Application of Decomposition Expression in Digital Video Object Segmentation

Jianfu Kong
School of Fine Arts and Design, Henan Vocational University of Science and Technology, China
jianfukong321@outlook.com

Abstract: *In the field of actual Video Object Segmentation (VOS), traditional techniques have poor adaptability and insufficient segmentation results. Therefore, based on existing problems, an Unsupervised Video Object Segmentation (UVOS) technique based on convolutional networks is proposed. Firstly, the method of decomposing expressions is used to handle the spatiotemporal relationship between the reference frame and the target frame, and video object reconstruction is achieved through similarity calculation. For target segmentation in motion scenes, a Single Linear Bottleneck Operator (SLBO) is introduced for feature extraction, and pooling compensation is used to optimize feature information loss. For general scene segmentation, a spatiotemporal similarity segmentation technique is introduced to achieve target video segmentation for complex scenes. In the foreground segmentation test of sports scenes, the Change Detection Benchmark Dataset 2014 (CDNet.20I4SM) dataset was selected to test the model's loss performance in different scenarios. In adverse weather scenario training, the proposed model tends to converge after 40 iterations, with a loss value of 0.276, which is superior to the Foreground image Segmentation (FgSegNet_), the Convolutional Networks for Biomedical Image Segmentation (MU Net), Cascade Convolutional Neural Network (Cascade CNN) models; In the accuracy test, the proposed FS-LBPC model tended to converge after 50 iterations, with a precision P-value of 0.963. It performed the best among the four segmentation models the FgSegNet_, MU Net, Cascade CNN, and a real-time Foreground Segmentation network based on single Linear Bottleneck and Pooling Compensation (FS-LBPC). Usually, the Densely Annotated VIdeo Segmentation (DAVIS16) dataset is selected for video scene segmentation, which has the best segmentation performance in horse racing and animal flight scenes, with segmentation accuracy of 0.976 and 0.965, respectively. In summary, the VOS technology has excellent application effects in practical scenarios, providing important technical references for the improvement of image and video processing and segmentation technology.*

Keywords: *Video object segmentation, unsupervised, deep learning, decomposing expression, feature, bottleneck operator, foreground segmentation.*

Received April 29, 2024; accepted October 14, 2024
https://doi.org/10.34028/iajit/22/1/3

1. Introduction

In the era of rapid development of information and data, the entertainment industry led by video is thriving. Video Object Segmentation (VOS) is a fundamental task in digital vision technology and plays an important role in the field of image vision [32]. VOS is an important computer vision task with broad application potential. Traditional VOS techniques mainly use methods based on frame difference, gradient, and region growth [28]. Although the above technologies have good applications in video object segmentation, they are generally suitable for general scenarios and do not accurately handle targets and boundaries, making them unable to meet the segmentation requirements of more complex scenes [19]. Therefore, in response to the problems faced by traditional technologies, a decomposition based Unsupervised Video Object Segmentation (UVOS) technique based on convolutional frameworks is proposed. By calculating the similarity between the target frame and the reference frame, the reconstruction of the video target is achieved. Simultaneously considering the difficulty of video segmentation in both motion and general scenes, a single linear bottleneck and spatiotemporal similarity calculation are introduced to achieve the processing of video segmentation problems in different scenes. There are two innovative points in the research content. Firstly, a separately expressed UVOS technique is proposed, which utilizes deep network structures to automatically learn feature representations, significantly improving the accuracy and stability of segmentation. Secondly, considering the issue of insufficient segmentation performance in different scenes, separate processing is carried out for both motion scenes and general scenes to improve the adaptability and segmentation performance of the research technology. There are two novelty points to this study. One is to optimize the problem of traditional techniques being unable to adapt to complex segmentation scenarios, and propose a segmentation technique based on improved Convolutional Neural Network (CNN) to achieve processing of more complex segmentation tasks. Secondly, it improves the problem of traditional CNNs being unable to process global information features by introducing linear bottleneck operators and pooling compensation optimization models to further enhance the application effectiveness of the technology. The main contribution of the research content is twofold. The first point is that the research technology has better segmentation performance, providing support for the improvement of VOS technology. The second point is that the proposed technology has better adaptability and processing ability, adding considerations for multiple segmentation scenarios, and providing effective technical references for VOS. The entire research work is divided into four parts, one of which is related research and discussion on the latest VOS technology; the second part is to establish an unsupervised video segmentation model based on convolutional networks and improve the model; the third part conducts performance testing and discussion on the video segmentation technology proposed by the research institute. The fourth part is a summary of the entire article.

2. Related Work

VOS is a computer vision technology that performs pixel segmentation on continuous frame images, which can accurately locate and track different objects or targets in the image. It is widely used in intelligent monitoring, automatic driving, medical imaging analysis and other fields, providing a powerful tool for automation and intelligence. Fan *et al*. [8] found that digital video has visual distortion issues, which can affect users' experience of video quality. To more accurately evaluate the impact of these distortions on vision, researchers proposed a new database and conducted a subjective quality assessment of the database. At the same time, they also evaluated several of the best performing video processing technologies and found that the video database can accurately evaluate real-time mobile videos, which has broad application prospects. Falaschetti *et al*. [7] conducted research on semantic segmentation in modern intelligent vehicles to address the challenges faced by modern intelligent vehicles in intensive operations. The semantic video segmentation scene on smart cars is complex, requiring high computational power and energy consumption. To achieve this goal, a low rank CNN architecture for real-time semantic segmentation is proposed, and tensor decomposition technology has been applied to the kernel of the universal convolutional layer, while combining UNet and ResNet models to optimize the architecture. Through experimental testing, the research technology has good stability and low power consumption capabilities. Wang *et al*.'s [29] study explored the impact of visual attention on the understanding of video object patterns. They found that in dynamic, task driven viewing processes, there is a strong correlation between the objects of human attention and clear judgments of the main objects. Based on these findings, researchers proposed a video solution and demonstrated its superiority and fast processing speed through experiments. Giraldo *et al*. [10] focused their research on video segmentation processing techniques. They proposed a semi supervised processing model that can achieve competitive results on both static and mobile camera videos, and requires less labeled data compared to current state-of-the-art methods. Ammar *et al*. [2] proposed a new method that utilizes a deep unsupervised anomaly detection framework and generative adversarial models to segment and classify moving objects in video sequences. This method has been evaluated on multiple datasets, demonstrating its effectiveness and superiority. Fu *et al*. [9] proposed a novel multimodal video instance segmentation method. This method combines motion information and appearance information to improve segmentation accuracy and achieves state-of-the-art segmentation performance on multiple datasets, demonstrating its superiority and robustness. Chan *et al*. [4] conducted research on video image segmentation techniques in the medical field and found that traditional image annotation is time-consuming and difficult to manually annotate. Deep learning-based video segmentation technology can effectively address this issue. Therefore, a deep learning video multi-scale segmentation framework is proposed to better obtain image scale information through encoders and decoders. Applying it to specific scenarios, this technology has excellent application effects and important research value for medical image processing. Logeshwaran *et al*. [18] conducted research on current photography techniques, where image segmentation, calibration, and pixel processing are key tasks in the field of photography. In order to improve the quality of shooting images, an enhanced video segmentation processing technique is proposed, which improves the overall video effect by increasing video resolution, color, and image contrast. Finally, the technology was applied to specific scenarios, and the results showed that compared to similar technologies, it has better performance in video processing. Zheng *et al*. [30] found that in the field of video segmentation, weakly supervised training cannot meet the requirements of high-quality video segmentation. Therefore, in order to address the aforementioned issues, a study was conducted on the characteristics of video segmentation, and a solution was proposed to address weak supervision. Among them, contrastive negative sample mining is introduced, using learnable Gaussian masks to generate positive samples, highlighting the most relevant video frames for querying, thereby improving video segmentation performance. Relevant experimental analysis shows that this technology has good application effects in practical scenarios, overcoming the shortcomings of traditional weakly supervised video processing.

In the field of video segmentation processing, unsupervised video segmentation technology has attracted much attention. The UVOS proposed by Lee *et al*. [14] is a binary labeling problem per pixel, which aims to separate foreground objects from the background in the video without using the ground truth mask of the foreground objects. In order to improve the performance of UVOS, a simple frame selector was proposed in the study, which can select a "simple" reference frame to make subsequent VOS simpler. In addition, a new framework called iterative mask prediction was proposed in the study. Tested on three UVOS benchmark sets, including the Densely Annotated VIdeo Segmentation (DAVIS16) dataset, Freiburg Berkeley Motion Segmentation (FBMS) dataset, and SegTrack dataset, the proposed models by all exhibit excellent performance. Zhou *et al*. [33] proposed a motion attention model based on flow edges to solve the UVOS problem. He uses motion focused encoders to combine learning space and time characteristics, and designs a flow edge connection module to hallucinate the edges of blurred or missing areas in the optical flow. The experimental results on two challenging common benchmark FBMS datasets show that the proposed scheme is advantageous compared to state-of-the-art methods. Vecchio *et al*. [27] found that integrating position prior into VOS has been proven to be an effective strategy for improving performance. However, their large-scale application is not feasible. Gamification can help reduce annotation burden, but it still requires user participation. To address this issue, they proposed a VOS framework that utilizes the combined advantages of user feedback for segmentation and gamification strategies. This framework reproduces the ability of humans to accurately locate moving objects and uses simulated feedback to drive decisions in fully convolutional deep segmentation networks. Experiments on the DAVIS-17 benchmark show that the model can provide users with prior knowledge. Huang *et al*. [11] conducted research on video segmentation and found that most methods rely on pixel-by-pixel manual annotation, which is very time-consuming and expensive. To solve this problem, researchers experimented with the method of achieving VOS through graffiti level supervision. However, using traditional network architecture and learning objective functions does not work well because the supervised information is sparse and incomplete. Therefore, they proposed the graffiti attention module and graffiti supervision loss as new elements to learn the VOS model to solve this problem. This method is close to the method that requires dense pixel by pixel annotation [16]. Raman *et al*. [23] proposed an algorithm to measure the efficiency of workflow scheduling in their research, which is used to solve traditional segmentation task problems. In the cloud environment, this technology schedules tasks to available resources through backfill algorithms and reduces the percentage of migration, compared to traditional "first come, first served" algorithms. In addition, they use the Berger model to measure the fairness of resource allocation and determine task reassignment based on the fairness value. Through experimental research, it has been proven that the proposed technology has excellent performance in both performance and efficiency. On the other hand, Zhu *et al*. [34] proposed a separable structural modeling method for semi Supervised Video Object Segmentation (SVOS). Unlike existing methods, this method not only captures the pixel level similarity relationship between the reference frame and the target frame, but also reveals the separable structure of the specified object in the target frame. This technology calculates a pixel-by-pixel similarity matrix using the representation of reference pixels and target pixels, and selects the highest-level reference pixels for target pixel classification. In the structural modeling branch, research techniques have extracted shared and individual features that can effectively represent the entire object and its components. In addition, this method is a fast algorithm that does not require online fine-tuning or any post-processing. This method has achieved excellent performance in terms of speed and accuracy through experiments.

Based on the above research, it can be seen that video segmentation technology has a large number of applications in many scenarios. Traditional supervised video segmentation has limitations in many scenarios, and traditional segmentation techniques face problems in feature extraction, background processing, and other aspects, which cannot meet specific scene requirements. In this regard, research mainly starts from the unsupervised direction, based on improved CNN network technology, to conduct research on VOS in specific scenarios, in order to improve the processing effect of computer vision technology and promote the effective application of intelligent video segmentation technology.

3. Construction of Digital VOS Model Based on Decomposition Expression

This section mainly conducts relevant research on UVOS technology and adopts unsupervised technology for VOS. On the basis of traditional UVOS technology, the decomposition method is adopted to improve the video segmentation effect. At the same time, target segmentation is targeted at different scenes, and VOS models are constructed for both motion scenes and general scenes to achieve effective processing of different targets.

3.1. Construction of VOS Model Based on Decomposition Method

VOS technology is one of the important research topics in the field of computing. Most traditional segmentation techniques are based on supervised segmentation, which cannot meet the segmentation requirements for special and complex scenes. SVOS technology can achieve good video segmentation results through a large number of segmentation label training. However, model training relies on a large amount of annotated data, which cannot meet specific video segmentation scenarios [15]. Considering the segmentation problem faced by SVOS technology, a video foreground segmentation model based on spatiotemporal similarity is introduced on the basis of decomposed expression to solve the problem [21]. VOS has different segmentation standards based on the type of segmentation, including pixel level consistency, motion information, spatial relationships, and other segmentation standards. In unsupervised segmentation, color is an important consideration and a relatively stable supervised signal. The input color is compared with the target frame image to determine the loss. Unsupervised target frame reconstruction is shown in Figure 1.

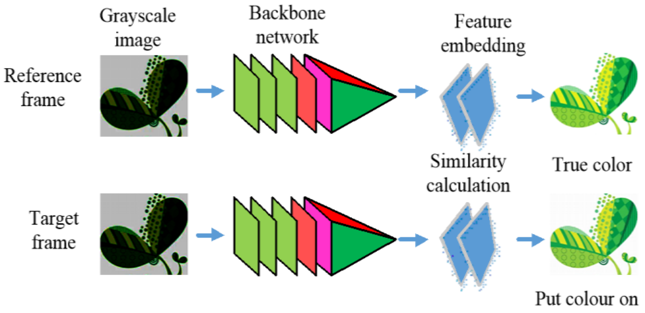


Figure 1. Target frame reconstruction process.

The specific principle is to transfer the color of the previous frame of the original image to the next frame according to the set rules, and use the stability of the color of the next frame as an unsupervised model signal to optimize the model. Define the reference frame as I_{t-1}, the target as I_t, and input $\{I_{t-1}, I_t\}$ into the model. The model embeds image features through backbone network, and embeds reference frame features as Equation (1).

$$y_{t-1} = R_{t-1}(I_{t-1};\theta_{t-1}) \tag{1}$$

In Equation (1), y_{t-1} is the backbone network reference frame output, and θ_{t-1} is the backbone network parameter. The target frame feature embedding is Equation (2) [1].

$$y_t = R_t(I_t;\theta_t) \tag{2}$$

In Equation (2), y_t refers to the target backbone network reference frame output, and θ_t refers to the target backbone network parameters. Target frame reconstruction involves calculating the similarity between the target frame and the reference frame at the corresponding position of the pixel, as Equation (3).

$$S_{t,t-1}^{i,j} = \frac{\exp(y_t^i, y_{t-1}^j)}{\sum_{p\in p1} exp(t_t^i, t_{t-1}^p)} \tag{3}$$

In Equation (3), y_t^i is the feature vector corresponding to the i-th pixel in the target frame. $P1$ is the set of adjacent target pixels corresponding to the reference frame. y_{t-1}^j is the feature vector corresponding to the jth pixel in the reference frame. $S_{t,t-1}$ represents the Matrix similarity between the reference frame and the target frame. Each row in $S_{t,t-1}$ corresponds to a target frame pixel, so its corresponding row is weighted and summed, and the sum result is used as the color value of the pixels in the reconstructed image, as Equation (4) [6].

$$\tilde{t}_t^i = \sum_{j\in p1} S_{t,t-1}^{i,j} V_{t-1}^j \tag{4}$$

In Equation (4), V_{t-1}^j is the color value of the jth adjacent pixel under the reference frame target pixel. In the study, the Huber loss function was used to optimize the guidance model, while Adam was used to optimize the model parameters. The results are Equation (5) [24].

$$Loss = \frac{1}{n}\sum_i H_i \tag{5}$$

In Equation (5), H_i represents the pixel loss of the i-th reconstructed target frame. The reconstruction of target frames represents weights through the similarity between adjacent and target pixels, and it is necessary to ensure the consistency of time and space in video processing, ensuring the stability of each frame is the key to construction [14]. In this regard, decomposition expression is introduced to improve the training effect of the model. The matching module will be divided into saliency and spatiotemporal modules [25]. The saliency module is responsible for enhancing the target of each frame, while the spatiotemporal module ensures spatiotemporal consistency. The purpose of decomposition expression is to learn different potential variables in a task, and the change of one variable does not affect other variables, but it will learn independently from each other [22]. The set of variables is Equation (6).

$$Z = \{Z_1, Z_2, \ldots, Z_n\} \tag{6}$$

Z_1, Z_2, …, Z_n in Equation (6) represents a potential learning variable, and the potential variable meets the requirements of Equation (7).

$$P(Z) = \prod_{k=1}^{n} P(Z_k) \tag{7}$$

In Equation (7), Z_k represents the potential learning variable. Meanwhile, different variables will match different visual cues, as Equation (8).

$$I = \sin(v_1, v_2, \cdots, v_n) \tag{8}$$

In Equation (8), sin() is a nonlinear mapping, and v_n is a visual cue of potential variable Z_k matching. Figure 2 shows the video segmentation process based on decomposition representation.

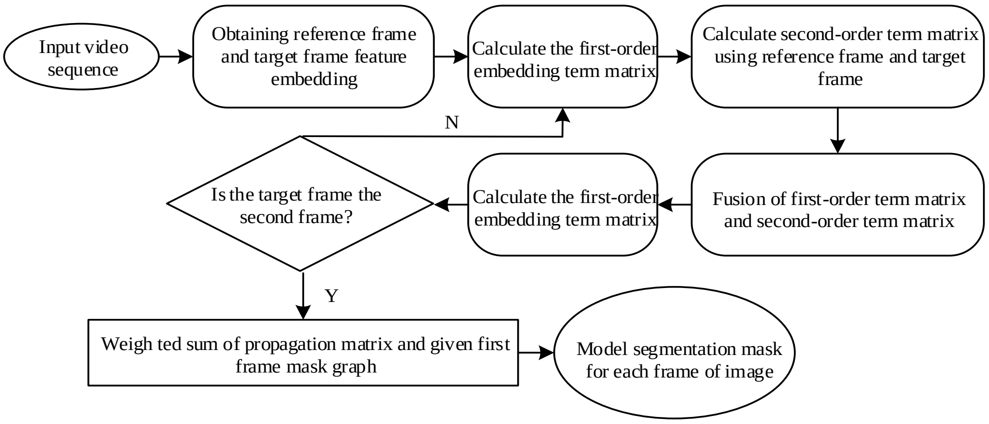

Input video sequence
Obtaining reference frame and target frame feature embedding
Calculate the first-order embedding term matrix
Calculate second-order term matrix using reference frame and target frame
N
Is the target frame the second frame?
Calculate the first-order embedding term matrix
Fusion of first-order term matrix and second-order term matrix
Y
Weigh ted sum of propagation matrix and given first frame mask graph
Model segmentation mask for each frame of image

Figure 2. Video segmentation process based on decomposition representation.

3.2. Constructing a Video Foreground Segmentation Model Considering Motion Scenes

For specific motion segmentation scenarios, the segmentation model based on decomposition representation cannot meet the real-time requirements, so a joint video segmentation strategy is adopted for optimization. In this regard, a real-time Foreground Segmentation Network based on single Linear Bottleneck and Pooling Compensation (FS-LBPC) segmentation method is proposed for motion scene segmentation. According to Lee *et al*. [14] research, the Linear Bottleneck and Pooling Compensation (LBPC) model utilizes an encoder decoder to achieve binary classification of video background and foreground pixels [5]. The encoder in the model will select the first four convolutional blocks of the VGG16 model for adjustment, and replace the convolution of the four convolutional blocks with a single linear bottleneck. Adopting the Tan *et al*. [26] framework, the decoder part consists of an activation function, double sampling, and convolution [13]. The model encoding and decoding structure is shown in Figure 3.

The model decoding is mainly performed by the pyramid pooling module for multi-scale feature learning, while upsampling processes the output information of pooling compensation on the previous decoding output, and finally outputs the class probability using the sigmoid function. Using a Single Linear Bottleneck Operator (SLBO) in the model to reduce the computational burden, reducing channel dimensions through convolution designed in the module, and subsequently recovering through convolution. ReLU function is used in the bottle neck sub layer recovery channel, and the helix is mapped to the dimensional space through ReLU function and Random matrix. The SLBO is represented as $\Phi(x)$, as Equation (9).

$$\Phi(x)=\left[A\cdot NoB\right]x \tag{9}$$

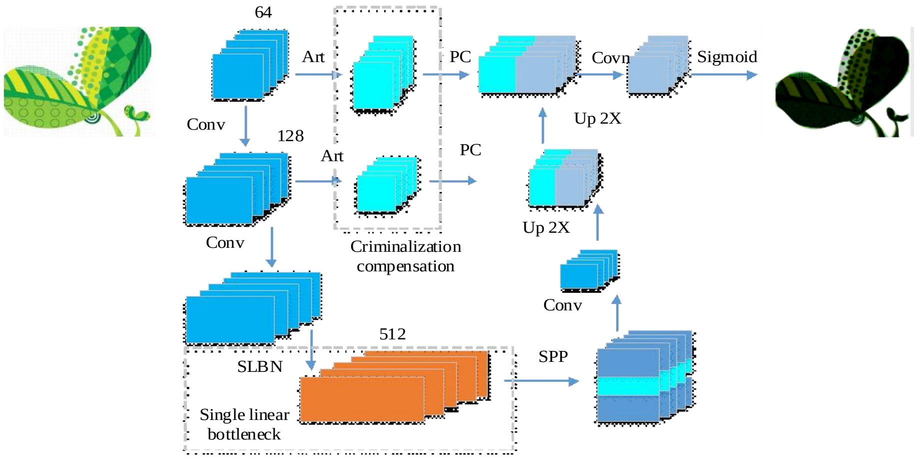
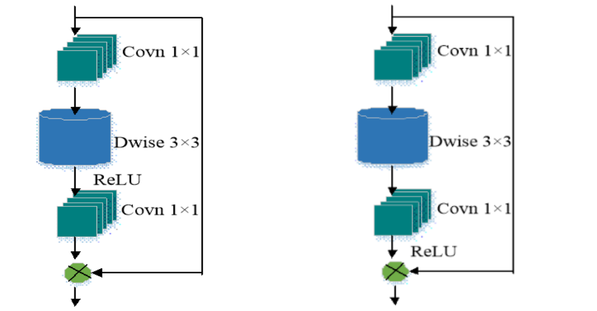


Figure 3. Structure diagram of model encoder and decoder.

In Equation (9), *A* and *B* are both standard convolutions. Among them, *A*·*N* is a linear transformation, *B* is a nonlinear transformation, and *N* is a separable convolution. In order to reduce the amount of convolution operations, a single bottleneck operator is introduced into the model, which aims to increase the depth and complexity of the network model while keeping the model parameters small, in order to better capture the feature information in the input data. The SLBO uses a module composed of four convolutions, as shown in Figure 4.



a) Residual bottleneck operator. b) Single line layer bottleneck operator.

Figure 4. Bottleneck operator structure diagram.

The residual bottleneck operator in Figure 4 has undergone linear processing on both ends, while SLBO only retains one end of the bottleneck as linear, improving the computational efficiency of the model. The research adopts pooling compensation to solve the problem of information loss caused by encoding decoding structure in compression processing, mainly applying skip link theory and attention mechanism to improve the processing of details. The model pooling operation can be understood as down sampling the data, expressed using *P*-norm as Equation (10).

$$\|A\| = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^{p}\right)^{\frac{1}{p}}, p >= 1 \tag{10}$$

In Equation (10), *m* and *n* represent the upper and lower limits of numerical values, *p* represents positive real numbers, and a_{ij} is the operation matrix. The research mainly compensates for lost information through compensation pooling while suppressing irrelevant information. The processed weighted feature map is Equation (11).

$$Y = \sigma(f^{k\times k}X)X \tag{11}$$

In Equation (11), σ represents the activation function, $f^{k\times k}$ represents the $k\times k$ convolution operation, and X represents the operated feature map. The detailed compensation for the entire pooling loss is Equation (12). Compensation pooling is used to compensate for lost information in the model and suppress irrelevant data. The details of the entire pooling loss compensation are shown in Figure 5.

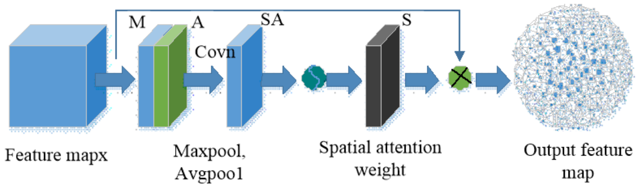


Figure 5. Schematic diagram of enhanced pooling loss details.

The pooling operation processes the feature maps before the maximum pooling layer in the encoder layer through an attention mechanism, while suppressing irrelevant features. Finally, the encoder and skip links weight the feature map to compensate for lost details.

3.3. Constructing a Video Segmentation Model Considering General Scenes

The previous section mainly discussed video segmentation techniques for moving scenes, but in actual segmentation, the target may be either moving or stationary. At the same time, the video scene also changes with the camera scene, and the segmentation requirements for general scenes are actually higher. In order to achieve the processing of general video scenes, a spatiotemporal Sequence Modeling and Similarity Learning (STS) model for video foreground segmentation based on spatiotemporal similarity is proposed. The structure of the STS model is shown in Figure 6.

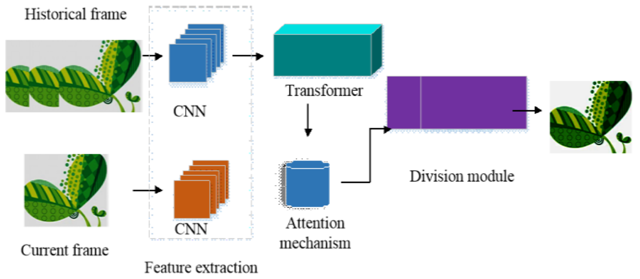


Figure 6. Schematic diagram of STS model structure.

The STS model mainly recognizes the appearance information of the target, without the use of optical flow calculation to achieve segmentation of the target. The model consists of twin encoders, decoders, segmentation modules, and similarity habits modules, with the core components being the similarity learning mechanism and transformer. The twin encoder is responsible for extracting features from image data, one for extracting historical frame data and the other for extracting current frame data, consisting of five convolutional modules. The Transformer module is mainly used for learning the features of historical frame backgrounds and foreground sequences, and provides reference for segmentation. The information generated by this module will undergo similar learning with the attention module to obtain more accurate segmentation feature maps, and the final binary classification results will be output by the segmentation module. The model uses Transformer to collect spatiotemporal information features of historical frames, and the main module will output corresponding values to the Transformer, responsible for predicting the sequence information of historical frame data. Define the decoder's last output as $z_0 \in \mathbb{R}^d$, which serves as a query to output historical frame object feature data. The multi attention output is Equation (12).

$$\hat{z}_{so} = \sum_{n=1}^{N} W_o^n (W_v^n z_0) \tag{12}$$

In Equation (12), W_o^n and W_v^n are both weight matrices. In the attention module, if any position feature in the Transformer encoder represents $y_{s,t} \in \mathbb{R}^d$, the output expression of the first cross attention module is Equation (13).

$$k_{s,t}^n = W_k^n y_{s,t} \tag{13}$$

In Equation (13), W_k^n represents the attention output weight corresponding to the first crossover module, and the output expression of the second crossover module is Equation (14).

$$v_{s,t}^n = W_v^n y_{s,t} \tag{14}$$

In Equation (14), W_k^n represents the attention output weight corresponding to the second crossover module. The final model decoder output is represented by $\hat{z}_\infty$, as Equation (15).

$$\hat{z}_\infty = \sum_{n=1}^{N} W_o^n \left[\sum_{t=1}^{T+1} \sum_{s=1}^{HW} \sigma\left(\frac{(W_q^n \hat{z}_{so})^T k_{s.t}^n}{\sqrt{d_n}}\right) \cdot v_{s,t}^n \right] \tag{15}$$

In Equation (15), W_q^n represents the corresponding weight output by the encoder, while H and W represent the height and width of the feature map, respectively. The transformer encoder is an important component in the transformer model, responsible for encoding and feature extraction of input sequences. The structure of the Transformer module is shown in Figure 7.

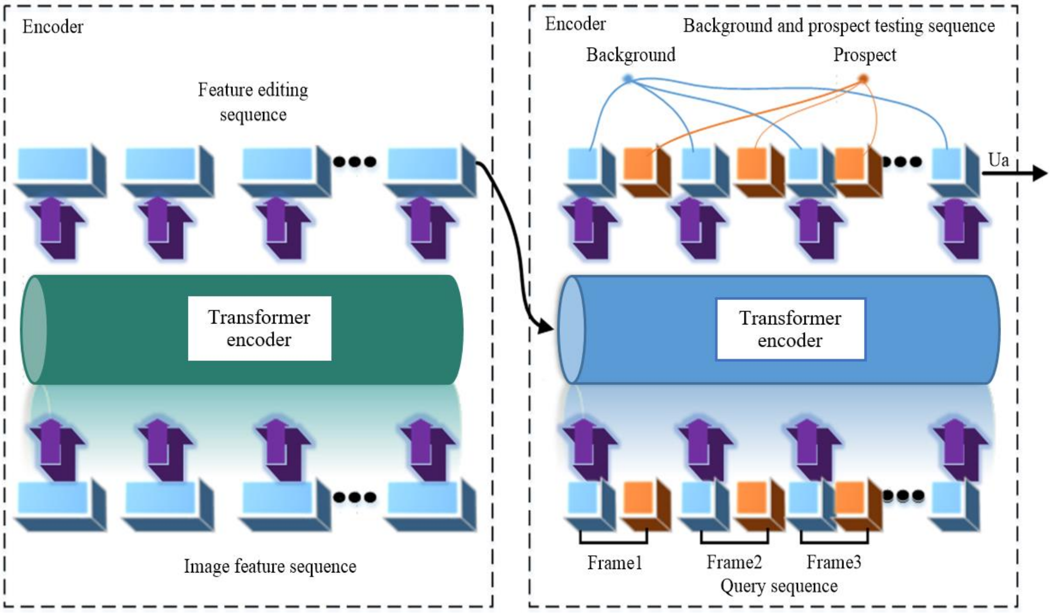


Figure 7. Schematic diagram of Transformer module structure.

The similarity learning between the output historical frame and the current frame results is mainly achieved by transforming the similarity matrix of the historical frame to obtain foreground features similar to the current frame. Finally, the final segmented feature map is obtained by fusing features into the segmentation module. Define $U_a \in \mathbb{R}^{W \times H \times C}$ as the historical frame map vector, C as the current frame map vector, $U_b \in \mathbb{R}^{W \times H \times C}$ as the number of feature map channels, and the affine matrix as S. In the affine matrix, each element represents a similarity between U_b^T and the historical frame U_a, as expressed in Equation (16).

$$S = U_b^T W U_a \in \mathbb{R}^{(WH) \times (WH)} \tag{16}$$

Assuming that the weight matrix W can represent a diagonal matrix, the diagonalization is Equation (17).

$$W = P^{-1} D P \tag{17}$$

In Equation (17), D represents a Diagonal matrix, and P represents a variable matrix. By changing the matrix, U_a and U_b are projected onto an orthogonal space. The calculation of the similarity distance between the two can remove the influence of ablation channel and reduce the feature Data redundancy. Finally, the affine output weight probability is calculated using the Softmax function, and the feature maps are concatenated according to the calculation dimension, which is input into the segmentation module to complete the target segmentation. The entire VOS process based on decomposition expression is shown in Figure 8.

The research proposes a target video segmentation technique based on decomposition expression, which uses decomposition expression to calculate the target frame and reference frame of the video sequence, and obtains more accurate video reconstruction target data through similarity calculation. At the same time, considering both motion and general scenes, FS-LBPC and STS models were used to process the foreground information of motion and general scenes respectively. Finally, the segmentation model was used to achieve segmentation processing of the target video.

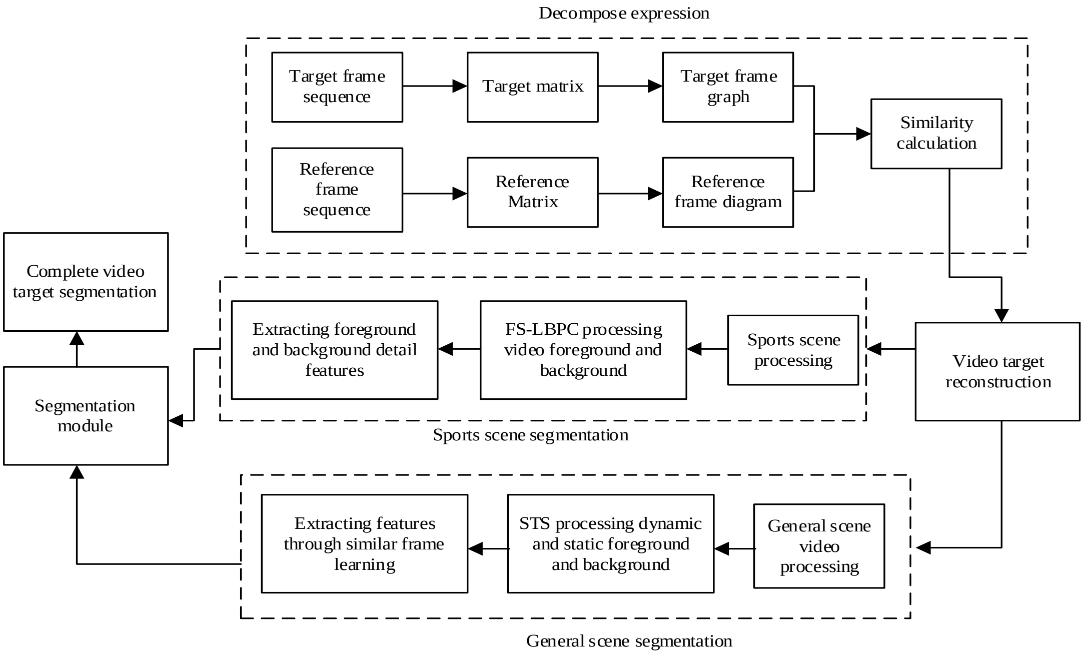

Decompose expression
Target frame sequence
Target matrix
Target frame graph
Reference frame sequence
Reference Matrix
Reference frame diagram
Similarity calculation
Complete video target segmentation
Extracting foreground and background detail features
FS-LBPC processing video foreground and background
Sports scene processing
Video target reconstruction
Segmentation module
Sports scene segmentation
Extracting features through similar frame learning
STS processing dynamic and static foreground and background
General scene video processing
General scene segmentation

Figure 8. VOS based on decomposition representation.

4. Algorithm Model Simulation Testing

This section will conduct experimental analysis from two aspects: sports scenes and general scenes, in order to discuss the application effect of the proposed technology in VOS scenes. The testing content includes comparison of segmentation stability, segmentation loss, segmentation accuracy, etc., in order to evaluate the comprehensive application effect of the model.

4.1. Video Foreground Segmentation Testing Considering Motion Scenes

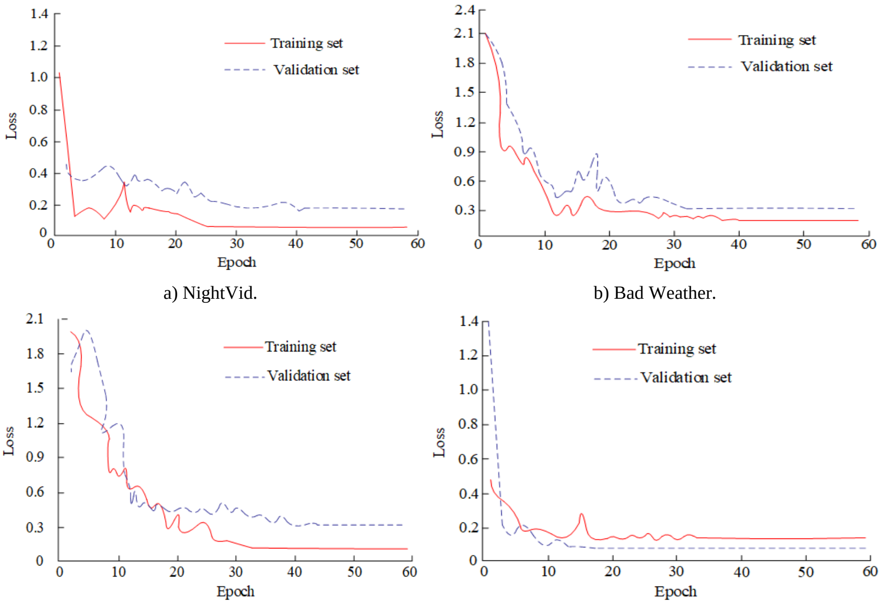
To verify the performance of the proposed model, experimental performance tests will be conducted on the universal dataset Change Detection Benchmark Dataset 2014 (CDNet.20I4SM) and University of California, San Diego (UCSD). The CDNet.20I4SM dataset includes scenes such as adverse weather, camera shake, low frame rate, heat and turbulence, and shadows. The UCSD dataset mainly consists of 18 video sequences with truth labels, including dynamic backgrounds, which test the training performance of the model [12]. The testing system platform is WINDOWS 10, the graphics card is NVIDIA RTX2070, the running memory is 32GB, and the processor is Intel I7 16 core. To ensure the accuracy of the experiment and avoid errors, all tests are completed in a unified software and hardware environment, and simulation testing is completed on the Matlab experimental platform. The Foreground Image Segmentation (FgSegNet_) model, the Convolutional Networks for Biomedical Image Segmentation (MU Net) model, and the Cascade Convolutional Neural Network (Cascade CNN) model are used as benchmark models. The FgSegNet model is a deep learning-based foreground segmentation network designed specifically for real-time video analysis. It utilizes convolution for multi-scale feature encoding to improve the accuracy and efficiency of foreground segmentation. The MU Net model is a CNN designed specifically for biomedical image segmentation, which is used in medical image processing fields such as cell segmentation and neural structure recognition. Cascade CNN is a facial image detection model that utilizes multiple simple networks cascaded into a strong classifier, which is particularly suitable for image task processing in complex backgrounds [31]. The experiment selects the optimal model training j basic parameters through multiple trainings, and the basic parameters for model training are shown in Table 1.

Table 1. Basic parameters for model training.

Model training configuration parameters	Numerical value
Iterations	100
Optimizer rho value	0.9
Optimizer epsilon	le-8
Initial Learning rate	1e-4

In addition, in the study, the Encoder module was selected as Imagenet, and the Optimizer was selected as MSProp. Select night scenes, adverse weather scenes, camera scenes, and turbulent weather scenes from the CDNet.20I4SM dataset to test the FS-LBPC model. The evaluation metrics include loss, precision, F-measure (F), Percentage of Wrong Classification (PWC), Frame rate Per Second (FPS), and segmentation accuracy.

The loss measures the difference between the predicted and true values of the model, which is used to guide model training; Precision evaluates the accuracy of prediction results and calculates the proportion of samples that are actually positive in the predicted positive category; the F-value is the harmonic average of precision and recall, used to comprehensively evaluate the performance of the model. When the F-value is high, the precision and recall of the model are relatively good; PWC measures the proportion of misclassified samples in the total sample, which is an intuitive indicator for evaluating the performance of classification models; FPS is the frame rate per second, which measures the smoothness of a video. The better the value, the better; the segmentation accuracy represents the proportion of correctly segmented pixels to the total number of pixels, evaluating segmentation performance. The higher the value, the better the segmentation effect [26]. In the study, four scenarios were selected from the CDNet.2014SM dataset for training loss testing, including NightVid, Bad Weather, CameraJit, and Turbulence. The segmentation loss is shown in Figure 9.



a) NightVid.

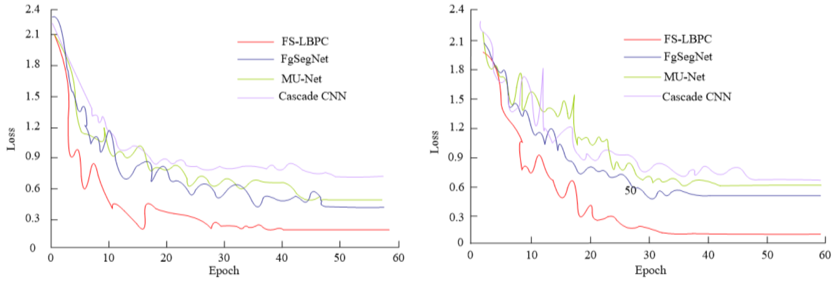
b) Bad Weather.

c) CamiraJit.

d) Turbulence.

Figure 9. Comparison of training losses based on FS-LBPC model.

Figure 9-a), (b), (c), and (d) show the training losses of FS-LBPC under night, adverse weather, camera, and turbulent weather scenarios, respectively. According to the training effects in different scenarios, as the number of iterations increases, the Loss values of the training set and the testing set both decrease continuously, and at around 40 iterations, the training set and the validation set gradually converge. In the comparison of loss in different scenarios, the loss value at night convergence is 0.089, and the validation set is 0.268. According to its training results, the requirements for model training vary in different scenarios. In adverse weather scenarios, the loss of model training set is significant, while in camera scenarios, the loss of model validation set is significant. From the test results, it can be seen that adverse weather scenarios are more complex and test the segmentation performance of the technology. For this, select adverse weather scenarios and camera scenarios for further training. Figure 10 shows the training loss situation of multiple models.

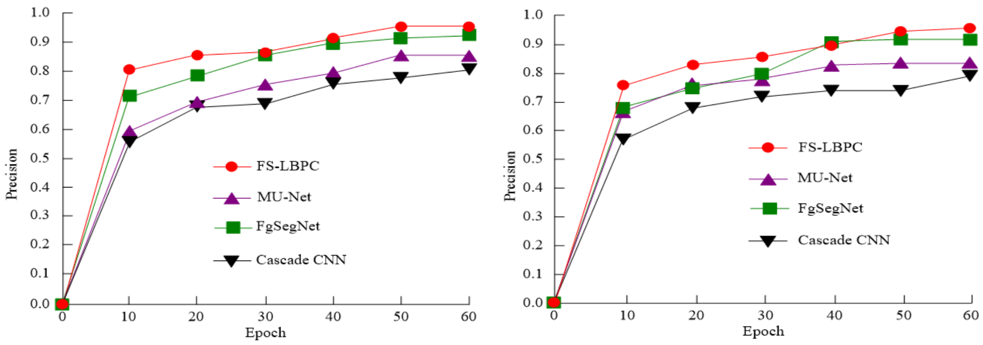


a) Bad Weather.

b) CamiraJit.

Figure 10. Multi model training loss situation.

Figure10-a) and (b) show the training results for adverse weather scenarios and camera scenario loss, respectively. In the training of adverse weather scenarios, there are significant differences in the training results of different segmentation models. The best performing model is FS-LBPC, which tends to converge after 40 iterations, and at this point, the Loss value is 0.276. Next in performance is FgSegNet, which tends to converge after 47 iterations, with a Loss value of 0.586 at this point. The worst performing is Cascade CNN, which tends to converge after 48 iterations, with a Loss value of 0.903 at this point. The best performing camera scene is FS-LBPC, which tends to converge after 32 iterations with a Loss value of 0.216. The worst case is Cascade CNN, which converges with a Loss value of 0.903 after 50 iterations. Select the Precision (P) indicator to evaluate the video segmentation effect of the model, as displayed in Figure 11.

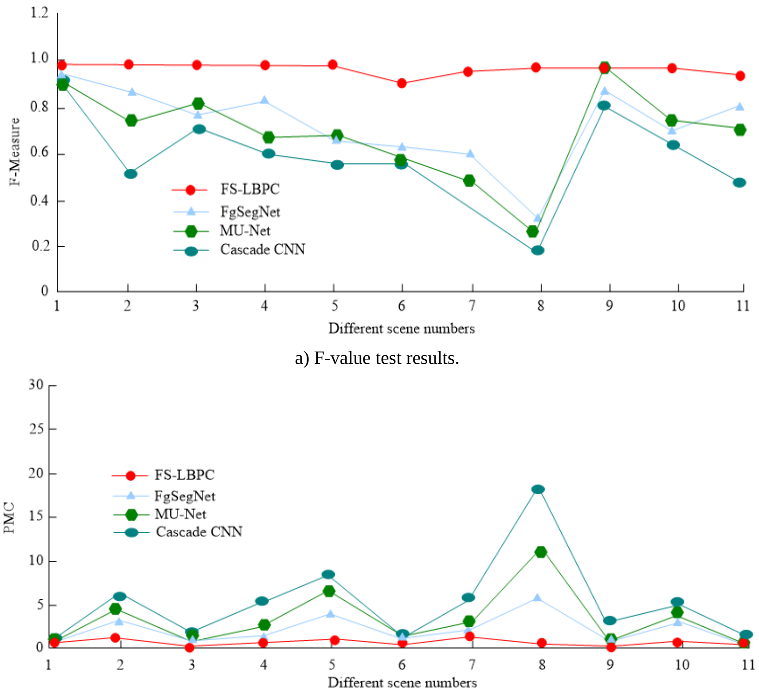


a) Bad Weather.

b) CamiraJit.

Figure 11. Test results of video segmentation precision for each model.

It is still selected adverse weather scenes and camera scenes for segmentation training, as shown in Figure 11-a) and (b), respectively. In adverse weather scenarios, the four models have different segmentation effects on videos, with the best performing being FS-LBPC, which tends to converge after 50 iterations with a P of 0.963; The *P* values for FgSegNet, MU Net, and Cascade CNN were 0.936, 0.863, and 0.806, respectively. In the camera scenario, similar to the test results in adverse weather scenarios, the P values for FS-LBPC, FgSegNet, MU Net, and Cascade CNN were 0.968, 0.916, 0.846, and 0.796. According to the test results, the FS-LBPC model can achieve the best Precision performance in adverse weather and camera scenarios, indicating its excellent appearance in different scenarios. Figure 12 shows the stability test results of video segmentation for multiple models.



a) F-value test results.

b) Percent of false classification.

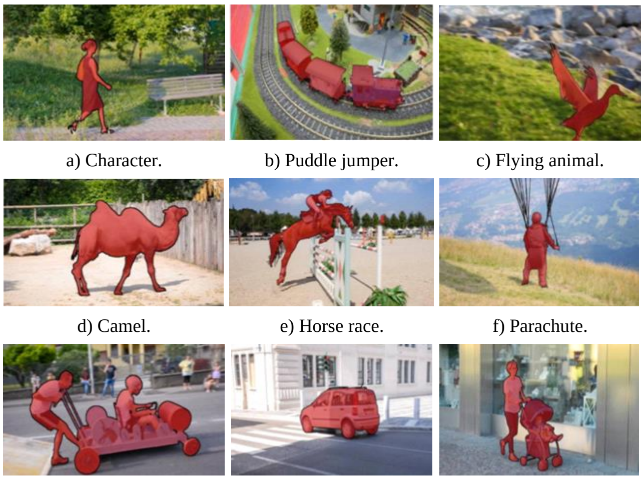
Figure 12. Model video segmentation stability test.

Figure 12-a) shows the F-value test results of multiple models. The experiment selected 11 types of video scenes from the CDNet2014 dataset and UCSD data for segmentation, and tested the segmentation stability of different models in different video scenes. The higher the F value, the better the video segmentation effect of the model. According to the curve results, FS-LBPC has high segmentation performance in scenes 1-11, with F values stable around 1.0 and above 0.9. The second-best performing model is the FgSegNet model, but there is significant fluctuation in Scenario 8, with an F1 value of 0.46. Overall comparison, FS-LBPC has good segmentation performance and stability. Figure 12-b) is the PWC value result of the model video segmentation classification error percentage. The higher the PWC value, the higher the accuracy of the model in image segmentation and the better the video segmentation effect. According to the results of the Figure 12-b) curve, the most stable video image segmentation is FS-LBPC. In 1-11 video segmentation scenarios, the model's misclassification percentage PWC value is less than 0.9, while FgSegNet, MU-Net, and Cascade CNN have significant stable fluctuations in Scenario 8. The PWC values of the three models in Scenario 8 are 5.3, 10.6, and 19.6, respectively. According to the test results, 11 scenarios were selected for comparison, and the F-value and PWC value of different technologies were compared. The FS-LBPC model performed better than the comparison technology in different scenarios, indicating its high video segmentation performance and stability.

4.2. Video Segmentation Testing Considering General Scenarios

Generally, the DAVIS 16 and the FBMS are chosen to verify the video segmentation effect of the model for video segmentation [3]. The DAVIS16 dataset is a widely used dataset in the field of video segmentation, which includes 50 types of video materials. The types of videos cover commonly seen blurred motion scenes, occluded motion scenes, and constantly changing appearance scenes in general videos. In addition, each scene in the dataset is labeled at the pixel level. Some scene graphs are exhibited in Figure 13.

STS training adopts random gradient descent optimization for model training. The initial Learning rate of the model is 2.5e-4, the maximum number of iterations of the model is set to 100, the batch size is set to 4, and the regularization coefficient is set to 1e-4, and the number of iterations is 60. The first step is to test the loss performance of STS, selecting the Filter Summary Net (FSNet) model and the CO-attention Siamese network (COSnet) model as benchmark models to verify the training effectiveness of the model. Among them, the FSNe model is a parameter sharing convolutional layer representation method called Filter Summary (FS), which aims to compress convolutional kernel parameters through one-dimensional representation and has good applications in the field of image processing. The COSnet model is an innovative twin neural network model that solves UVOS tasks through a shared attention mechanism and has good segmentation capabilities [17]. The results are shown in Figure 14.



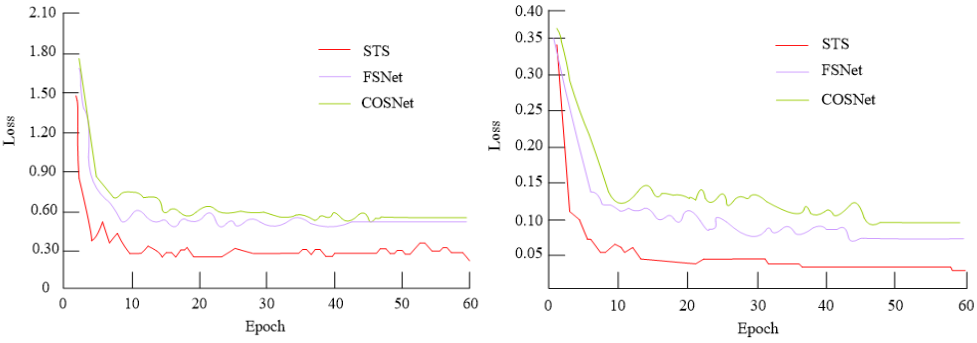
a) Character. b) Puddle jumper. c) Flying animal.

d) Camel. e) Horse race. f) Parachute.

g) Play. h) Car. i) Stroller.

Figure 13. Scene graph of DAVIS16 dataset.

Figure 14 shows the training loss results of different scene models. Among them, Figure 14-a) represents the model video segmentation loss result for the car scene. The best performing STS gradually converges after 25 iterations, with a Loss value of 0.320. The second best performance is FSNet, which gradually converges after 40 iterations, with a Loss value of 0.605. The worst performing COSnet gradually converges after 45 iterations, with a Loss value of 0.756. The data results indicate that STS has excellent segmentation performance and the overall loss is the lowest in video segmentation for car sports events. Figure 14-b) is the model video segmentation loss result for the character scene. Compared to complex car motion scenes, the loss of each model in character scenes is lower. The best performing loss performance is STS, followed by FSNet, and the worst is COSnet. When the three models tend to converge, the Loss values are 0.036, 0.092, and 0.146, respectively. Therefore, STS has better video segmentation performance. Table 2 shows the test results of video segmentation accuracy for different models.



a) Car. b) Character.

Figure 14. Training loss results of different scene models.

Table 2. Video segmentation effects of different models.

Evaluation model	Video segmentation scene	Segmentation accuracy	Split frames (fps）	Split F-value
STS model	Car scene	0.963	52	0.855
	Character scenes	0.975	51	0.864
	Parachute scene	0.986	53	0.868
	Camel scene	0.995	54	0.896
	Horse racing scene	0.976	50	0.894
	Flying Animal Scenes	0.965	51	0.886
	Play scenes	0.983	52	0.846
FSNet model	Car scene	0.923	47	0.802
	Character scenes	0.926	46	0.796
	Parachute scene	0.934	47	0.756
	Camel scene	0.905	45	0.756
	Horse racing scene	0.924	46	0.735
	Flying Animal Scenes	0.903	47	0.756
	Play scenes	0.914	50	0.765
COSnet model	Car scene	0.856	35	0.653
	Character scenes	0.846	32	0.634
	Parachute scene	0.843	31	0.681
	Camel scene	0.863	35	0.646
	Horse racing scene	0.825	30	0.674
	Flying Animal Scenes	0.806	30	0.675
	Play scenes	0.843	32	0.694

Table 2 selected 8 classic scenes from the DAVIS16 dataset for video segmentation training, including car movement, characters, parachute descent, camel walking, horse racing, animal flight, and character play. It evaluates the video segmentation performance of each model using segmentation accuracy and the number of video segmentation frames. From the data in Table 2, there are significant differences in the effectiveness of model video segmentation in different scenarios. High speed motion scenes, such as horse racing, animal flying scenes, test the comprehensive segmentation ability of the model. The segmentation accuracy and the number of segmentation frames of the model in high-speed motion scenes will be affected. However, STS has the best comprehensive performance. In the horse racing and animal flight scenes, the segmentation accuracy of the model is 0.976 and 0.965 respectively; the segmentation accuracy of FSNet and COSnet in both scenarios is 0.924, 0.903, and 0.825, 0.806, respectively. At the same time, compare the number of split frames and F value in the horse racing and animal flying scenes: in horse racing, the frame rate values of STS, FSNet and COSnet are 50fps, 46fps and 30fps respectively, and the F values are 0.894, 0.735 and 0.674 respectively. In the animal flight scene, the frame rate values of STS, FSNet and COSnet are 51 fps, 47 fps and 30 fps respectively, and the F values are 0.886, 0.756 and 0.675 respectively. Therefore, STS has excellent video segmentation performance. Figure 15 shows the segmentation effect of the STS model.

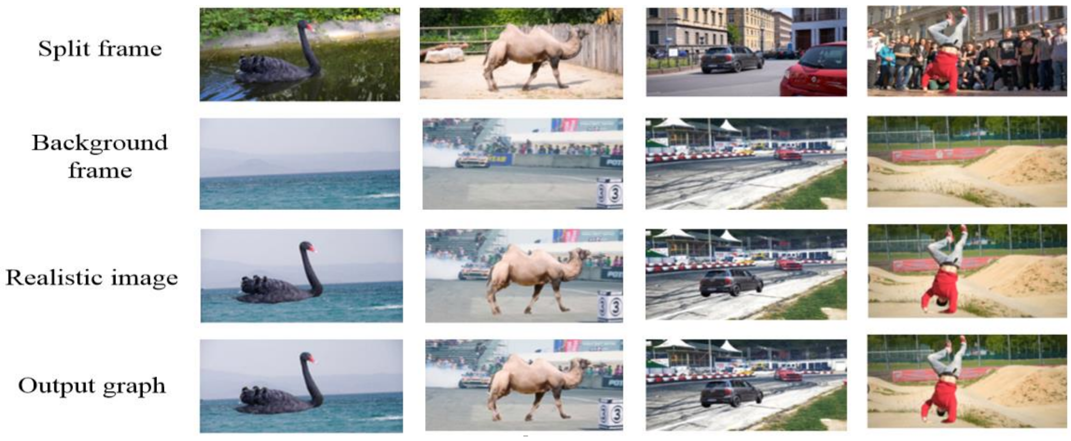


Figure 15. STS model video segmentation results.

Figure 15 shows four sets of video segmentation results, including segmented frame images, background frame images, and real images. From the results, the proposed model can effectively segment video scenes.

5. Discussion

VOS is one of the important tasks in the field of computer vision, which aims to process video elements. It has a wide range of applications in various fields, such as video surveillance, intelligent transportation, medical image analysis, etc. For VOS tasks, researchers have proposed various methods and models to address this issue. An unsupervised decomposition expression VOS technique is proposed in the study, which considers complex segmentation scenarios. Compared with traditional segmentation techniques such as unsupervised and supervised VOS techniques, this technique has higher segmentation accuracy and stronger adaptability [20].

Apply the proposed technology to specific experimental scenarios and select night scenes, adverse weather scenes, camera scenes, and turbulent weather scenes from the CDNet.2014SM dataset for experimental analysis. In sports scene analysis, FGSegNet model, MU Net model, and Cascade CNN model are selected as benchmark models. The proposed FS-LBPC model performs best in adverse weather and camera scenes, with accuracy rates of 0.963 and 0.968, respectively. However, the accuracy of other models is relatively low, with the worst being the Cascade CNN model, which has an accuracy of only 0.806. This indicates that the proposed FS-LBPC model has excellent performance in video segmentation. In addition, the researchers also tested the video segmentation stability of multiple models. According to the results of F-value and classification error percentage index, the proposed FS-LBPC model has high segmentation performance and stability in different video scenes. However, other models exhibit significant fluctuations in certain scenarios.

In addition, considering video segmentation testing in general scenarios, the DAVIS16 dataset and FBMS dataset were selected to verify the performance of the model. Based on the evaluation of training loss results and segmentation accuracy, the proposed STS model performs well in different scenarios, with lower loss and higher segmentation accuracy. Especially in high-speed motion scenarios, the STS model performs better than other models.

In summary, according to the research results, the proposed FS-LBPC model and STS model perform well in video segmentation tasks, with high segmentation accuracy and stability. These models have broad application prospects and can help improve the efficiency and accuracy of video segmentation, which is of great significance for research and application in related fields.

6. Conclusions

In the field of intelligent vision, VOS is one of the key contents of machine vision research. Traditional VOS focuses on supervised learning and relies on a large number of tag data. In order to overcome the problem of traditional supervised learning, this paper proposes a video segmentation method based on decomposition representation, which completes the recognition of spatio-temporal information and significant data through decomposition module. Considering the video segmentation of motion scenes and general scenes, the former proposes a joint pooling compensation technique and a single linear bottleneck video foreground segmentation technique, while the latter uses spatiotemporal similarity features to segment general video scenes by calculating similarity. The accuracy of model segmentation in adverse weather scenarios was tested using motion foreground segmentation. The precision P-values of FS-LBPC, FgSegNet, MU-Net, and Cascade CNN were 0.968, 0.916, 0.846, and 0.796, respectively. In general video scene segmentation testing, the training loss performance of the model is tested in the car scene: the proposed STS model performs best, gradually converging after 25 iterations, with a loss value of 0.320. In the test of video segmentation accuracy of different models, STS performs best in horse racing and animal flight scenes, with segmentation accuracy of 0.976 and 0.965 respectively; the segmentation accuracy of FSNet and COSnet in both scenarios is 0.924, 0.903, and 0.825, 0.806, respectively. Given this, the proposed video segmentation model has excellent segmentation performance. However, there are also shortcomings in the research. The unsupervised model designed only performs lightweight processing on some scenes, and in the future, some scenes can be compressed and pruned to improve the segmentation effect of the model.

References

[1] Ahmad J., Muhammad K., Lloret J., and Baik S., “Efficient Conversion of Deep Features to Compact Binary Codes Using Fourier Decomposition for Multimedia Big Data,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3205-3215, 2018. DOI:10.1109/TII.2018.2800163

[2] Ammar S., Bouwmans T., Zaghden N., and Neji M., “Deep Detector Classifier (DeepDC) for Moving Objects Segmentation and Classification in Video Surveillance,” *IET Image Process*, vol. 14, no. 8, pp. 1490-1501, 2020. https://doi.org/10.1049/iet-ipr.2019.0769

[3] Bian J., Zhan H., Wang N., Li Z., Zhang L., Shen C., Chen M., and Reid I., “Unsupervised Scale-Consistent Depth Learning from Video,” *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2548-2564, 2021. https://link.springer.com/article/10.1007/s11263-021-01484-6

[4] Chan S., Huang C., Bai C., Ding W., and Chen S., “Res2-UNeXt: A Novel Deep Learning Framework for Few-Shot Cell Image Segmentation,” *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 13275-13288, 2022. https://link.springer.com/article/10.1007/s11042-021-10536-5

[5] Das P., Karaoglu S., and Gevers T., “Intrinsic Image Decomposition Using Physics-Based Cues and CNNs,” *Computer Vision and Image Understanding*, vol. 223, pp. 103538, 2022. https://doi.org/10.1016/j.cviu.2022.103538

[6] Deepak K., Chandrakala S., and Mohan C., "Residual Spatiotemporal Autoencoder for Unsupervised Video Anomaly Detection," *Signal Image Video Process*, vol. 15, no. 1, pp. 215-222, 2021. https://link.springer.com/article/10.1007/s11760-020-01740-1

[7] Falaschetti L., Manoni L., and Turchetti C., "A Low-Rank CNN Architecture for Real-Time Semantic Segmentation in Visual SLAM Applications," *IEEE Open Journal of Circuits and Systems*, vol. 3, pp. 115-133, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9773325

[8] Fan J., Liu B., Zhang K., and Liu Q., "Semi-Supervised Video Object Segmentation Via Learning Object-Aware Global-Local Correspondence," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8153-8164, 2022. DOI:10.1109/TCSVT.2021.3098118

[9] Fu Y., Yang L., Liu D., Huang T., and Shi H., "Compfeat: Comprehensive Feature Aggregation for Video Instance Segmentation," *in Proceedings of the 35th AAAI Conference on Artificial Intelligence*, Virtual, pp. 1361-1369, 2021. https://doi.org/10.1609/aaai.v35i2.16225

[10] Giraldo J., Javed S., and Bouwmans T., "Graph Moving Object Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2485-2503, 2020. DOI:10.1109/TPAMI.2020.3042093

[11] Huang P., Han J., Liu N., Ren J., and Zhang D., "Scribble-Supervised Video Object Segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 2, pp. 339-353, 2022. DOI:10.1109/JAS.2021.1004210

[12] Hussein M., Puyal J., and Lines D., Sehgal V., Toth D., Ahmad O., Kader R., Everson M., Lipman G., Fernandez-Sordo J., Ragunath K., Esteban J., Bisschops R., Banks M., Haefner M., Mountney P., Stoyanov D., Lovat L., and Haidry R., "A New Artificial Intelligence System Successfully Detects and Localises Early Neoplasia in Barrett's Esophagus by Using Convolutional Neural Networks," *United European Gastroenterology Journal*, vol. 10, no. 6, pp. 528-537, 2022. https://onlinelibrary.wiley.com/doi/epdf/10.1002/ueg2.12233

[13] Khan R., Kifayat Ullah., Pamucar D., and Bari M., "Performance Measure Using a Multi-Attribute Decision-Making Approach Based on Complex T-Spherical Fuzzy Power Aggregation Operators," *Journal of Computational and Cognitive Engineering*, vol. 1, no. 3, pp. 138-146, 2022. https://doi.org/10.47852/bonviewJCCE696205514

[14] Lee Y., Seong H., and Kim E., "Iteratively Selecting an Easy Reference Frame Makes Unsupervised Video Object Segmentation Easier," *in Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Vancouver, pp. 1245-1253, 2022. https://doi.org/10.1609/aaai.v36i2.20011

[15] Li D., Li R., Wang L., Wang Y., Qi J., Zhang L., Liu T., Xu Q., and Lu H. C., "You Only Infer Once: Cross-Modal Meta-Transfer for Referring Video Object Segmentation," *in Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Vancouver, pp. 1297-1305, 2022. https://doi.org/10.1609/aaai.v36i2.20017

[16] Lin F., Xie H., Liu C., and Zhang Y., "Bilateral Temporal Re-Aggregation for Weakly-Supervised Video Object Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4498-4512, 2022. DOI:10.1109/TCSVT.2021.3127562

[17] Liu W., Lin G., Zhang T., and Liu Z., "Guided Co-Segmentation Network for Fast Video Object Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1607-1617, 2020. DOI:10.1109/TCSVT.2020.3010293

[18] Logeshwaran J., Kiruthiga T., Aravindarajan V., and Ravi S., "SVPA-the Segmentation Based Visual Processing Algorithm (SVPA) for Illustration Enhancements in Digital Video Processing (DVP)," *ICTACT Journal on Image and Video Processing*, vol. 12, no. 3, pp. 2669-2673, 2022. DOI:10.21917/ijivp.2022.0379

[19] Lu X., Wang W., Shen J., Crandall D., and Luo J., "Zero-Shot Video Object Segmentation with Co-Attention Siamese Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2228-2242, 2020. DOI:10.1109/TPAMI.2020.3040258

[20] Luo H., Sun B., Zhou H., and Cao W., "Image Segmentation with Multi-feature Fusion in Compressed Domain based on Region-Based Graph," *The International Arab Journal of Information Technology*, vol. 20, no. 2, pp. 159-169, 2023. https://doi.org/10.34028/iajit/20/2/2

[21] Pu S., Zhao W., Chen W., Yang S., Xie D., and Pan Y., "Unsupervised Object Detection with Scene-Adaptive Concept Learning," *Frontiers of Information Technology and Electronic Engineering*, vol. 22, no. 5, pp. 638-651, 2020. https://link.springer.com/article/10.1631/FITEE.2000567

[22] Qi J., Gao Y., Hu Y., Wang X., Liu X., Bai X., Belongie S., Yuille A., Torr P., and Bai S., "Occluded Video Instance Segmentation: A Benchmark," *International Journal of Computer Vision*, vol. 130, no. 8, pp. 2022-2039, 2022.

https://link.springer.com/article/10.1007/s11263-022-01629-1

[23] Raman N., Wahab A., and Chandrasekaran S., "Computation of Workflow Scheduling Using Backpropagation Neural Network in Cloud Computing: A Virtual Machine Placement Approach," *The Journal of Supercomputing*, vol. 77, no. 9, pp. 9454-9473, 2021. https://link.springer.com/article/10.1007/s11227-021-03648-0

[24] Shahrezaei I. and Kim H., "Fractal Analysis and Texture Classification of High-Frequency Multiplicative Noise in SAR Sea-Ice Images Based on a Transform-Domain Image Decomposition Method," *IEEE Access*, vol. 8, pp. 40198-40223, 2020. DOI:10.1109/ACCESS.2020.2976815

[25] Shakeel N. and Shakeel S, "Context-Free Word Importance Scores for Attacking Neural Networks," *Journal of Computational and Cognitive Engineering*, vol. 1, no. 4, pp. 187-192, 2022. https://doi.org/10.47852/bonviewJCCE2202406

[26] Tan Z., Liu B., Chu Q., Zhong H., Wu Y., Li W., and Yu N., "Real Time Video Object Segmentation in Compressed Domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 175-188, 2021. DOI:10.1109/TCSVT.2020.2971641

[27] Vecchio G., Palazzo S., Giordano D., Rundo F., and Spampinato C., "MASK-RL: Multiagent Video Object Segmentation Framework through Reinforcement Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5103-5115, 2020. DOI:10.1109/TNNLS.2019.2963282

[28] Vinayaraj P., Sugimoto R., Nakamura R., and Yamaguchi Y., "Transfer Learning with CNNs for Segmentation of PALSAR-2 Power Decomposition Components," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, no. 5, pp. 6352-6361, 2020. DOI:10.1109/JSTARS.2020.3031020

[29] Wang W., Shen J., Lu X., Hoi S., and Ling H., "Paying Attention to Video Object Pattern Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2413-2428, 2021. DOI:10.1109/TPAMI.2020.2966453

[30] Zheng M., Huang Y., Chen Q., and Liu Y., "Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining," *in Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Vancouver, pp. 3517-3525, 2022. https://doi.org/10.1609/aaai.v36i3.20263

[31] Zhou T., Porikli F., Crandall D., Gool L., and Wang W., "A Survey on Deep Learning Technique for Video Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, no. 6, pp. 7099-7122, 2023. DOI:10.1109/TPAMI.2022.3225573

[32] Zhou T., Wang S., Zhou Y., Yao Y., Li J., and Shao L., "Motion-Attentive Transition for Zero-Shot Video Object Segmentation," *in Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, pp. 13066-13073, 2020. https://doi.org/10.1609/aaai.v34i07.7008

[33] Zhou Y., Xu X., Shen F., Zhu X., and Shen H., "Flow-Edge Guided Unsupervised Video Object Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8116-8127, 2022. DOI:10.1109/TCSVT.2021.3057872

[34] Zhu W., Li J., Lu J., and Zhou J., "Separable Structure Modeling for Semi-Supervised Video Object Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 330-344, 2021. DOI:10.1109/TCSVT.2021.3060015

Jianfu Kong male, born in July 1986 in Zhoukou city, Henan Province, Han Dynasty, Bachelor's Degree in Animation in 2009, Research Direction: 3D Animation. Work experience: From August 2009 to August 2019, he worked with Zhoukou Vocational College of Science and Technology, teaching in the Department of Arts, as a secretary. From 2019 to now, he has worked in Henan Vocational University of Science and Technology, as a League member, as a party secretary. Academic situation: Research on the Methods of Constructing Harmonious Campus and Gratitude Education in Vocational Colleges, Zkyjky [2013]-015. Research on the Role of Youth in the Construction of the Innovation Zone for the Inheritance of Chinese Historical Civilization, QSNYJ2013225, Second Prize, Henan Provincial Social Science Federation, Communist Youth League, Henan Provincial Committee. Research on the Influence of Confucianism on Zhongyuan culture, SKL-2015-1076, Henan Federation of Social Sciences. Reform Ideas for the Integration of Industry and Education between Private Universities and Cultural Industry Bases under the New Situation, HNMXJ2020123, Henan Private Education Association. New Reflections on the Development of Vocational Education, "Literary Education, March 2014 Issue 2 on Dramatic Language and its Innovation," published by "A Hundred Prose Writers" in February 2014. Utility model patent: A script clip for animation production, patent number: ZL 2017 2 0210105.9, certificate number: 6529183.