IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 15, 2021, accepted September 3, 2021, date of publication September 10, 2021,
date of current version September 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3112014

Semi-Supervised Video Object Segmentation Based on Local and Global Consistency Learning

HUAGANG LIANG, LIHUA LIU, YING BO, AND CHAO ZUO
College of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China

Corresponding author: Huagang Liang (hgliang@chd.edu.cn)

This work was supported by the Key Research and Development Plan of Shaanxi Province of China under Grant 2019ZDLGY03-01.

ABSTRACT Due to the variety of video types and different quality on the Internet, it brings more challenges to video processing algorithms such as video object segmentation. Most existing video object segmentation methods rely on modules in other fields as an additional structure of the segmentation model. The combination of modules can improve the accuracy of the model, but it will also reduce the algorithm speed. This paper proposes a semi-supervised video object segmentation method based on local and global consistency learning, which does not rely on additional structures to achieve fast segmentation. First, we extract the embedding features of the image based on GhostNet which is the lightweight network. By using the embedded features of pixels, the graph model is established based on the similarity between pixels. Second, we adopt the local-global consistency learning framework to construct the label conduction model. Third, to optimize the memory occupation and inference speed of the model, we propose a sampling strategy for reference frames by considering local and global information. Finally, we establish a high-speed monitoring video dataset to verify the practical application effect of the method. Our method achieves a result of 69.5% $J\&F$ mean with 46 FPS on DAVIS 2017 dataset. At the same time, this paper constructed a high-speed monitoring video dataset. The algorithm obtained 68.2% $J\&F$ on this dataset, indicating that the method has good generalization and robust performance in practical applications.

INDEX TERMS Deep learning, video object segmentation, conduction model, high-speed monitoring video.

I. INTRODUCTION

With the advancement of modern science and technology, video has become the main form of media communication and social interaction. Some mainstream media have also joined the ranks of video media, marking that video has become one of the most extensive information carriers for audiences. At the same time, due to the application of video in social and security supervision and other fields, video processing technologies have also been developed rapidly, such as temporal video segmentation (TVS) [27] and video object segmentation (VOS) [24]. Therefore, how to improve the accuracy and generalization ability of video segmentation has become an important research topic. Among them, video object segmentation has a wide range of applications in areas such as autonomous driving [28], [29], and video editing [13]. Video object segmentation is mainly divided into unsupervised VOS, semi-supervised VOS, and interactive VOS. The main difference lies in the way users participate in the testing phase.

This paper focuses on semi-supervised VOS. The semi-supervised VOS is to divide the pixels in the video frame into two subsets of the foreground and the background according to the object pixels in the given first frame and generate the mask. It is also the core issue of behavior recognition and video retrieval. Compared with image segmentation, video segmentation should not only consider the appearance of the object but also pay attention to the spatial and temporal changes of the object in the video sequence. This requires the proposed method can fully utilize the temporal information to set up the interdependence between frames and learn the appearance, motion, and scale of the object in different frames. At present, VOS technology has many difficulties in practical application, such as appearance deformation, scale change, occlusion, fast motion, and low resolution.

At present, there have been a lot of achievements and progress in video object segmentation technology based on deep learning. OSVOS [2], OSVOS-S [4], PReMVOS [6],

The associate editor coordinating the review of this manuscript and approving it for publication was Kok-Lim Alvin Yau.

etc., take the object annotation of the first frame given in the semi-supervised VOS task as prior knowledge to guide the segmentation of the current frame. The above methods greatly improved the segmentation performance by optimizing the spatial smoothness in the specific area. However, the information obtained from the first frame may not be optimal for the deformation, occlusion, and reappearance of the object in the video. Therefore, some algorithms [3], [13], [14] uses the similarity of embedded features of pixels to guide model segmentation. These methods do not need to fine-tune the first frame, and by establishing a connection between the current frame and the previous frame, the segmentation performance is significantly improved and better temporal smoothness is achieved. However, this method of spreading local sparse information has a drift problem. When the prediction of the current frame has errors, the prediction errors of the subsequent frames will continue to add up.

Based on these problems, we propose a semi-supervised video object segmentation algorithm based on local and global consistency learning. We use a local and global consistency learning algorithm to process video frames, establish a label transmission model between frames, and take longer-term information into the video segmentation model. The sampling method of the reference frame is proposed to reduce the calculation amount of the algorithm and the time cost of model inference. Finally, to verify the actual application performance of the algorithm, in addition to the DAVIS and Youtube-VOS datasets, this paper constructs a high-speed monitoring video dataset, taking the vehicles on the road as the foreground target in the figure, including the rapid movement of the vehicles, scale changes, and low resolution. Complicated conditions such as speed, mutual occlusion, etc., are used as a supplement to the generalization of the verification algorithm.

This paper is organized as follows. Section II presents a survey on the related work. Section III describes the proposed framework of the video object segmentation approach. The datasets and experimental results on three datasets are presented in Section IV. In Section V, we conclude from the results and discuss our future work plans.

II. RELATED WORK

The popular VOS methods include the single-frame processing method, propagation-based method, and remote spatio-temporal method.

A. THE SINGLE-FRAME PROCESSING METHOD

It is to separate the video into multiple images without considering the temporal information. Fully Convolution Network (FCN) [1] was a classical image segmentation model, which directly predicted the category of pixels through up-sampling to obtain the object mask. Then, OSVOS [2] applied FCN to the video segmentation task, which processed each frame of the video individually. OSVOS-S [4] added an instance-level segmentation module in OSVOS to distinguish each object. A classifier was used to generate the semantic prior information, and another was used to model the appearance of the object. Then, by fusing semantic and appearance information, the model outputted the segmentation mask of the current frame. OnAVOS [5] was also a method based on OSVOS, which selected the credible region in the test sequence as per the reliability and spatial configuration to enhance the training data and adapt the model to the appearance changes of the object. The PReMVOS [6] first employed Region Proposal Network (RPN) and ROIAlign [7] to obtain the rough recommendation of the object, and segmented the object region after clipping, then tracked each instance in the first frame combined with the optical flow of the object, Re-ID feature embedding vector and spatial constraints. To deal with the problems of object deformation, occlusion, disappear and reappear, DyeNet [8] and BubbleNets [9] attempted to improve the performance by searching for more superior initial frames. The above methods greatly improved the segmentation performance by optimizing the spatial smoothness in the specific area. However, due to ignoring the temporal information, they had poor stability in general when the object changes in scale, deformation, and occlusion.

B. THE PROPAGATION-BASED METHOD

It adopted the similarity of embedded features between pixels to guide segmentation. PML [10] transformed pixels in the reference frame into the embedding space and then predicted the pixel category. VideoMatch [11] used a siamese network to extract pixel features and then used the similarity of pixels to match the pixel category. The above two methods selected the first frame as the reference frame. MaskTrack [12], RGMP [13], and FEELVOS [14] added the information of the last frame to the segmentation model, then established the local relevance. These methods improved the segmentation performance by making associations with the first frame or the last frame. However, the above methods only established inter-frame dependencies locally, which will lead to the error expansion with the increase of the number of prediction frames, so the robustness of the algorithm needs to be further improved. TVOS [15] modeled all frames before the current frame and sampled the reference frame, but the complicated network and rough sampling process result in the loss of the segmentation speed and accuracy.

C. REMOTE SPATIO-TEMPORAL METHOD

It was to optimize the dense long-distance space-time, taking into account spatial and temporal information. Tsai *et al.* [16] employed the multi-label Markov Random Field (MRF) to represent the video, then completed the segmentation by solving the minimum energy label allocation. CNN-MRF [17] utilized Convolutional Neural Network (CNN) to encode the spatial correlation of pixels and then described temporal information by optical flow. Finally, a new MRF model was established by combining spatial and temporal correlation, which did not rely on additional modules. BVS [18] designed new energy to

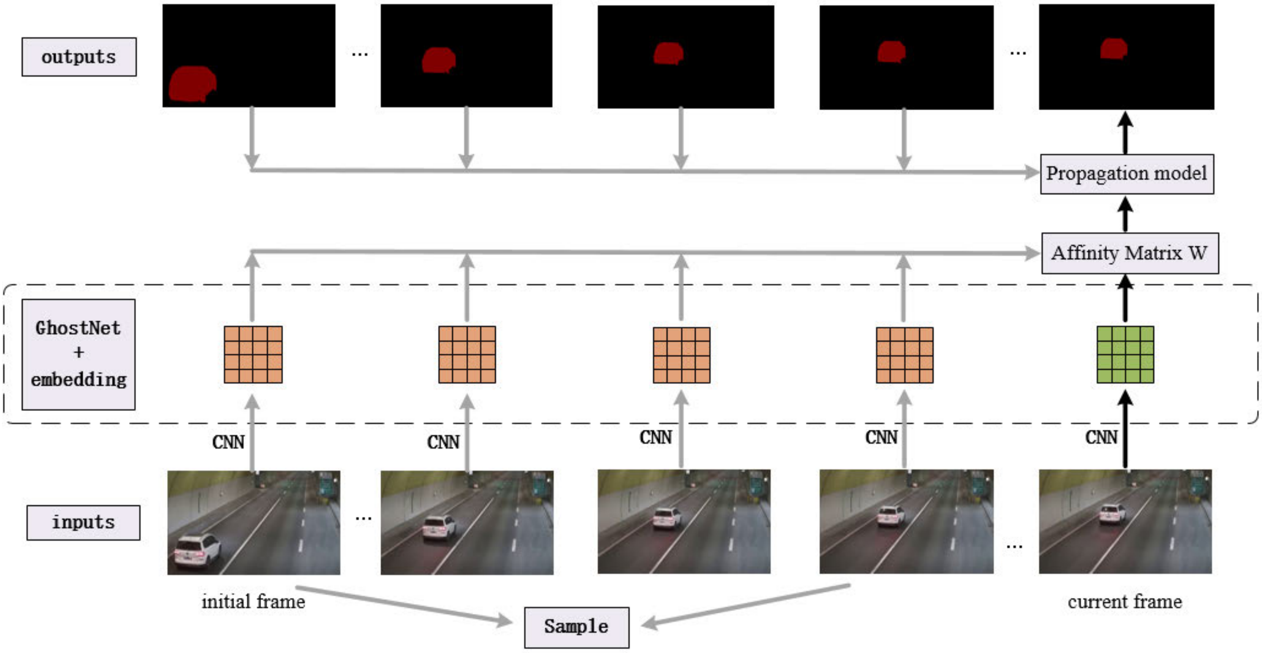


FIGURE 1. A global framework for semi-supervised video object segmentation based on local and global consistency learning. It mainly includes four parts: conduction model, affinity matrix, video frame sampling, and appearance embedding model.

approximate the long-distance spatio-temporal connection between pixels, which only contained few variables. These models have high computational complexity, and their performance cannot be compared with the current learning algorithm.

III. METHODOLOGY

A. FRAMEWORK OF METHOD

Fig. 1 is the schematic diagram of the video object segmentation method proposed in this paper. The process of video object segmentation based on local and global consistency learning mainly includes label propagation model, affinity matrix W, sampling method, and appearance embedding model. First, we construct the online video segmentation model through the local and global consistency learning framework, which is the label propagation model, to establish the connection between the current frame and the historical frame. The affinity matrix $\boldsymbol{W}$ in the propagation model is to describe the similarity measure between pixels. The sampling method obtains a certain number of reference frames from historical frames to improve the inference speed. The appearance embedding model learns the object appearance features in the video.

B. THE FRAMEWORK OF LOCAL AND GLOBAL CONSISTENCY LEARNING

Semi-supervised learning has two important prior consistency hypothesizes [19]: 1) Points with similar locations tend to have the same label; 2) Points with similar internal structures tend to have the same label. Learning with Local and Global Consistency (LLGC) is a graph-based learning algorithm [20], [21]. It constructs a graph model according to the correlation between samples, then obtains a classification function based on the graph model and optimizes it, and finally predicts the label of unlabeled data. The essence of the LLGC is to smooth the classification function, which makes the labels of each sample spread to the adjacent samples until it reaches a stable state.

For the classification problems, it is assumed that there is a sample set $D = \{(x_1, y_1), (x_2, y_2), (x_l, y_l), x_{l+1}, \ldots, x_n\}$, where the $x_i(i \leq l)$ are labeled samples and the $x_u(l+1 \leq i \leq n)$ are unlabeled, the label set $L = \{1, \ldots, c\}$. The task of the classification algorithm is to predict the label of unlabeled samples. In the LLGC algorithm, F is defined as a series of non-negative n × c matrices, representing the label probability corresponding to the sample set D. F can be regarded as a vector function and the label of x_i is the category of the column where the maximum value is in F_i. Define matrix $Y \in F$, when x_i is marked as $y_i = j$, $Y_{ij} = 1$, otherwise the $Y_{ij} = 1$, so matrix Y is consistent with the initial label of the sample.

The steps of the LLGC are as follows:

1. Define the affinity matrix W, which represents the spatial relationship between samples. When $i = j$, the affinity matrix W_{ii} is set to 0 to prevent the sample point from transmitting the label information to itself. When $i \neq j$, the affinity matrix W is shown in equation 1. is defined as the norm. σ is a constant, usually set to the mean value of the distance between each sample.

$$W_{ij} = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \tag{1}$$

TABLE 1. The existing method, brief methodology, highlights, and limitations of the related works.

The existing method	Brief methodology	Highlights	Limitations
The single-frame processing method	OSVOS-S[4], PReMVOS[6], DyeNet[8], etc.	Optimize the spatial smoothness in specific areas	Ignore the temporal information of the video and poor stability
The propagation-based method	PML[10], VideoMatch[11], FEELVOS[14], etc.	Consider the spatio-temporal correlation of video	Lack of global information about the video sequence
Remote spatio-temporal method	CNN-MRF[17], BVS[18], etc.	Consider long-term temporal and spatial information	High computational complexity

2. Construct the matrix $S = D^{-1/2}WD^{-1/2}$. Where D is a diagonal matrix, and D_{ii} is equal to the sum of the i-th row of W.

3. Initialize $F(0)$, and iterate equation 2 until convergence. α takes (0,1) to weigh the proportion of labeled samples and unlabeled samples.

$$F\left(t+1\right) = \alpha SF\left(t\right) + \left(1-\alpha\right)Y \tag{2}$$

4. F^* represents the limit of $F(t)$, then the label of the sample point x_i is $y_i = arg \max_{j \leq c} F^*_{ij}$.

The LLGC algorithm regards the sample set D as a graph $G = (V, E)$. The vertex set V is composed of samples X. The edge E is weighted by W, indicating the similarity between samples. The matrix W is symmetrically normalized to obtain the matrix S, which is conducive to the convergence of the iterative calculation. During the iteration, each sample receives the label information from its adjacent samples and maintains the initial label information of the sample set. Finally, the most label information received by unlabeled samples is the prediction of the model for these samples.

C. VIDEO OBJECT SEGMENTATION APPROACH

The LLGC algorithm can be used to construct the label conduction model, which is used to predict the object mask of unknown video frames. The stable segmentation of the video object needs to rely on the dense pre-frame information, and the video data processing is different from the point data: (1) Since the video frame is processed sequentially, the inference of the current frame cannot rely on the subsequent video frames, the model must be predicted online. (2) The video is composed of multiple single-frame images, and each image has thousands of pixels, so the designed similarity measurement between pixels should be simple and efficient.

1) ONLINE SEGMENTATION METHOD

According to the iterative of LLGC, a label conduction regularization framework is proposed, and the loss function $Q\left(F\right)$ related to the sample label matrix $\hat{y}$ is defined as shown in equation 3.

$$Q\left(\hat{y}\right) = \sum_{i,j}^{n} w_{ij} \left\| \frac{\hat{y}_i}{\sqrt{d_i}} - \frac{\hat{y}_j}{\sqrt{d_j}} \right\|^2 + \mu \sum_{i=1}^{l} ||\hat{y}_i - y_i||^2 \tag{3}$$

w_{ij} represents the similarity between pixel i and j and d_i represents the sum of row i in the affinity matrix of pixel j and reference pixel j. The first term represents the smoothing constraint, which makes the sample points with similar locations more likely to be in the same label category. The latter term represents a consistency constraint, which can be used as the penalty when the prediction result is inconsistent with the initial label. This constraint includes labeled data and unlabeled data. μ is a positive parameter that balances the two constraints.

Therefore the classification problem can be expressed as the following equation:

$$\hat{y} = \text{argmin} Q\left(\mathbf{y}\right) \tag{4}$$

For the VOS task, when inferring the t frame, the front $t-1$ frames of the video have been predicted. Therefore, the iterative process can be realized in the order of video sequence. According to equation 2, the iterative equation of the video can be approximated as the following equation:

$$\hat{y}\left(t+1\right) = S_{1:t\rightarrow t+1}\hat{y}\left(t\right) \tag{5}$$

$S_{1:t\rightarrow t+1}$ represents the similarity matrix S, which calculates the similarity between the pixels from the first frame to the t frame and the pixels from the $t+1$ frame. Since there is no label before the first frame, the initial label value Y item is omitted for the $t+1$ frame.

For $t+1$ frame, the label propagation process of the above equation can be expressed as minimizing the smoothing term in the loss function. i is the index of t+1 frame, j is the index of all previous frames.

$$Q^{t+1}\left(\hat{y}\right) = \sum_{i}\sum_{j} w_{ij} \left\| \frac{\hat{y}_i}{\sqrt{d}_i} - \frac{\hat{y}_j}{\sqrt{d}_j} \right\|^2 \tag{6}$$

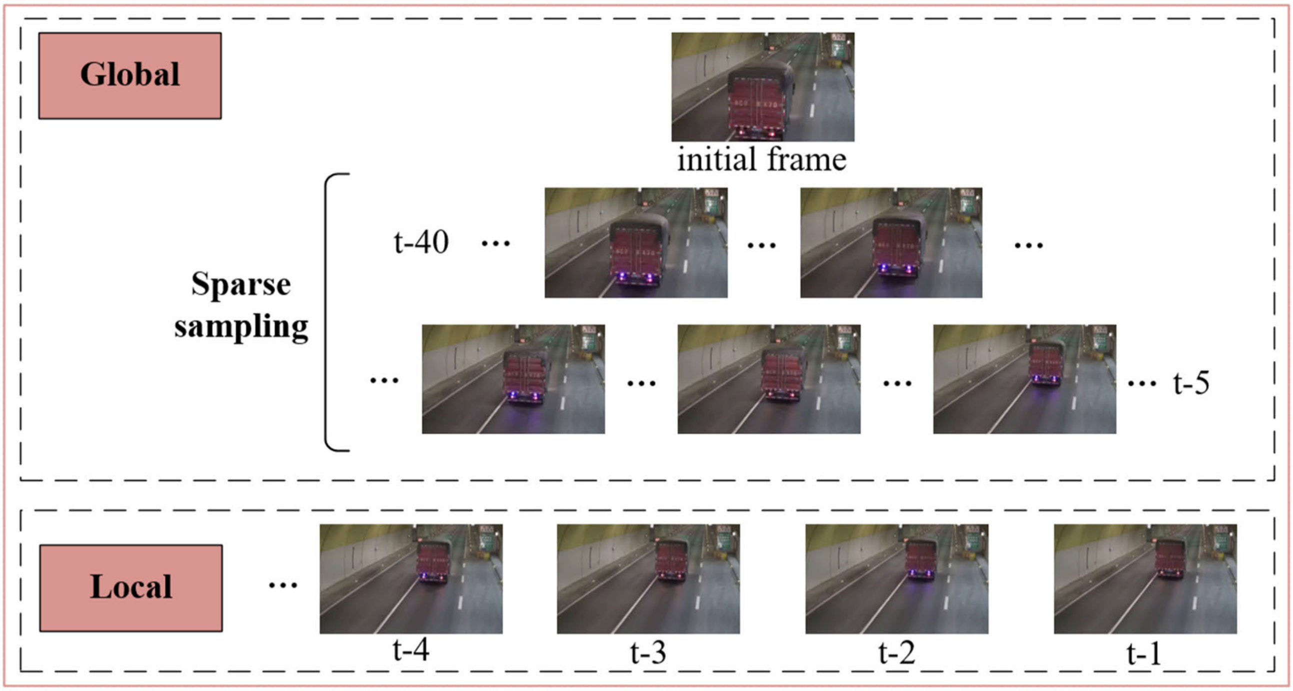


FIGURE 2. Schematic diagram of the sampling method. Intensive sampling is used in the sequence closer to the current frame, while sparse sampling is used in the long-distance frame.

2) SIMILARITY MEASURE

The similarity measure is the core of the label propagation model, which defines the similarity between two pixels, and is the basis to predict the pixel category. Therefore, the quality of video object segmentation depends heavily on similarity measures.

To establish a smooth classification function, the similarity measures should consider both semantic and spatial information. The similarity measure w_{ij} is shown in equation 7, which includes an appearance item and a space item. f_i, f_j are the embedded feature values extracted from the pixels p_i, p_j through CNN. $loc(i)$ represents the spatial position of pixel i. The spatial term is controlled by the locality parameter σ.

$$w_{ij}= \exp\left(f_i^T, f_i\right) \cdot \exp\left(-\frac{\|loc\,(i) - loc\,(j)\|^2}{\sigma^2}\right) \tag{7}$$

3) SAMPLING METHOD

For video sequences of hundreds of frames or more, the computation of similarity matrix S on all previous frames has high complexity, so reasonable simplification methods need to be adopted to increase the model rate. Since the video has the characteristics of the small difference between close range frames and the weak connection between distant range frames, sampling methods need to take into account both close and distant range frames. Close range frames can achieve local information association, while distant range frames can contact global information so that the local and global information is integrated to ensure the accuracy and robustness of the model. In this paper, small amount frames are sampled according to the temporal redundancy in the video.

Specifically, We save the 40 frames before the current frame as historical frames. 9 frames are sampled from the historical frames: the first frame; 4 consecutive frames before the current frame; and the other 4 frames are sparsely sampled in the sequence between 40 and 5 frames from the current frame. That is, dense sampling is used in the sequence closer to the current frame, while sparse sampling is adopted in the long-distance frame. When the current frame t is less t han 10, all frames must be sampled.

During sparse sampling, the sampling frames are selected based on the similarity between frames. Firstly, embedding features of the video frames from $t-40$ to $t-5$ and the current frames are reduced by using the feature hashing method [22], that is, the feature vector is mapped to the low-dimensional space using a hash function. The hash function uses the non-cryptographic hash function MurmurHash3. Since the number of sparsely sampled candidate video frames is not more than 35, the number of columns of the hash matrix output by the hash function is set to 6, which can represent $2^6 - 1$ feature vectors at most to avoid hash collisions in low-dimensional spaces. Figure 3 is a schematic diagram of dimensionality reduction using the feature hash method.

Then we measure the distance of the low-dimensional vector output by the feature hash and utilize the Hamming distance [23] to calculate the distance between the vectors. By calculating the Hamming distance between the current frame and the corresponding embedding vector of the previous $t - 40$ to $t - 5$ frames in the low-dimensional space,

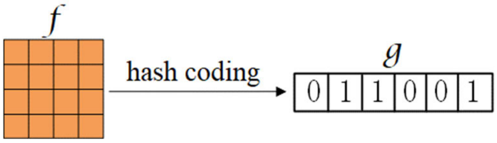


FIGURE 3. Dimensionality reduction through feature hashing.

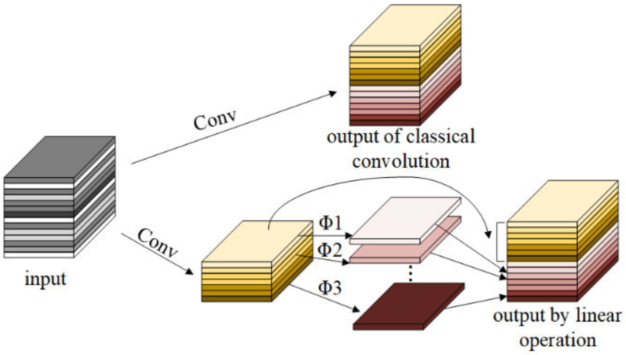


FIGURE 4. Schematic diagram of ordinary convolution and Ghost module.

the first 4 frames with the smaller value are taken as the sparse sampling frames.

D. APPEARANCE EMBEDDED MODEL

The CNN is adopted to learn the appearance feature of the object in the video so that the method can adapt to the short-term and long-term changes caused by the movement, deformation, and scale change. In this paper, we employed the lightweight CNN GhostNet as the backbone network of the appearance embedding model. Similar feature map pairs are generally obtained in deep CNN, and the key idea of GhostNet is to obtain these similar feature maps by simpler calculation. Firstly, a small number of convolution kernels are used to manipulate the image, and then the redundant feature maps are mapped from the feature maps derived from the convolutions with cheap linear operations. It reduces parameters and computation with a low cost of precision without changing the output size and channel number of the feature maps.

The appearance embedding model learns the appearance of the object from the training data. Firstly, given the object pixel x_i, and all pixels in the previous frame are regarded as references. f_i, f_j represents the embedded features extracted in the backbone network of the pixel x_i and its reference pixel x_j, and the predicted label $\hat{y}_i$ of the pixel x_i can be expressed as equation 8.

$$\hat{y}_i = \sum_j \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_i^T f_k\right)} \cdot y_j \tag{8}$$

The reference indexes j and k indicate the time range before the current frame. The VOS task is to classify pixels, so the cross-entropy loss function is applied to optimize the appearance learning model. In Equation 9, x_i is all the pixels in the current frame, and y_i, $\hat{y}_i$ are the labels and predicted labels of the pixels.

$$L = -\sum_i \log P\left(\hat{y}_i = y_i \mid x_i\right) \tag{9}$$

IV. EXPERIMENTAL RESULTS

In this section, we introduce the details of the evaluation, including the datasets, settings of the experiments, and obtained results.

A. DATA PREPARATION

We use DAVIS 2017 [24], Youtube-VOS [25], and a self-built high-speed monitoring video database to verify the effectiveness of the method. DAVIS is the most representative database in the field of VOS, covering complex scenes such as object deformation, occlusion, rapid movement, blur and defocus. DAVIS database has high quality and all frames of the video sequence are marked with high-resolution pixels. Youtube-VOS is the largest and most comprehensive dataset in the VOS field. The videos are from the YouTube video website. The target category and scene are comprehensive, and the video shooting equipment is very diverse, so the video quality and resolution are very different, which is close to the actual application of video segmentation. In addition, we select the monitoring video on the high-speed lane as the sample, take the driving vehicle in the video as the foreground object, and intercept the video clips to construct the high-speed monitoring video dataset. The fixed position of cameras results in a limited viewing angle. Since the vehicle is in a state of high-speed motion, to maintain the continuity between video frames, we set the frame rate of the extracted frame to 1, then convert the video into images, and utilize the image annotation tool named labelme [26] for pixel-level annotation of video frames. According to the proportion of vehicles in the video, we divide the vehicles in the road into two labels: small buses (cars) and large trucks (trucks), and multiple vehicles of the same category require instance differentiation. The high-speed monitoring video dataset contains 30 video sequences, and the entire dataset has 1425 frames, with an average of 47.5 frames per video. The dataset is randomly divided into 18 training sets and 12 test sets.

Considering the cost of annotation, we expand the video frames and the corresponding annotations of the self-built dataset. The clockwise 180° rotation and mirroring method are used to transform the video frame, and then the obtained image is recombined into a new video sequence. Figure 5 shows the original high-speed monitoring video sample and its corresponding expanded sample. Table 2 shows the basic information of the above three datasets.

B. IMPLEMENTATION

The proposed approach is implemented in PyTorch. It is running on a computer equipped with Intel(R) Core

TABLE 2. Summary of DAVIS 2017, Youtube-VOS, and high-speed monitoring video data sets.

Data set	Number of videos	Total number of frames	The average number of video frames	characteristic
DAVIS 2017	150	10459	69.7	The video quality is high, including basic segmentation scenes such as appearance distortion, scale change, and occlusion.
Youtube-VOS	4972	197272	44.3	The amount of data is large, the target categories are many, and the video resolutions are different.
High-speed monitoring video	30	1425	47.5	The video resolution is low, the background is single, and it mainly contains fast motion, scale changes, and occlusion scenes.

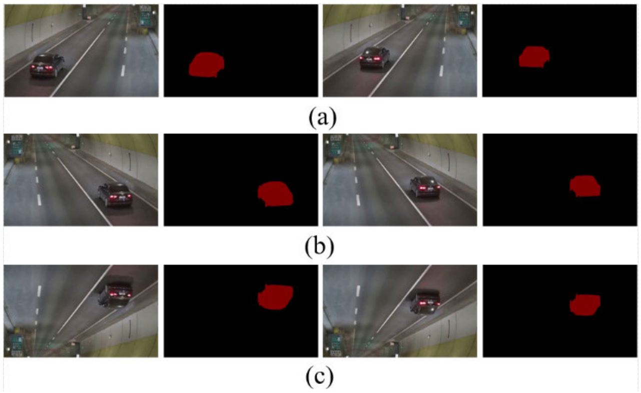


FIGURE 5. Sample expansion example. (a) is the original video frame; (b) is after mirror flipping; (c) is after 180° rotation.

(TM)i7-7700K CPU, an NVIDIA GeForce GTX 1080 TI GPU, and 16GB of RAM under Ubuntu16.04.

1) TRAINING PHASE

In this section, GhostNet is employed as the backbone of the method to extract the appearance embedding features of the object. To better retain the high-pixel features, the Ghost module adopts a convolution operation with a step length of 1 in the fourth and fifth stages of the network. At the same time, a layer of 1 × 1 convolution kernel is added at the end of the network to generate 256-dimensional embedding features. When training the appearance embedding model, the GhostNet weights are pre-trained on the ImageNet dataset, and then fine-tuned on different datasets. Iterated 240 epochs on the Davis 2017 dataset and self-built high-speed monitoring video dataset, and 30 epochs on the Youtube-VOS dataset.

In this paper, the input image is randomly flipped and randomly cropped to 256 × 256 size for data enhancement. The SGD stochastic gradient descent training strategy is adopted, the number of batches is set to 2, the initial learning rate is 0.0025, and cosine annealing is used to reduce the learning rate. Use NVIDIA GeForce GTX 1070 GPU to optimize 60 hours, 160 hours, and 17 hours on Davis 2017, Youtube-VOS, and high-speed monitoring datasets respectively.

2) INFERENCE PHASE

On the test video, the proposed method is used for online derivation. Based on this prior knowledge, we set an adaptive object action prior, that is, the value of σ in w_{ij} varies with the distance between the sampling frame and the current frame. When the frame obtained by dense sampling is close to the current frame, σ is 8. And for frames obtained by sparse sampling, σ is 21.

C. COMPARISON OF SAMPLING METHODS

Effective use of local and global information is the key to ensuring the accuracy of the video segmentation model. To verify the effectiveness of the proposed sampling method, this section uses different reference frame sampling strategies in the training phase and the test inference phase to study the impact of local and global information on the model's segmentation accuracy. We have selected: 1) 1 reference frame before the target frame; 2) 3 consecutive frames before the target frame; 3) 9 consecutive frames before the target frame; 4) Even in the first 40 frames of the target frame Sampling 9 frames; 5) The proposed sampling method in section 3.3.3. Table 3 shows the results of the regional similarity J of the different sampling methods on the DAVIS 2017 validation set.

In table 3, when the test sequence is inferred, the more consecutive frames before the target frame are selected, the higher the segmentation accuracy of the model, which indicates that the dense sampling before the target frame is helpful to improve the accuracy of the segmentation model. At the same time, when training the appearance embedding model, the model with 9 consecutive frames before the frame has the best segmentation effect, reaching a region similarity of 68.5, which shows that the long-distance frame does not improve the segmentation accuracy of the target frame. The method of uniformly sampling 9 frames has limited improvement in the algorithm. This may be because when the video is too long, the long-distance frame changes greatly from the current frame, which leads to worse training effects. The accuracy of the 9-frame sparse sampling method in this paper is between the uniform sampling and continuous sampling methods.

TABLE 3. Comparison of results of different sampling methods.

Training/inference	1 frame	3 frame	9 frame	9 uniform sampling	Our sampling method
1 frame	52.6	57.5	60.3	60.7	62.8
3 frame	53.1	58.5	62.5	62.5	64.9
9 frame	57.8	60.6	65.4	65.4	68.5
9 uniform sampling	52.5	59.7	62.2	64.1	65.7
Our sampling method	57.2	60.4	64.5	66.3	67.4

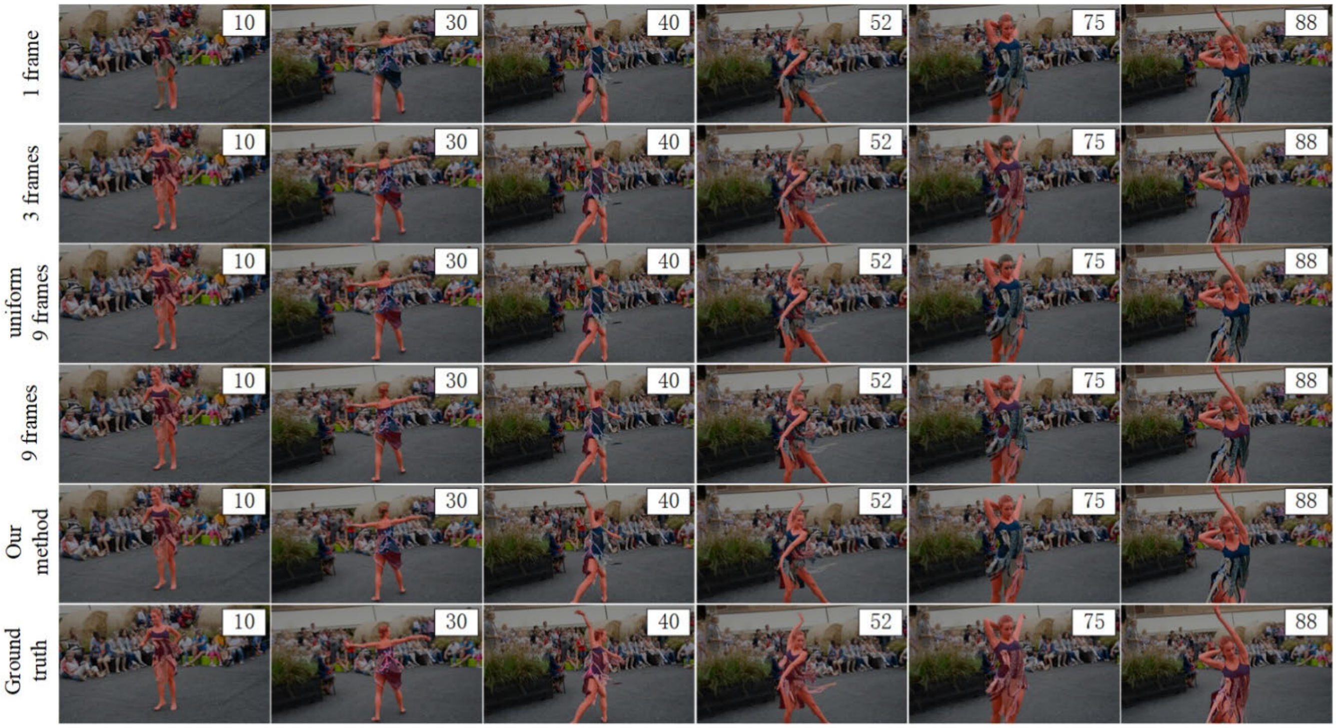
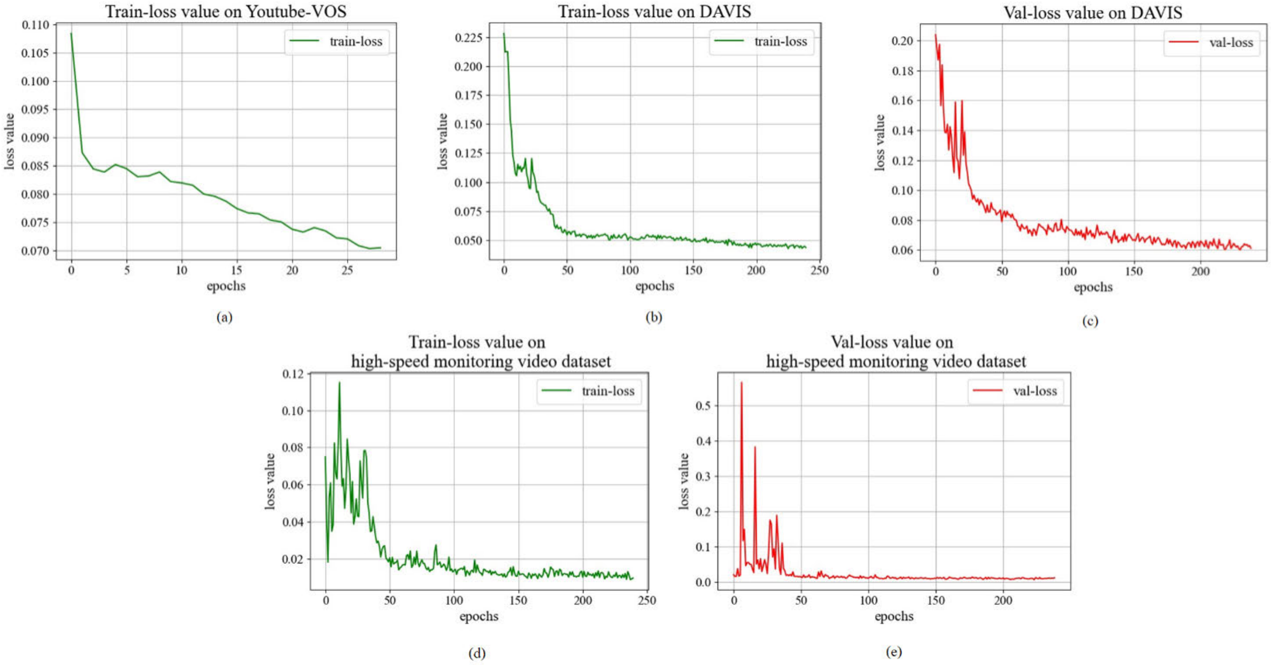


FIGURE 6. Comparison of different sampling methods.

Figure 6 shows the test video segmentation results of different sampling strategies. The sampling method of continuously sampling 9 frames has a more stable segmentation result than the method of sampling 3 frames and 1 frame, indicating that the continuous frame before the current frame is sampled, and the local information of the video sequence is taken into account. The more the number of sampling frames, the more the segmentation effect The better. Compared with the uniform sampling method and the continuous sampling method, the segmentation effect is unstable. The 9-frame sparse sampling method used in this article can segment the target more completely after the video is divided into 50 frames. This is because the sampling method in this article considers both the long-distance frame and the short-distance frame, and is compared with the first frame of the video. Establish a dependency relationship, ensure that the initial information is not lost, and take into account the local and global information of the video so that the model has a better segmentation effect and robustness.

D. SEGMENTATION RESULTS

The experiments in this section were carried out on the DAVIS 2017 data set, Youtube-VOS data set and high-speed monitoring video data set. The results of this model on different data sets were obtained and compared with other advanced algorithms. Figure 7 shows the change of the loss function value of the model in this paper when it is trained on different data sets.



(a)

(b)

(c)

(d)

(e)

FIGURE 7. **Loss changes during model training.**

According to Figure 7, the loss value of the model in this paper can be stable on different data sets. When training on the high-speed monitoring video data set and DAVIS data set, the loss value of the training set and the validation set was relatively large at first. By continuously adjusting the learning rate to adapt to the sample, the loss value dropped rapidly after training to about 60 epochs. After epochs reached 60 times, the loss changes began to stabilize. The Youtube-VOS dataset only labels the first frame on the validation set, so only the loss of the training set changes. After training for two epochs, the loss value drops rapidly, and then after 25 epochs, the loss change gradually stabilizes.

The loss value during training is finally stable in a small range, indicating that the model training effect is good. The segmentation results of the model on different data sets are analyzed below.

1) DAVIS 2017 VALIDATION SET

To verify the effect of the segmentation algorithm based on the conduction model, Table 4 shows the test results of different segmentation algorithms on the DAVIS 2017 validation set in recent years. FT means that the algorithm needs to be fine-tuned online in the first frame, $J\&F$ means the average of regional similarity and contour similarity, and FPS is the number of frames processed by the model per second. Figure 8 shows a comparison of performance and speed for semi-supervised video object segmentation methods on the DAVIS 2017 validation set. Among the algorithms that do not need to be fine-tuned, the method in this paper achieves a regional similarity of 67.4 and a contour accuracy of 71.6. The number of frames processed by the algorithm reaches 46 frames per second. The rate of the model is the highest, and the accuracy exceeds that of other algorithms without fine-tuning except the TVOS algorithm. This is because the appearance embedding model in this paper uses GhostNet as the appearance learning network, which has fewer network parameters and a small memory footprint, but the network feature learning ability has declined. The PReMVOS algorithm has the highest accuracy, reaching an area similarity of 73.9 and a contour similarity of 81.7. However, the network is more complicated, the algorithm calculation is large, and the running speed of the model is not high. The method in this paper has the same accuracy as the DyeNet and CNN-MRF algorithms that need to be fine-tuned and has a greater speed advantage. Compared with FEELVOS, the regional similarity of the method in this paper is slightly higher. This is because although FEELVOS considers the information of the previous frame and the first frame, the inter-frame dependence established is very sparse, and the calculation method of pixel matching is complicated, and the algorithm is real-time. not tall.

2) YOUTUBE-VOS VERIFICATION SET

Table 5 shows the segmentation results of the algorithm on the Youtube-VOS verification set. There are two different evaluation methods of seen and unseen on this data set, which respectively represents the existing target category and the unknown target category in the training set. Unseen can be used to verify the generalization performance of the algorithm to general foreground targets. The method in this paper surpasses all algorithms except PREMVOS, and the average accuracy Overall is equivalent to PREMVOS.

TABLE 4. Comparison of advanced methods on DAVIS 2017 dataset.

Method	FT	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$&$\mathcal{F}$	FPS
OnAVOS[5]	√	61.0	66.1	63.6	0.08
DyeNet[8]	√	67.3	71.0	69.1	0.43
CNN-MRF[17]	√	67.2	74.2	70.7	0.03
PReMVOS[6]	√	73.9	81.7	77.8	0.03
VideoMatch[11]	×	56.5	68.2	62.4	2.86
RGMP[13]	×	64.8	68.8	66.7	3.57
FEELVOS[14]	×	65.9	72.3	69.1	1.96
TVOS[15]	×	69.9	74.7	72.3	37
Our approach	×	67.4	71.6	69.5	46

TABLE 5. Comparison of advanced methods on Youtube-VOS dataset.

Method	Overall	$\mathcal{J}$ (Seen)	$\mathcal{F}$(Seen)	$\mathcal{J}$(Unseen)	$\mathcal{F}$(Unseen)
RGMP[13]	53.8	59.5	-	45.2	-
OnAVOS[5]	55.2	60.1	62.7	46.6	51.4
OSVOS[2]	58.8	59.8	60.5	54.2	60.7
PReMVOS[6]	66.9	71.4	75.9	56.6	63.7
Our appoarch	66.4	66.2	68.6	61.4	69.2

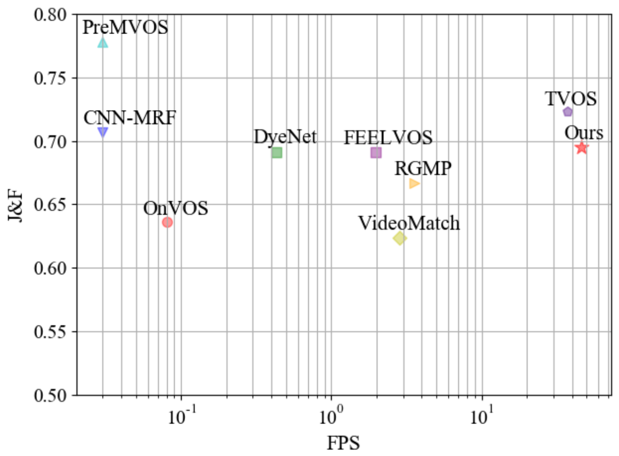


FIGURE 8. A comparison of performance and speed for semi-supervised video object segmentation methods on the DAVIS 2017 validation set.

TABLE 6. Comparison of advanced methods on high-speed monitoring video dataset.

Method	FT	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$&$\mathcal{F}$
OnAVOS[5]	√	60.3	64.5	62.4
DyeNet[8]	√	65.9	69.4	67.6
CNN-MRF[17]	√	66.1	73.8	69.9
PReMVOS[6]	√	72.3	79.6	75.9
VideoMatch[11]	×	53.0	65.3	59.1
FEELVOS[14]	×	64.6	70.8	67.7
Our approach	×	66.3	70.2	68.2

Compared with the two single-frame processing models of OSVOS and OnAVOS, as well as the RGMP method that only considers the prediction mask and features of the previous frame, the accuracy advantage of the algorithm in this paper is more significant under the unseen category than under the seen category, indicating that it is especially For random videos of unknown categories, methods that do not consider timing information or only consider local sparse information are less robust and difficult to adapt to changes in new videos. For the segmentation of unseen category, the method in this paper achieves a regional similarity of 61.4 and a contour accuracy of 69.2, which surpasses all other methods in the table, indicating that the algorithm in this paper can adapt to the target of unknown categories in the video and has good generalization performance.

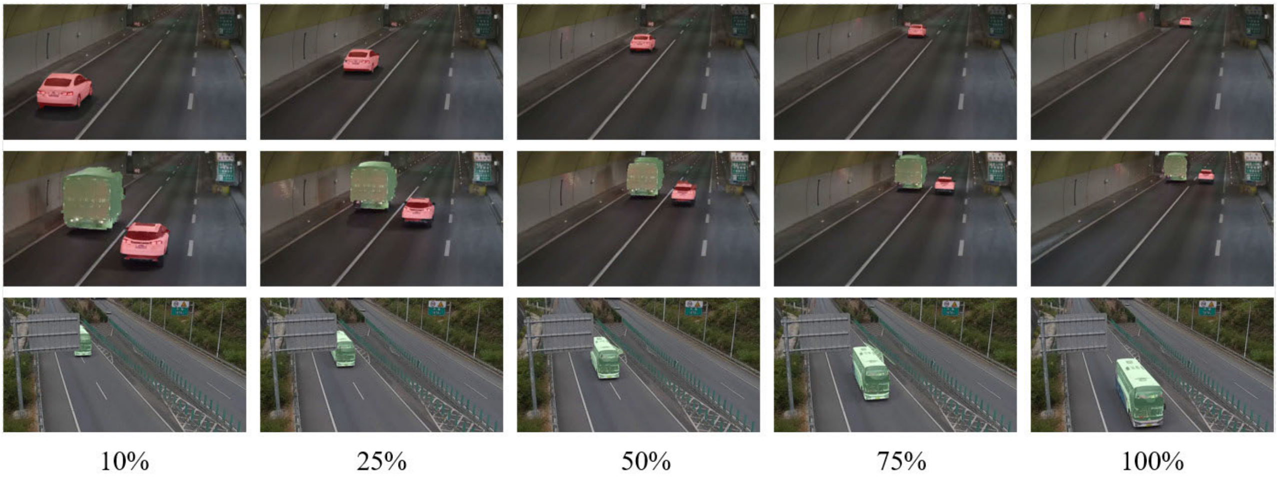


FIGURE 9. Part of the results of the method in the high-speed monitoring video dataset.

3) HIGH-SPEED MONITORING VIDEO DATA SET

To verify the actual application performance of the proposed method, we tested the method on the constructed high-speed monitoring video dataset. Table 6 shows the results of different algorithms on the high-speed monitoring video dataset. The proposed method has the best effect in algorithms that do not require fine-tuning of the first frame, reaching the region similarity of 66.3 J and the contour accuracy of 70.2 F, and the average accuracy $J\&F$ is also the highest, reaching 68.2. The algorithm PReMVOS has the highest accuracy among all the algorithms in the table, with an average accuracy $J\&F$ of 75.9. Compared with VideoMatch considering pixel similarity matching in the first frame, the local and global dependencies established by the proposed method can effectively improve the accuracy of the algorithm. Figure 9 is the segmentation result of this algorithm on the video sequence. It can be seen that the proposed method has a good segmentation effect when the object is fast-moving and scale transformation, and when the object is occluded in the third row, it can ensure better edge segmentation. The results show that the proposed method has better robustness and generalization in practical applications.

V. CONCLUSION

In this paper, we propose a semi-supervised video object segmentation method based on local global consistency learning. This method combined the classic semi-supervised learning method with the video object segmentation task and applied the graph-based learning algorithm framework to process video. Instead of establishing dependencies between the previous frame or the initial frame, we used more unlabeled frames to improve the robustness and generalization of the method. Then we proposed the sampling method which takes into account the local and global information of the video. It not only reduced the complexity and memory consumption but also ensures the segmentation stability of the model. To verify the practical application performance of the method, in addition to the DAVIS-2017 and Youtube-VOS datasets, we constructed a high-speed monitoring video dataset. The experimental results showed that the proposed method had a great segmentation effect and high prediction speed in the methods without fine-tuning, indicating that it had good practical application value.

At present, the real-time and lightweight model is a trend. This method adopted GhostNet as the backbone network to reduce the parameters and designed a sampling method to reduce memory occupation and computation. However, the accuracy of the model was insufficient. The backbone network more suitable for image segmentation tasks should be used to improve the segmentation accuracy. At the same time, the background of the high-speed monitoring video dataset constructed in this paper was relatively single, and the samples were fewer, which were not comprehensive to measure the generalization of the method. Therefore, further research can be conducted on scene selection and annotation of video.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[2] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 221–230.

[3] Y. Yin, D. Xu, X. Wang, and L. Zhang, "Directional deep embedding and appearance learning for fast video object segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 15, 2021, doi: 10.1109/TNNLS.2021.3054769.

[4] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1515–1530, Jun. 2019.

[5] P. Voigtlaender and B. Leibe, "Online adaptation of convolutional neural networks for video object segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 116.1–116.13, doi: 10.5244/C.31.116.

[6] J. Luiten, P. Voigtlaender, and B. Leibe, "PReMVOS: Proposal-generation, refinement and merging for video object segmentation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Perth, NSW, Australia, 2018, pp. 565–580.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 2961–2969.

[8] X. Li and C. C. Loy, "Video object segmentation with joint re-identification and attention-aware mask propagation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 90–105.

[9] B. A. Griffin and J. J. Corso, "BubbleNets: Learning to select the guidance frame in video object segmentation by deep sorting frames," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8914–8923.

[10] Y. Chen, J. Pont-Tuset, A. Montes, and L. V. Gool, "Blazingly fast video object segmentation with pixel-wise metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1189–1198.

[11] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "Videomatch: Matching based video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 54–70.

[12] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2663–2672.

[13] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7376–7385.

[14] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen, "FEELVOS: Fast end-to-end embedding learning for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9481–9490.

[15] Y. Zhang, Z. Wu, H. Peng, and S. Lin, "A transductive approach for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 6949–6958.

[16] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, Nov. 2012.

[17] L. Bao, B. Wu, and W. Liu, "CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5977–5986.

[18] N. Marki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 743–751.

[19] D. Zhou, O. Bousquet, T. N. Lal, and J. Weston, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2014, pp. 321–328.

[20] X. Wang, X. Zhang, and Y. Cheng, "Barebones learning with local and global consistency," *Control Decis.*, vol. 26, no. 11, pp. 1726–1730, 2011.

[21] Z. Xie, Y. Li, and S. Zheng, "Weakly supervised hand segmentation based on local and global consistency learning," *Comput. Appl. Softw.*, vol. 36, pp. 204–210, Apr. 2019.

[22] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, "Feature hashing for large scale multitask learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2009, pp. 1113–1120.

[23] S. Shi, "Research on the locality sensitive hashing," M.S. thesis, Dept. Prof. Comput. Softw. Theory, Xidian Univ., Xi'an, China, 2013.

[24] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 724–732.

[25] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "Youtube-VOS: Sequence-to-sequence video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 585–601.

[26] K. Wada. (2016). *Labelme: Image Polygonal Annotation with Python*. [Online]. Available: https://github.com/wkentaro/labelme

[27] S. H. Abdulhussain, S. A. R. Al-Haddad, M. I. Saripan, B. M. Mahmmod, and A. Hussien, "Fast temporal video segmentation based on krawtchouk-tchebichef moments," *IEEE Access*, vol. 8, pp. 72347–72359, 2020.

[28] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. M. Lopez, "Vision-based offline-online perception paradigm for autonomous driving," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 231–238.

[29] K. Saleh, M. Hossny, and S. Nahavandi, "Kangaroo vehicle collision detection using deep semantic segmentation convolutional neural network," in *Proc. Int. Conf. Digit. Image Computing: Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–7.

HUAGANG LIANG received the B.S. degree from Xi'an University of Technology, Shaanxi, China, in 2003, and the M.S. and Ph.D. degrees from Fukui University, Japan, in 2006 and 2009, respectively. He is currently an Associate Professor with the College of Electronic and Control Engineering, Chang'an University. His research interests include pattern recognition, machine vision, and intelligent transportation.

LIHUA LIU received the B.S. degree from Lanzhou University of Technology, Gansu, China, in 2019. She is currently pursuing the M.S. degree in control science and engineering with Chang'an University. Her research interests include pattern recognition and deep learning.

YING BO received the B.S. degree from Anhui University of Science and Technology, Anhui, China, in 2019. She is currently pursuing the M.S. degree in control engineering with Chang'an University. Her research interests include target tracking and deep learning.

CHAO ZUO received the B.S. degree from North China University of Water Resources and Electric Power, Henan, China, in 2018. She is currently pursuing the M.S. degree in control science and engineering with Chang'an University. Her research interests include pattern recognition, object detection, and segmentation.