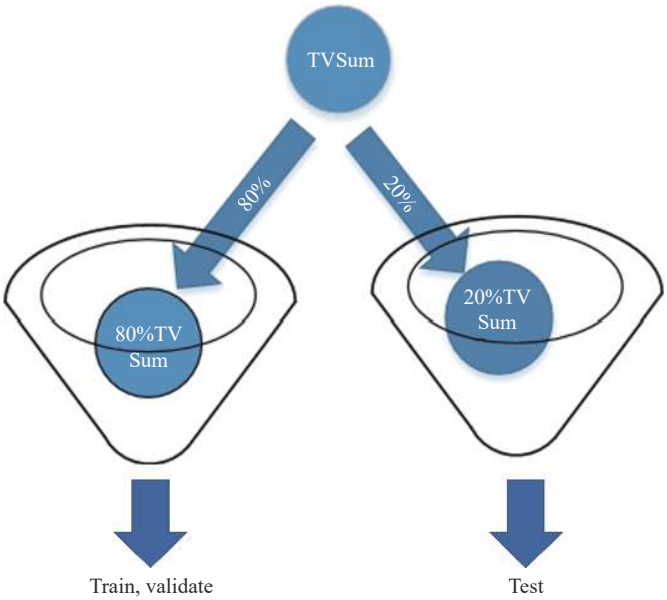

Figure 7. Canonical experiment dataset processing diagram.

Augmented experiment: YouTube and OVP are used as extended data to be added to the training and verification process for testing network performance. As shown in Figure 8, with TVSum as the basic data set, 80% of the data in TVSum and SumMe, YouTube and OVP are randomly selected to form the training set, and the remaining 20% of the samples in the TVSum data set are used as the test set. If the data set is based on SumMe, 80% of the data in SumMe, TVSum, YouTube and OVP constitutes the training set, while the remaining 20% of the samples in SumMe data set are taken as the test set.

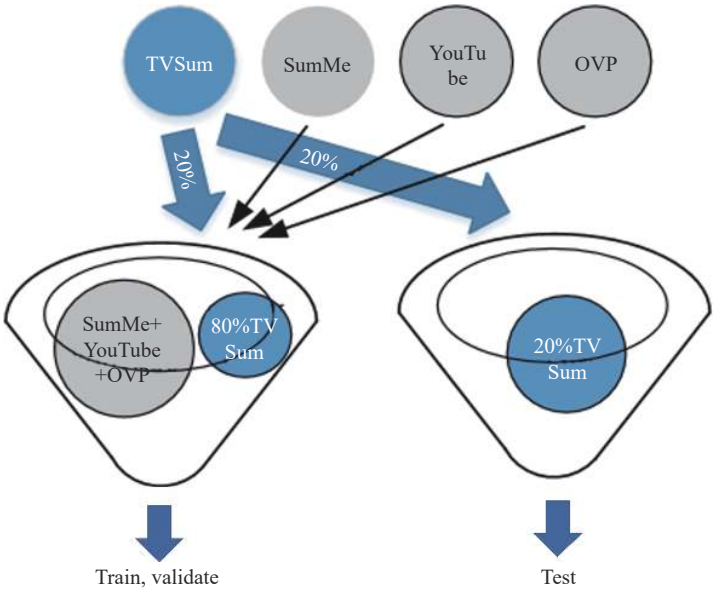

Figure 8. Augmented experiment dataset processing diagram.

Transfer experiment: This setting is to test whether the model can still generate good abstracts from the video without any relevant annotations. Therefore, as shown in Figure 9, we evaluate the learned model on a given dataset and use three different datasets that are unrelated to it as the training set. If the TVSum dataset is selected as the given dataset for testing network performance, SumMe is combined with YouTube and OVP to form a training set. If the SumMe dataset is selected as the given dataset, the other three datasets are combined to form a training set. This approach allows us to assess whether the model can still maintain good performance under the condition of unlabeled samples.

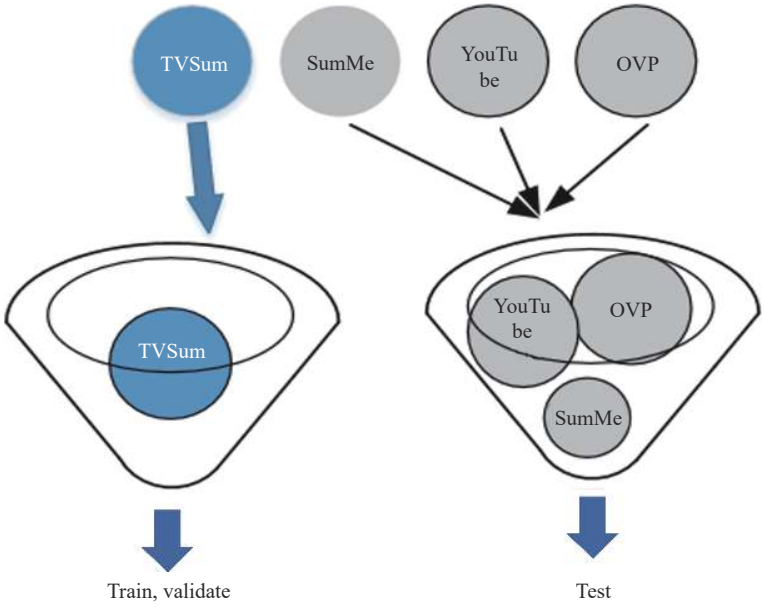

Figure 9. Transfer experiment dataset processing diagram.

4.2. Evaluation metrics

In this paper, the video frame importance score obtained through the video summarization network will be used to determine whether or not to use it for summarization fragments. Therefore, for network performance measurement, a harmonic average based on precision and repetition rate is used, known as the F_1 metric. According to existing video summarization networks, this metric is commonly used as a measure of network performance. In order to maintain consistency and fairness in comparison, this article also introduces the F-scores as the basis for evaluating the effectiveness of network summarization.

We calculate the F-scores by the network's accuracy P and completeness R, and use it as a metric to evaluate the network performance: S_G is the summary generated by the network while S_U is the summary generated by user annotation.

The following check rates are used to calculate how many frames of information of interest to the user are contained in the filtered video summary:

$$P = \frac{|S_G \cap S_U|}{|S_G|} \#(11) \tag{11}$$

Use the completeness rate to calculate how many video clips that users consider to contain important information are filtered into the video summary:

$$R = \frac{|S_G \cap S_U|}{|S_U|} \tag{12}$$

The F1 metric is based on the summed average of the calculated accuracy and completeness rates:

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

5. Analysis of experimental results

In this paper, we compare the proposed method with the current state-of-the-art video summarization methods GAN_dpp [44], UnpariedVSN [45], DR-DSN [46], SUM-FCN [10], PCDL [47], SUM-GAN-VAAE [15], SUM-GAN-AAE [15], DSR-RL-GR [48], DSR-RL-LSTM [48], Tessellation [49], and SUM-GAN-GEA [3]. The specific data is shown in Table 1:

Table 1 Performance of TVSum and SumMe with the new network (%)

Methods	SumMe	TVSum
GAN_dpp	39.1	51.7
UnpariedVSN	47.5	55.6
DR-DSN	41.4	57.6
SUM-FCN	41.5	52.7
PCDL	42.7	58.4
SUM-GAN-VAAE	45.7	57.6
SUM-GAN-AAE	48.9	58.3
DSR-RL-GR	50.3	60.2
DSR-RL-LSTM	43.8	61.4
Tessellation	41.4	64.1
SUM-GAN-GEA	53.4	61.3
OURS	**53.4**	**63.0**

We compare the performance of our model with nine other models on the SumMe and TVSum datasets. As presented in Table 1, the summary performance of our model (for videos) compares favorably with the existing state-of-the-art methods on both datasets. In particular, there is a significant improvement on the SumMe dataset. However, the performance of the Tessellation method on the TVSum data set is better than ours. The reason is that it is a data set customization technology and has better performance on a specific data set. Its performance on SumMe is far inferior to our model. The SUM-GAN-GEA model has the same score as ours on the SumMe dataset, but our model is superior to it on TVSum. We believe that the temporal dependencies of obtaining videos on different dimensions are helpful in the generation of the final video summaries.

In addition to the experiments on the two base datasets, we have also conducted enhancement experiments and migration experiments using the supplemental datasets and compared the performance with methods DR-DSN [46], DSR-RL-GR [48] and DSR-RL-LSTM [48], as shown in Table 2.

Table 2 Results of different methods for basal, enhancement and migration experiments (%)

Dataset	Method	Canonical	Augmented	Transfer
SumMe	DR-DSN	41.4	42.8	42.4
	DSR-RL-GR	50.3	48.5	48.7
	DSR-RL-LSTM	43.8	43.3	44.2
	OURS	**53.4**	**50.1**	**50.8**
TVSum	DR-DSN	57.6	58.4	57.8
	DSR-RL-GR	60.2	59.2	58.9
	DSR-RL-LSTM	61.4	60.2	59.9
	OURS	**63.0**	**61.3**	**61.1**

As shown in Table 2, from the data of both the basic experiment and the other two groups of supplementary experiments, the proposed network performs better than other networks. As the performance of the network on the TVSum dataset in isolation, the performance of the network video summary proposed in the basic experiment setting improves by about 2% compared with the optimal network at present. In the migration experiment setting, the network proposed in this chapter shows notable improvements ranging from a minimum of 1% to a maximum of 3% compared with other networks. In the enhanced settings, the proposed network enhances with performance improvements ranging from a minimum of 1% to a maximum of less than 3%. Based on the horizontal data comparison results, it becomes evident that the network model performs the best under the basic experimental setting. The performance of the network under the enhanced experimental setting is 2% worse than that under the basic experiment, but still better than that under the transfer experimental setting.

The performance effect of the network on the SumMe dataset is consistent with that on the TVSum dataset. The difference is that the network performance in the basic experimental setting is 3% higher than that of the optimal network. In the enhanced experimental setting, the proposed network video summary improves by less than 2% compared with the optimal network. In addition, the proposed network improves by 2% in the migration experiment setting. Based on the experimental results integrated on two datasets, our network's performance in the basic experi-

mental setting is always the best group. The network performance in the enhanced experimental setting is 3% lower than that in the basic experimental setting, while better than that in the transfer experiment.

In summary, from the experimental data of the network enhancement effect under the three experimental settings, the enhancement effect of the network model proposed under the basic experimental settings is more obvious than that of the other two groups of supplementary experiments. Most of the network performs in the basic experiments better than those in the two groups of supplementary experiments, and the enhancement effect is also the most obvious in the basic experiments, followed by the enhancement experiments. The migration experiment is the least obvious. However, the effect of any data network on SumMe dataset is not as good as the TVSum dataset. It can be seen from the experimental data that under the basic experimental setting, the video summary effect of the network on the SumMe dataset is 10% less than the TVSum dataset. In addition, the performance of the network in the TVSum dataset is 10% higher than that in the SumMe dataset under the enhanced experiment setting and the migration experiment setting. Based on the characteristics of these two data sets, the possible reason is that the video length and the number of videos in TVSum are longer than the former, so the amount of data contained in TVSum dataset is large. Thus, the network can obtain more effective information for training, which leads to more accurate prediction. From the above data results, it can be concluded that the U-shaped structure can help the network to obtain more information in the video, and thus improve the performance of the network video summary.

In addition to the above data results, Figure 10 shows the video frame information contained in the summary generated by the video.

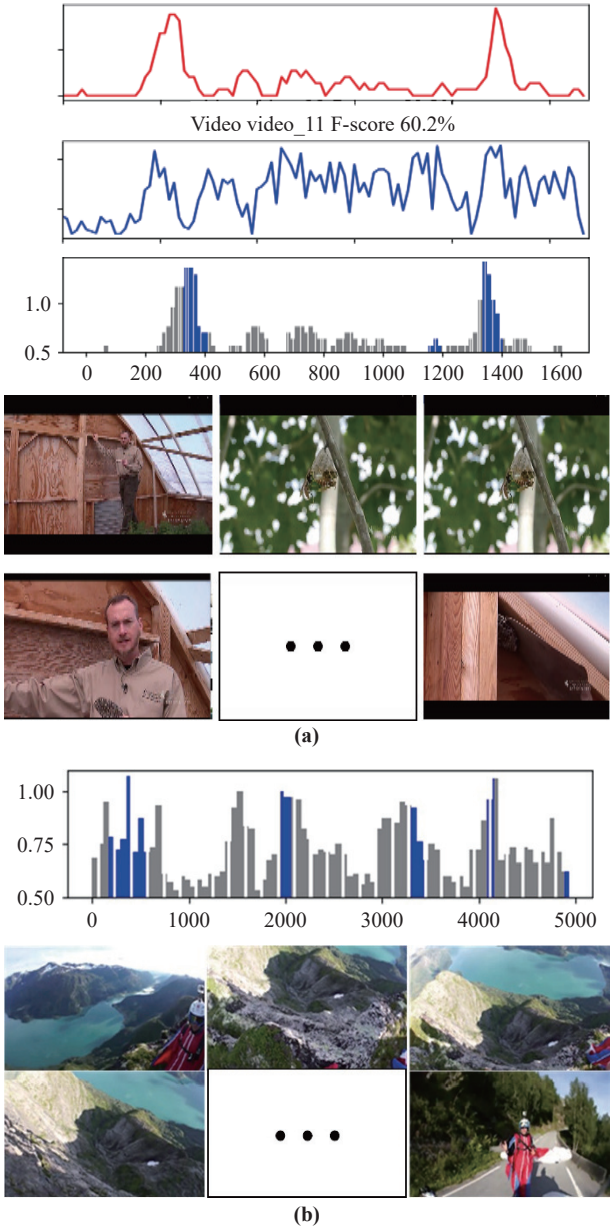

Figure 10. Visualization results of video_11 **(a)** in the SumMe dataset and video_43; **(b)** in the TVSum dataset.

The gray part of Figure10 shows the annotated labels of the video frames, representing the importance of each frame labeled by 20 individuals. The blue part indicates the importance of each frame predicted by our network. We select a video in each of the TVSum and SumMe datasets and analyze the effect of the network on its video summarization. It can be found that the regions with higher probability scores in the graph can be accurately predicted by the network. Thus, accurate extraction of valid information from the video can be achieved. However, the network only extracts the regions where the high probability scores are concentrated, so it is not very clear whether the extracted information is comprehensive, which is one of the issues we want to explore in the future.

6. Conclusion

In this paper, we have designed a U-shaped non-local network to extract information from videos across various dimensions and to discern the temporal dependencies among them. The probability of selecting each frame for the summary has been determined based on these conditional relationships. By analyzing the probability scores assigned to each frame, we have calculated the scores for each video segment. Ultimately, we have filtered out several video segments that meet our criteria to form the video summary. Our experiments have been conducted on two base datasets to demonstrate that understanding the temporal dependencies between different dimensions of video information enhances the effectiveness of video summarization. Additionally, we have incorporated two supplementary datasets to conduct enhancement and migration experiments, with the experimental data indicating improved performance of our model. In future work, we aim to explore additional techniques to capture the temporal dependencies of video information across different dimensions for improving video summarization models.

Author Contributions: Shasha Zang: Conceptualization, Methodology, Software, Writing – Original Draft. **Haodong Jin:** Project Administration, Writing – Review & Editing, Manuscript Submission. **Qinghao Yu:** Data Curation, Formal Analysis, Investigation. **Sunjie Zhang:** Validation, Visualization, Resources. **Hui Yu:** Supervision, Writing – Review & Editing.

Funding: No external funding was received for this study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zang, S.S.; Yu, H.; Song, Y.; *et al*. Unsupervised video summarization using deep Non-Local video summarization networks. Neurocomputing, **2023**, *519*: 26−35. doi: 10.1016/j.neucom.2022.11.028
2. Liang, G.Q.; Lv, Y.B.; Li, S.C.; *et al*. Video summarization with a dual-path attentive network. Neurocomputing, **2022**, *467*: 1−9. doi: 10.1016/j.neucom.2021.09.015
3. Yu, Q.H.; Yu, H.; Wang, Y.X.; *et al*. SUM-GAN-GEA: Video summarization using GAN with gaussian distribution and external attention. Electronics, **2022**, *11*: 3523. doi: 10.3390/electronics11213523
4. Alfasly, S.; Lu, J.; Xu, C.; *et al*. FastPicker: Adaptive independent two-stage video-to-video summarization for efficient action recognition. Neurocomputing, **2023**, *516*: 231−244. doi: 10.1016/j.neucom.2022.10.037
5. Liu, T.R.; Meng, Q.J.; Huang, J.J.; *et al*. Video summarization through reinforcement learning with a 3D spatio-temporal U-Net. IEEE Trans. Image Process., **2022**, *31*: 1573−1586. doi: 10.1109/TIP.2022.3143699
6. Ming, Y.; Hu, N.N.; Fan, C.X.; *et al*. Visuals to text: A comprehensive review on automatic image captioning. IEEE/CAA J. Autom. Sin., **2022**, *9*: 1339−1365. doi: 10.1109/JAS.2022.105734
7. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Netw. Learn. Syst., **2021**, *32*: 604−624. doi: 10.1109/TNNLS.2020.2979670
8. Wang, F.Y.; Miao, Q.H.; Li, X.; *et al*. What does ChatGPT say: The DAO from algorithmic intelligence to linguistic intelligence. IEEE/CAA J. Autom. Sin., **2023**, *10*: 575−579. doi: 10.1109/JAS.2023.123486
9. Zhang, K.; Chao, W.L.; Sha, F.; *et al*. Video summarization with long short-term memory. In *Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 766–782. doi:10.1007/978-3-319-46478-7_47
10. Rochan, M.; Ye, L.W.; Wang, Y. Video summarization using fully convolutional sequence networks. In *Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 347–363. doi:10.1007/978-3-030-01258-8_22
11. He, K.M.; Zhang, X.Y.; Ren, S.Q.; *et al*. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, 2016; pp. 770–778. doi:10.1109/CVPR.2016.90
12. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput., **1997**, *9*: 1735−1780. doi: 10.1162/neco.1997.9.8.1735
13. Wang, X.L.; Girshick, R.; Gupta, A.; *et al*. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: New York, 2018; pp. 7794–7803. doi:10.1109/CVPR.2018.00813

14. Apostolidis, E.; Metsai, A.I.; Adamantidou, E.; *et al*. A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery, Nice, France, 21 October 2019*; ACM: New York, 2019; pp. 17–25. doi:10.1145/3347449.3357482
15. Apostolidis, E.; Adamantidou, E.; Metsai, A.I.; *et al*. Unsupervised video summarization via attention-driven adversarial learning. In *Proceedings of the 26th International Conference on Multimedia Modeling, Daejeon, South Korea, 5–8 January 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 492–504. doi:10.1007/978-3-030-37731-1_40
16. Liu, D.; Wen, B.H.; Fan, Y.C.; *et al*. Non-local recurrent network for image restoration. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 3–8 December 2018*; Curran Associates Inc.: Morehouse Lane, Red Hook, 2018; pp. 1680–1689.
17. Aung, N.; Kechadi, T.; Chen, L.M.; *et al*. IP-UNet: Intensity projection UNet architecture for 3D medical volume segmentation. arXiv preprint arXiv: 2308.12761, 2023.
18. Shahi, K.; Li, Y.M. Background replacement in video conferencing. Int. J. Netw. Dyn. Intell., **2023**, *2*: 100004. doi: 10.53941/ijndi.2023.100004
19. Haq, H.B.U.; Asif, M.; Bin, M. Video summarization techniques: A review. Int. J. Sci. Technol. Res., **2020**, *9*: 146−153.
20. Li, Y.; Lee, S.H.; Yeh, C.H.; *et al*. Techniques for movie content analysis and skimming: Tutorial and overview on video abstraction techniques. IEEE Signal Process. Mag., **2006**, *23*: 79−89. doi: 10.1109/MSP.2006.1621451
21. Li, Y.M.; Bhanu, B. Utility-based camera assignment in a video network: A game theoretic framework. IEEE Sensors J., **2011**, *11*: 676−687. doi: 10.1109/JSEN.2010.2051148
22. Prangl, M.; Szkaliczki, T.; Hellwagner, H. A framework for utility-based multimedia adaptation. IEEE Trans. Circuits Syst. Video Technol., **2007**, *17*: 719−728. doi: 10.1109/TCSVT.2007.896650
23. Li, Y.; Kuo, C.C.J. *Video Content Analysis Using Multimodal Information: For Movie Content Extraction, Indexing and Representation*; Springer: New York, 2003. doi:10.1007/978-1-4757-3712-7
24. Brauwers, G.; Frasincar, F. A general survey on attention mechanisms in deep learning. IEEE Trans. Knowl. Data Eng., **2023**, *35*: 3279−3298. doi: 10.1109/TKDE.2021.3126456
25. Niu, Z.Y.; Zhong, G.Q.; Yu, H. A review on the attention mechanism of deep learning. Neurocomputing, **2021**, *452*: 48−62. doi: 10.1016/j.neucom.2021.03.091
26. Zazo, R.; Lozano-Diez, A.; Gonzalez-Dominguez, J.; *et al*. Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks. PLoS One, **2016**, *11*: e0146917. doi: 10.1371/journal.pone.0146917
27. Chen, H.W.; Kuo, J.H.; Chu, W.T.; *et al*. Action movies segmentation and summarization based on tempo analysis. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, NY, USA, 15–16 October 2004*; ACM: New York, 2004; pp. 251–258. doi:10.1145/1026711.1026752
28. Smith, M.A.; Kanade, T. *Video Skimming for Quick Browsing Based on Audio and Image Characterization*; Carnegie Mellon University: Pittsburgh, PA, USA, 1995.
29. Rasheed, Z.; Sheikh, Y.; Shah, M. On the use of computable features for film classification. IEEE Trans. Circuits Syst. Video Technol., **2005**, *15*: 52−64. doi: 10.1109/TCSVT.2004.839993
30. Yeung, M.M.; Yeo, B.L. Time-constrained clustering for segmentation of video into story units. In *Proceedings of 13th International Conference on Pattern Recognition, Vienna, Austria, 25–29 August 1996*; IEEE: New York, 1996; pp. 375–380. doi:10.1109/ICPR.1996.546973
31. Li, X.; Li, M.L.; Yan, P.F.; *et al*. Deep learning attention mechanism in medical image analysis: Basics and beyonds. Int. J. Netw. Dyn. Intell., **2023**, *2*: 93−116. doi: 10.53941/ijndi0201006
32. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241. doi:10.1007/978-3-319-24574-4_28
33. Liu, L.L.; Cheng, J.H.; Quan, Q.; *et al*. A survey on U-shaped networks in medical image segmentations. Neurocomputing, **2020**, *409*: 244−258. doi: 10.1016/j.neucom.2020.05.070
34. Siddique, N.; Paheding, S.; Elkin, C.P.; *et al*. U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access, **2021**, *9*: 82031−82057. doi: 10.1109/ACCESS.2021.3086020
35. Mnih, V.; Kavukcuoglu, K.; Silver, D.; *et al*. Playing atari with deep reinforcement learning. arXiv preprint arXiv: 1312.5602, 2013.
36. Rezaei, M.; Tabrizi, N. A survey on reinforcement learning and deep reinforcement learning for recommender systems. In *Proceedings of the 4th International Conference on Deep Learning Theory and Applications, Rome, Italy, 13–14 July 2023*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 385–402. doi:10.1007/978-3-031-39059-3_26
37. Zeng, N.Y.; Li, H.; Wang, Z.D.; *et al*. Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip. Neurocomputing, **2021**, *425*: 173−180. doi: 10.1016/j.neucom.2020.04.001
38. Deng, J.; Dong, W.; Socher, R.; *et al*. ImageNet: A large-scale hierarchical image database. In *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009*; IEEE: New York, 2009; pp. 248–255. doi:10.1109/CVPR.2009.5206848
39. Liu, S.M.; Xia, Y.F.; Shi, Z.S.; *et al*. Deep learning in sheet metal bending with a novel theory-guided deep neural network. IEEE/CAA J. Autom. Sin., **2021**, *8*: 565−581. doi: 10.1109/JAS.2021.1003871
40. Song, Y.L.; Vallmitjana, J.; Stent, A.; *et al*. TVSum: Summarizing web videos using titles. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*; IEEE: New York, 2015; pp. 5179–5187. doi:10.1109/CVPR.2015.7299154
41. Gygli, M.; Grabner, H.; Riemenschneider, H.; *et al*. Creating summaries from user videos. In *Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 505–520. doi:10.1007/978-3-319-10584-0_33
42. de Avila, S.E.F.; Lopes, A.P.B.; da Luz, A. Jr.; *et al*. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognit. Lett., **2011**, *32*: 56−68. doi: 10.1016/j.patrec.2010.08.004
43. Gong, B.Q.; Chao, W.L.; Grauman, K.; *et al*. Diverse sequential subset selection for supervised video summarization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 8–13 December 2014*; MIT

Press: Cambridge, 2014; pp. 2069–2077.
44. Mahasseni, B.; Lam, M.; Todorovic, S. Unsupervised video summarization with adversarial LSTM networks. In *Proceedings of 2017 IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, 2017; pp. 202–211. doi:10.1109/CVPR.2017.318
45. Rochan, M.; Wang, Y. Video summarization by learning from unpaired data. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019*; IEEE: New York, 2019; pp. 7902–7911. doi:10.1109/CVPR.2019.00809
46. Zhou, K.Y.; Qiao, Y.; Xiang, T. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans Louisiana USA, 2–7 February 2018*; AAAI Press: Palo Alto, 2018; p. 929. doi:10.1609/aaai.v32i1.12255
47. Zhao, B.; Li, X.L.; Lu, X.Q. Property-constrained dual learning for video summarization. IEEE Trans. Neural Netw. Learn. Syst., **2020**, *31*: 3989−4000. doi: 10.1109/TNNLS.2019.2951680
48. Phaphuangwittayakul, A.; Guo, Y.; Ying, F.L.; *et al*. Self-attention recurrent summarization network with reinforcement learning for video summarization task. In *Proceedings of 2021 IEEE International Conference on Multimedia and Expo, Shenzhen, China, 5–9 July 2021*; IEEE: New York, 2021; pp. 1–6. doi:10.1109/ICME51207.2021.9428142
49. Kaufman, D.; Levi, G.; Hassner, T.; *et al*. Temporal tessellation: A unified approach for video analysis. In *Proceedings of 2017 IEEE International Conference on Computer Vision*, Venice, Italy, 22–29 October 2017; IEEE: New York, 2017; pp. 94–104. doi:10.1109/ICCV.2017.20
50. Potapov, D.; Douze, M.; Harchaoui, Z.; *et al*. Category-specific video summarization. In *Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 540–555. doi:10.1007/978-3-319-10599-4_35

Citation: Zang, S.; Jin, H.; Yu, Q.; *et al*. Video Summarization Using U-shaped Non-local Network. *International Journal of Network Dynamics and Intelligence*. 2024, 3(2), 100013. doi: 10.53941/ijndi.2024.100013