Segmentation of Motion Objects in Video Frames using Deep Learning

Feng JIANG[1], Jiao LIU[2], Jiya TIAN[3*]
School of Electrical Information, Changchun Guanghua University, Changchun 130033, China[1]
Student Affairs Office, Changchun Guanghua University, Changchun 130033,China[2]
School of Information Engineering, Xinjiang Institute of Technology, Aksu 843100, China[3]

***Abstract*—The segmentation of the moving objects in the video sequences is one of the most usable series in the machine vision field, which has absorbed the consideration of researchers in the latter decades. It is a challenging task, especially when there are several motion objects in the video, and then the system needs to discover the objects that should be segmented among the trail. Therefore, in this article, we present a new method to segment several motion objects at the same time. In this work, the propagation of the credence of the confidently-estimated frames by fine-tuning the DCNN model with the other frames is the main idea. We exert a DCNN model (which is pre-trained) for the frames to estimate the class of the object; then, we gather the frames where the approximation is locally or globally reliable. In the following, we apply a collection of the frames of CE as the training set to fine-tune the pre-trained network with the existing examples in a video. Our proposed model provides acceptable results, which are better than the results of similar models. These comparisons are made in the dataset of YouTube-VOS. Also, our presented approach is applied in the dataset of DAVIS-2017 and the obtained results are better than the results of the similar works.**

Keywords*—Segmentation; video processing; motion objects; deep convolutional neural network (DCNN)

I. Introduction

With the growth of technology and the presence of machines in human life, the various applications of this technology have increased daily. Currently, with the increase in computing capabilities along with the low price of the cameras, image perception is an important part of many applications. One of these applications in image processing is the segmentation of motion objects. The segmentation of the object is the separation of the background and the objects in a trail of video images with a specific purpose [1]. The discovery and segmentation of the moving objects on the trail of videos is a prerequisite step for the high-level systems of machine vision, such as stewardship systems, robotics, and so on. The accuracy of the mentioned systems depends on the segmentation method used. For example, in a surveillance system that uses the information of the movement model for the recognition of people, it should be possible to continuously segment and track the moving objects with high accuracy through the installed cameras in the desired location. Then, by analyzing the received information about the movement and the location of these people, in case of unfortunate events such as falling down which occur, the system can be notified automatically to the relevant centers such as the emergency [2]. With the consideration of the important mentioned applications in the above and many other applications of object segmentation, in the current article, we propose a novel approach for the segmentation of the object in video trials.

Therefore, the main purpose of the segmentation of the video is to separate the foreground from the background with respect to a video trail [3]. Recently, new approaches have been proposed to segment all motion objects in a video and produce larger datasets. This work leads to more challenging tasks [4]. Most of the presented methods in this field evaluate the frames separately [5], and they do not remark on the dimension of the temporal to obtain the affiliation among the successive frames. Recently, an encoder-decoder architecture has been presented based on RNN [6] and is similar to our proposed method.

Therefore, in this paper, the key idea is the propagation of the CE frame credence into another frame using the fine-tuning of the model of DCNN. So, we exert the DCNN model (which is pre-trained) for the frames to estimate the class of the object, and then, we gather the frames where the estimation as globally or locally is reliable. In the following, we exert a collection of the frames of CE as the training set to fine-tune the used pre-trained model with the examples in the videos. Also, we confine the used model of DCNN [7] to only the video. For example, we perform the model centralization in the particular examples in the input video. We, in this procedure, only use the CE region labels and permit the CE frames to determine the un-estimated regions. In addition, we use the feeble labels to prevent the degradation of the model by a few incorrect labels. Our procedures for the generation of the self-consistent datasets and the use of the CE frames for the updation of the system can retrieve the unspecified parts or the classified sections from the frames of UE, which contain several objects.

The article continuation is as the below: Section II characterizes the related works and their overview. Section III characterizes the details of our presented method. The evaluation details and the details of the performed tests are provided in Section IV. In this section also, we provide the visual outcomes and the numerical outcomes of the done tests. In Section V, we provide the suggestions and conclusions.

II. Related Works

Due to the wide applications of the segmentation of motion objects in the arena of machine vision, researchers have studied and have presented different methods for this task in recent years. Among the comprehensive performed works in this field is the presented work in [8], which has reviewed and classified the proposed segmentation methods. In [9]–[12], the different methods for the segmentation and the tracking of the motion objects have been investigated. Usually, the segmentation is done based on the obtained information for a series of special characteristics from the objects. These characteristics include the below cases: the edge, the texture, the color of the objects, the movement information, the corner points, the appearance of the objects, etc. Any segmentation algorithm based on the application can use any of these characteristics or a combination of these characteristics. For example, in the algorithms that segment and track the objects based on the object contour, the edge feature [13] is used. In [14], the motion objects were segmented based on the difference between the existing edges between two consecutive frames. The detectors of the corner points in the literature on object segmentation are Moravec [15], Harris [16], KLT [17], and SIFT [18].

In addition, the segmentation of moving objects based on deep learning techniques has received regard in the association of the research in the latter years. It can be due to the emergence of novel segmentation datasets and new challenges: Berkeley (2011), SegTrack (2013) [19], Berkeley Freiburg (2014) [20], DAVIS (2016-2017) [21], and YouTubeVOS (2018) [22]. These datasets provide the biggest content of the tagged videos.

The later works, such as [15], use the optical flow for the temporal adaptation after the use of the fields of Markov random, which is the basis on the taken specifications of a CNN model. The other suggestion for the obtention of the coherence of the temporal is the use of the boded masks on prior frames as a guide for the subsequent frames [7]. The proposed method in [23] disseminates the information using spatiotemporal features. Finally, the proposed method in [24] uses an architecture of the encoder-decoder RNN that employs the LSTM for the learning of the trail.

In the segmentation of the objects in the video, the learning with the single-shot is found as the use from an alone tagged frame for the estimation of the residual frame's segmentation in a sequence. Also, the learning with the zero-shot is found as the construction models, which do not require the initialization for the generation of the masks of the segmentation of the object in the trail of the video. There are multiple articles in the literature which is emphasized the first mask for the input to can propagate via the trail [3], [7], [10], [25], and [26]. Generally, the approaches with the single-shot outperform in comparison to the approaches with the zero-shot because the first segmentation is formerly taken, so there is no need for the estimation of the mask of the first segmentation of the abrasion. Most of the proposed systems emphasize online learning, which is the adaption of the weights with the first frame and associated masks. Usually, the methods of online learning achieve better outcomes, but they need more computing time. On the learning with the zero-shot, for the estimation of the segmentation of the object on a video, multiple papers have used the object saliency [8], [27], [28] or they have used the object suggestion methods outputs [12], or they have used network with two-stream. The exploitation of the motion templates on the videos is perused at [29], but the article of [14] formulates the 3D representation conclusion of a planar object and the motion segmentation. In addition, foreground segmentation which is the basis of the sample embedding is presented in [16].

Also, optical flow computation is one of the fundamental tasks in computer vision. Deep learning methods allow efficient computation of optical flow, both in supervised learning on synthetic data [42], and in the self-supervised [39] setting. Additionally, in [40], the authors propose to highlight the independently moving object by compensating for the background motion, either by registering consecutive frames, or explicitly estimating camera motion. Another line of work has tackled the problem by explicitly leveraging the independence, in the flow field, between the moving object and its background. For instance, [41] proposes an adversarial setting, where a generator is trained to produce masks, altering the input flow, such that the inpainter fails to estimate the missing information.

Finally, in [43-45], two protocols have attracted increasing interest from the vision community, namely, semi-supervised video object segmentation (semi-supervised VOS), and unsupervised video object segmentation (unsupervised VOS). The former aims to re-localize one or multiple targets that are specified in the first frame of a video with pixel-wise masks, and the latter considers automatically separating the object of interest (usually the most salient one) from the background in a video sequence.

III. Proposed Method

To present our method, we consider an important hypothesis. We presume that the video contains at least some frames of CE such that it is useful for the improvement of the uncertainly-estimated frame outcomes. In the presented method, the main idea is the propagation of the credence of the frames of CE using the fine-tuning of DCNN. Therefore, our method includes the below stages: the election of the CE frames, the production of the label mapping, and the matching of a model with the input video. In the following subsections, we characterize the desired algorithm of these stages. Fig. 1 displays the general format of our presented approach.

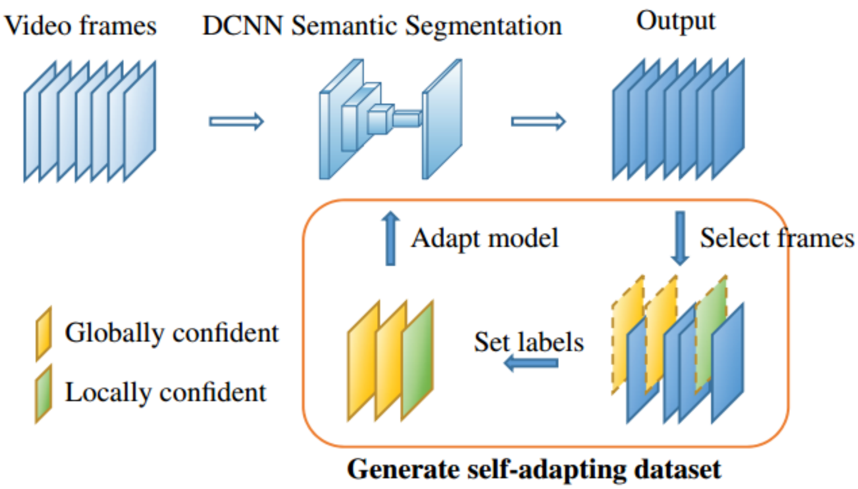

Fig. 1. The general format of our presented approach.

A. Steps of Our Presented Method

Fig. 1 displays the general steps of our presented method. In this sub-section, we will characterize the presented approach details to segment the motion objects in the video frames. F shows the collection of the indices of the frames; also W displays the collection of the desired poor labels from the input video. The presented method starts using the DCNN model θ which is pre-trained (for the frame $f \in F$), then we apply $Softmax$ for the computation of the probability $P(x_i|\theta)$ where i-th pixel is the organ of the class $x_i \in O$ which in it, O represents a collection of the classes of the object and the classes of the background. The mapping of the semantic label S can be measured by the use of the $argmax$ for each pixel i:

$$\mathrm{S(i)} = \arg\max_{xi} \mathrm{P}(x_i|\theta) \quad (1)$$

We gather the self-adaptive dataset G for adaption of the model of DCNN with the input video, which G includes the frames of CE and the related tag mappings. We gather the global CE frames and the local CE frames. Then we calculate corresponding label mappings G^g and G^l for the make of a self-consistent collection. Algorithm 1 summarizes the processes of the frame selection and the label calculation. First, we apply the analysis of the connected part in each mapping from the class S for the generation of the collection of the candidate object zones. For k-th mapping of $R_k \in \mathrm{R}$, the confidence of $C(R_k)$ measures the evaluated zones, which in it, the operator $C(\cdot)$ catches the label mapping as the input. Then, it calculates the mean probability of which pixels are labeled as objects. The label mapping has the labels of the related class.

In the following, we construct the mapping of the label G_f^g with the setting of the zone label when the confidence value trespasses an upper threshold t_0. Also, we adjust the label of the background for each pixel where $P(x_i = bg|\theta)$ (for being the background) is the greater than the threshold of t_b.

For the completion of G_f^g, the residual undefined zones must be processed. We, for this goal, let the residual pixels be labeled as "ignored." The pixels of the unspecified "ignored" are not attended to in the calculation of the loss value for the updation of the model. Also, we relinquish all pixels which have tags that are not on the collection of W. We surcharge the global frames of CE with G_f^g which have one safe zone for self-consistent dataset G.

Since the elected frames may be distributed temporally, our model can be overcome by the frames which are elected in a short time. For the reduction of the obtained error and for the regularization of the model, it is recommended to select the local frames of CE which have the best confidence of the object in each interval τb. We determine the local CE frames and their label mapping G^l as follows: For each frame f, we create a label mapping G_f^l using label keeping of total pixels if and only if $S(i)$ to be consisted on W, when we set the background as the prior. In the following, we compute the frame confidence by the computation of $C(G_f^l)$ and then we consider the frame with the highest confidence during each part of the frame τb as the local CE. Let the local frame of CE formerly not elected as the global frame of CE; then we surcharge it into the self-adaptive dataset G.

With the consideration of the self-adaptive dataset G which is computed by the mentioned processes, finally, we reconcile the model θ with the video. This task is done using the fine-tuning of the model into θ'. In the following, we calculate the novel label mapping using θ' for each frame.

Algorithm 1. Procedures for Selecting the Frames and Calculating the Labels

```
Input: DCNN model $\boldsymbol{\theta}$, a set of weak labels $\boldsymbol{W}$
Local best confidence $\boldsymbol{d = 0}$
for $\boldsymbol{f \in F}$ do
    Initialize $\boldsymbol{G_f^g}$ , $\boldsymbol{G_f^l}$ to ignored label
    Compute $\boldsymbol{P(x|\theta)}$ and $\boldsymbol{S = arg\,max_x\, P(x|\theta)}$
    Compute set $\boldsymbol{R}$ of connected components in $\boldsymbol{S}$
    for $\boldsymbol{R_k \in R}$ do
        if $\boldsymbol{S\ (i)\ \notin W, i \in R_k}$ then continue
        if $\boldsymbol{C(R_k)\ >\ t_0}$ then
            Set $\boldsymbol{G_f^g\ (i) = S\ (i), \forall i \in R_k}$
        Set $\boldsymbol{G_f^l\ (i) = S(i)\ , \forall i \in R_k}$
    Set $\boldsymbol{G_f^g\ (i) = G_f^l(i) = 0, \forall i\ st. P(x_i =\ bg|\theta) > t_b}$
    if $\boldsymbol{C(G_f^g) > 0}$ then $\boldsymbol{G \leftarrow G \cup \{G_f^g\}}$
    if $\boldsymbol{C(G_f^l) > d}$ then
        Update $\boldsymbol{t = f}$ and $\boldsymbol{d = C(G_f^l)}$
    if $\boldsymbol{f\ mod\ \tau b = 0}$ then
        if $\boldsymbol{G_t^g \notin G}$ then
            $\boldsymbol{G\ \leftarrow\ G\ \cup\ \{G_t^l\}}$
        Initialize $\boldsymbol{d = 0}$
Fine-tune DCNN model $\boldsymbol{\theta}$ to $\boldsymbol{\theta}\ '$using the set $\boldsymbol{G}$
```

B. Development of Proposed Method for Un-Supervised Video

Our presented approach can be used for the processing of unsupervised video. This task can easily be applied to unsupervised videos using the limitation of line eight in Algorithm 1. This omission means that the model doesn't manage whether the class emerges really on the video. So, we adjust all tags of the CE zones even if the tags are wrong. The experiments show that most processed videos have the same outcomes as the weakly-supervised videos is so much that the pixel tags specified by a great probability typically match the true tags. However, exceptions occur, which these exceptions are related to incorrect labels. They can degrade the model, and they can reduce the accuracy compared to the settings of weakly supervised.

C. Implement the Post-Processing for the Correction

Since the DCNN output is not sufficient to accurately characterize the object therefore, we apply the fully-connected CRF [30]. We apply the DCNN output for the single expression. Also, we apply the pixel's positions and the pixels' colors for the calculation of the even expressions (similar to [31]). We, finally, modify the label mapping via the practices of morphology (such as erosion and dilation).

IV. Tests and the Outcomes Evaluation

In this sub-section, first, we will present the implementation details and the performed tests. Also, we will introduce the used dataset. In the following, the tests' results are presented, and an analytical evaluation is done.

A. Details of Implementation

The tests in this article are performed as the single-shot and the zero-shot. Also, the designed tests are done using two datasets: YouTube-VOS [32] and DAVIS-2017 [33]. The first dataset, YouTube-VOS, contains 474 films on the set of the validation and 3471 films on the set of the training. It is the biggest dataset in the field of the segmentation of the video object. In addition, the training dataset contains 65 unique groups of the object, which are considered as the observed groups. Also, in the dataset of the validation, there are 91 groups of the object that consist of 26 unseen groups and all seen groups.

On the other hand, the dataset of DAVIS-2017 includes 60 films for the training dataset, 30 films for the validation dataset, and 30 films for the test dataset. In both datasets, the videos contain several objects, and their duration is between three to six seconds. The Python programming language has been used for the implementation of these tests. The presented method is implemented in a machine with Core (TM) i7 CPU 3.0 GHz Intel(R) and 8G RAM. The convolutional network is implemented on GPU, and the used graphic card in this method is NVIDIA GEFORCE 840M. The tests are analyzed by the use of the normal analysis criteria: (1) the accuracy of the contour F and (2) the similarity of the region J. On the YouTube-VOS dataset, these criteria are divided into two sub-criteria, depending on whether groups already have been seen with a network (F_{seen} and J_{seen}) or have not been seen by the model (J_{unseen} and F_{unseen}). The concept of the seen (or the unseen) means that, these categories are included in the set of training (or are not included).

B. Experiments and Results for the YouTube-VOS Dataset

As mentioned, the tests and the results are presented in two modes: the single-shot and the zero-shot. The single-shot mode consists of the object segmentation of a video according to the mask of the objects on the initial frame. But zero-shot mode involves the video objects segmentation without the previous data about that which of the objects must be segmented. It means that no object mask is obtained. This work is more complicated than the single-shot mode because the network must identify and then segment the objects that appeared in the film.

Table I shows the obtained results on the validation dataset of YouTube-VOS for the single-shot mode. All presented models in this study were trained using an 80-20 split for the training dataset. Fig. 2 shows some qualitative results, which in it, we can view, which our proposed method better maintains the segmentation of the objects over time. The proposed network can learn how the fixing the faults that may arise in the deduction. Table I can view which this approach is strong and has a suitable performance. Fig. 3 displays some qualitative outcomes which compare our trained approach over the mask of the ground truth and our trained approach over the concluded mask.

Table II displays the comparison of our presented model and similar approaches using the entire training dataset of YouTube-VOS. As it is clear, our proposed model has analogous outcomes with the mentioned model in [32]. The proposed method has an awhile worse turnover for the similarity of the region J. However, it has an awhile better turnover for the accuracy of the contour F. The proposed network performs better than the remaining advanced methods [25], [34]–[36] for the observed categories. Also, depending on the number of examples in the videos, Table III displays the related outcomes to the similarity of the region J and the accuracy of the contour F. We can view which objects for segmentation be fewer, then the work is easier, and we get better outcomes for the trails with only one or two annotated objects. Fig. 4 displays the qualitative outcomes for our presented approach for different trials from the validation set of YouTube-VOS. It contains the samples by the different samples number. Note that which samples are segmented correctly. However, there are different samples of a similar group on the video trail (the leopard, the sheep, the bird, the fish, or the person), or there are cases that vanish from the trail (a sheep on third row and a dog in fourth row).

So far, we have presented the test outcomes for our presented approach in the single-shot mode for the YouTube-VOS dataset. Also, we provide the outcomes for the mentioned dataset on the zero-shot mode. It should be noted that today, there is no designed special dataset for zero-shot segmentation. Although the YouTubeVOS dataset and the DAVIS-2017 dataset can be used to train and evaluate the models without the use of the provided annotations in initial frame, these datasets have this restriction in which the total appeared objects on the film are not annotated. In the YouTube-VOS dataset, specifically, a maximum of five object instances in each video are annotated. This makes sense when the objects are given for the segmentation but may be problematic for the zero-shot segmentation because the model can correctly segment the objects which are not annotated on the dataset. Fig. 5 displays some samples in it and some annotations of the missing object. Notwithstanding the stated quandary on the annotations of the missing object, for this mode, we have trained our network using the existing object annotations in this dataset.

Table IV displays the obtained outcomes for the validation dataset of YouTube-VOS on the segmentation with the zero-shot. Similar to the segmentation problem in single-shot mode, the proposed model has a good performance for object segmentation in the video. Fig. 6 displays some outcomes for segmentation in the zero-shot mode on the validation dataset of YouTube-VOS. Note that the masks are not obtained, so the system must detect the objects that must be segmented.

C. Experiments and Outcomes for the DAVIS-2017 Dataset

We test our pre-trained model (which is trained using the YouTube-VOS dataset) on a different dataset: DAVIS-2017. As shown in Table V, if the pre-trained network is done directly for the DAVIS-2017 dataset, our presented approach performs better than the rest of the approaches that do not use online learning. In addition, when the proposed model is adjusted for the training dataset of DAVIS-2017, the proposed method outperforms some methods (for example, OSVOS [3]). Fig. 7 displays the obtained visual outcomes on the dataset of DAVIS-2017 in the single-shot mode.

But in the zero-shot mode, by reviewing the articles and the research literature, it was found that there are no formal outcomes for the zero-shot mode on the DAVIS-2017 dataset so that we can compare our proposed method with it. The segmentation with zero-shot mode only is remarked for the DAVIS-2016 dataset, which in the unsupervised approaches have been made for it. Using the YouTube-VOS dataset on the zero-shot mode, if our model is directly done on the DAVIS-2017 dataset, our pre-trained model yields an average region similarity equal to $J = 22.4$ and an average contour accuracy equal to $F = 28.0$. If this pre-trained network is fine-tuned using the validation dataset of DAVIS-2017, then it results in a while better efficiency: $F = 30.6$ and $J = 24.7$. This poor efficiency of the zero-shot segmentation on the DAVIS-2017 dataset can be illustrated by the poor efficiency of the YouTube-VOS dataset in the unseen groups. Fig. 8 displays the visual outcomes of the test dataset of DAVIS-2017, which in it the mask of the object is not obtained.

D. Discussion

The results obtained from the qualitative experiments showed that our proposed method maintains the segmentation of the objects as better over time. This is because the proposed network can learn how to fix the errors that may occur in the deduction. Also, the results showed that when we test our pre-trained model on a different dataset, our proposed approach outperforms other approaches that do not use the online learning. Furthermore, when the proposed model is adjusted to DAVIS-2017 training dataset, the proposed method outperforms other approaches. Also, the various tests on YouTube-VOS showed that if our model is run directly on the DAVIS-2017 dataset, our pre-trained model yields an average regional similarity of J=22.4 and an average contour accuracy of F= 28.0. If this pre-trained network is fine-tuned by using the DAVIS-2017 validation dataset, it gives the better performance for some time: F=30.6 and J=24.7. This poor performance of the zero-shot segmentation on the DAVIS-2017 dataset can be illustrated by the poor performance of the YouTube-VOS dataset on the unseen groups. Finally, the presented method in this article has a specific limitation that occurs sometimes. This limitation occurs when a video does not match the hypothesis that we stated at the beginning (at least one zone of the object collates with the classes of the pre-trained object, or at least one frame has the true tag). Mostly, these samples happen due to the size of very small the objects on a video. The lack of a frame for improvement of other frames gives similar outcomes to the base model. The researchers can remark on this limitation in their subsequent works.

TABLE I. OBTAINED RESULTS FOR VALIDATION DATASET OF YOUTUBE-VOS IN THE SINGLE-SHOT MODE

YouTube-VOS Dataset in the Single-Shot Mode				
	J_{seen}	J_{usseen}	F_{seen}	F_{unseen}
Proposed Method	64.0	45.1	67.9	51.2

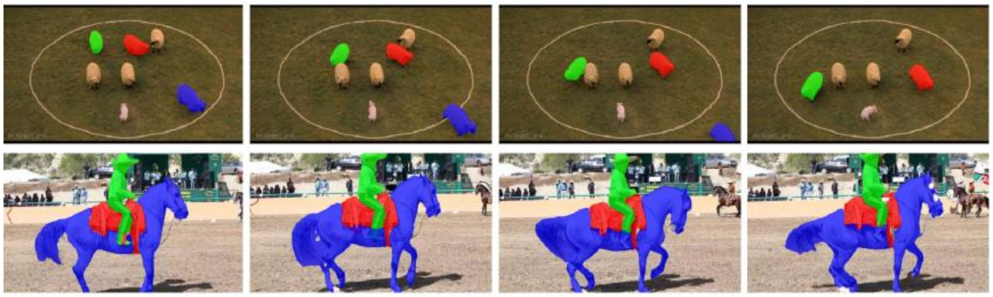

Fig. 2. Qualitative results for our presented approach in a single-shot mode for the YouTube-VOS dataset.

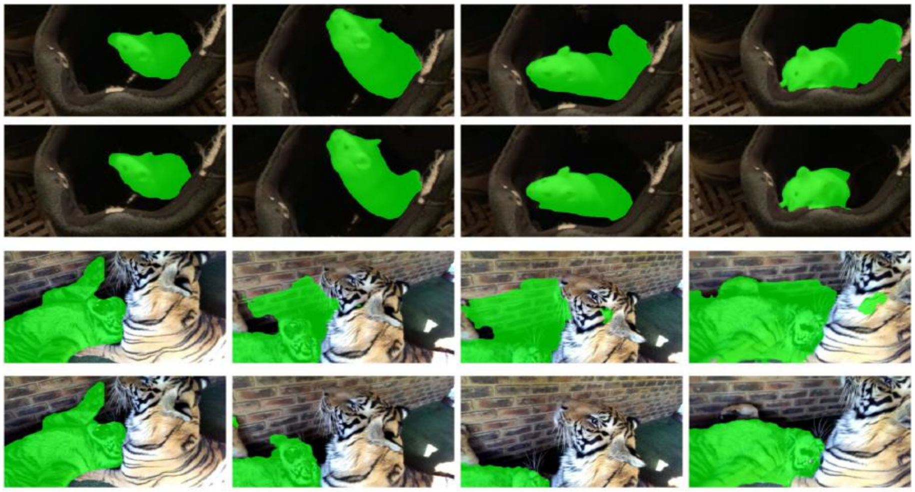

Fig. 3. Some comparative qualitative outcomes for our presented approach on the single-shot mode for the YouTube-VOS dataset.

TABLE II. COMPARISON OF OUR PRESENTED APPROACH WITH THE ADVANCED APPROACHES FOR THE SINGLE-SHOT MODE ON THE VALIDATION SET OF YOUTUBE-VOS. THE TERM OL REFERS TO ONLINE LEARNING

YouTube-VOS Dataset in Single-Shot Mode					
	OL	J_{seen}	J_{usseen}	F_{seen}	F_{unseen}
OSVOS [3]	Yes	59.8	54.2	60.5	60.7
MaskTrack [25]	Yes	59.9	45.0	59.5	47.9
OnAVOS [35]	Yes	60.1	46.6	62.7	51.4
OSMN [36]	No	60.0	40.6	60.1	44.0
S2S w/o OL [32]	No	66.7	48.2	65.5	50.3
Proposed Method	No	64.8	45.4	68.4	51.9

TABLE III. ANALYSIS OF OUR PRESENTED APPROACH DEPENDING ON THE NUMBER OF SAMPLES IN THE SEGMENTATION OF THE SINGLE-SHOT

Number of Samples (YouTube-VOS)					
	1	2	3	4	5
J_{mean}	78.9	63.3	51.2	50.6	56.9
F_{mean}	76.1	68.1	56.4	62.9	66.8

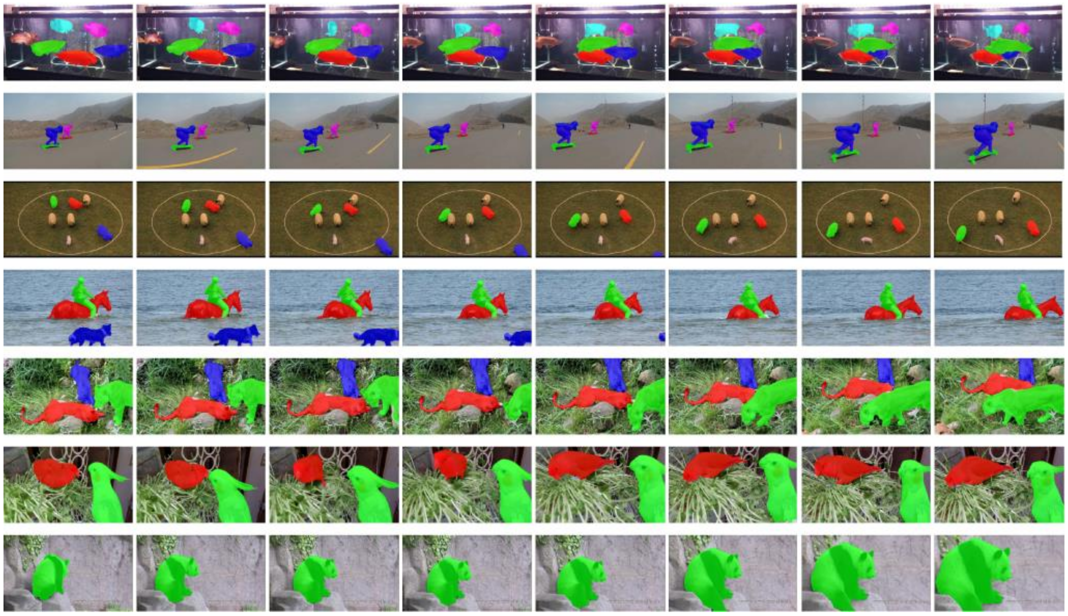

Fig. 4. Some visual results for our presented approach for the different trials from the YouTube-VOS validation set.

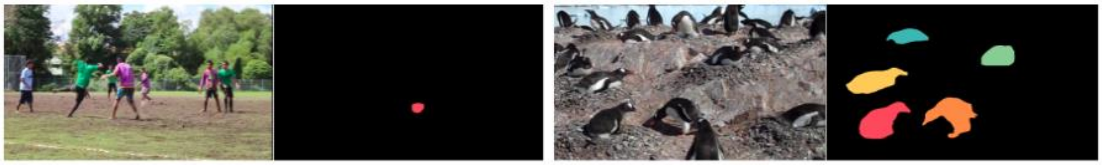

Fig. 5. The examples of the missing object annotations.

TABLE IV. THE RESULTS OF THE PERFORMED TESTS FOR THE SEGMENTATION WITH THE ZERO-SHOT MODE FOR THE DATASET OF YOUTUBE-VOS

YouTube-VOS Dataset in the Zero-Shot Mode				
	J_{seen}	J_{usseen}	F_{seen}	F_{unseen}
Proposed Method	45.2	24.1	45.9	24.2

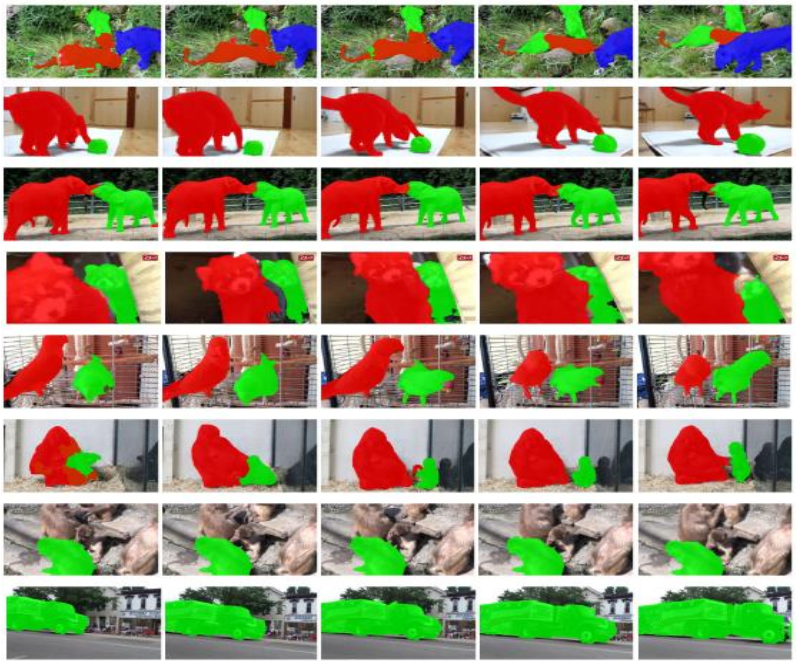

Fig. 6. Qualitative outcomes for segmentation with the zero-shot mode for the dataset of YouTube-VOS.

TABLE V. COMPARISON OF OUR PRESENTED APPROACH OVER ADVANCED METHODS FOR SEGMENTATION IN THE SINGLE-SHOT MODE IN THE DAVIS-2017 DATASET. NOTE THAT OL REFERS TO ONLINE LEARNING

DAVIS-2017 Dataset in the Single-Shot Mode			
	OL	**J**	**F**
OSVOS [3]	Yes	47.0	54.8
OSVOS-S [37]	Yes	52.9	62.1
CINM [38]	Yes	64.5	70.5
OnAVOS [35]	Yes	52.9	62.1
OSMN [36]	No	37.7	44.9
FAVOS[4]	No	42.9	44.2
Proposed Method	No	48.8	53.3

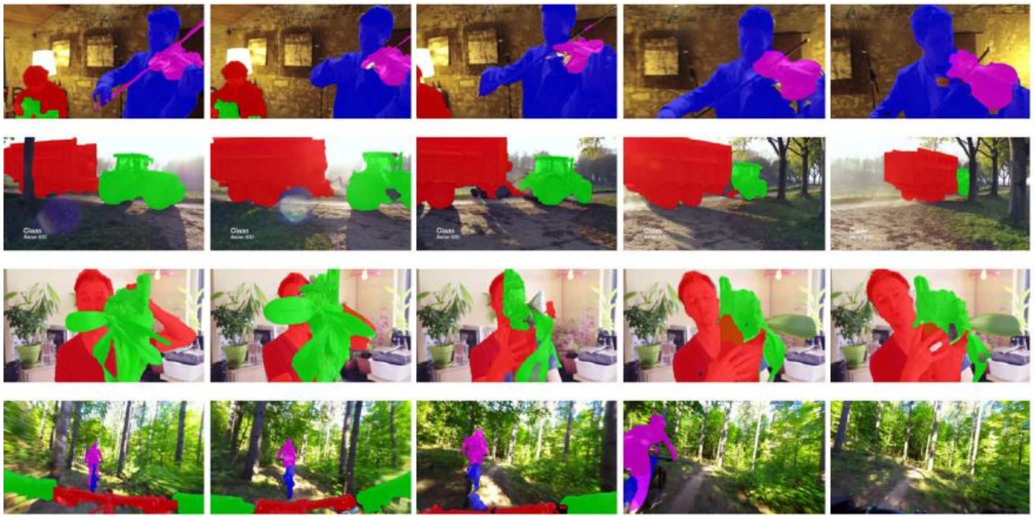

Fig. 7. Visual outcomes of the single-shot segmentation on the dataset of DAVIS-2017.

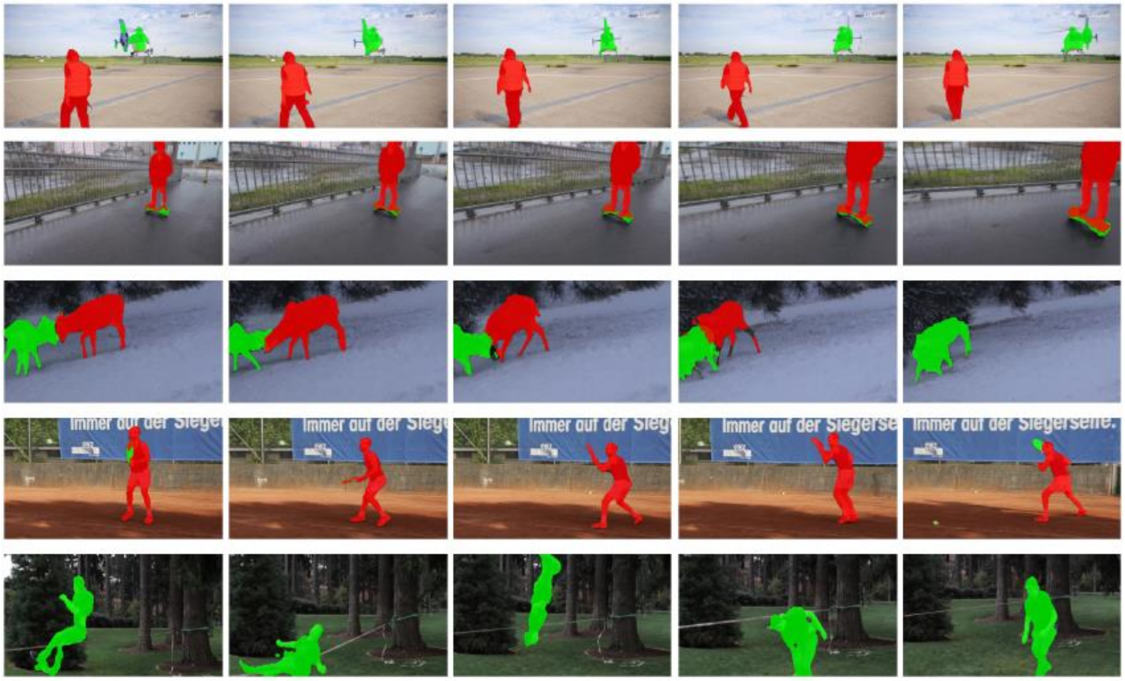

Fig. 8. Visual outcomes of the test dataset of DAVIS-2017 in the zero-shot segmentation mode.

V. CONCLUSIONS AND SUGGESTIONS

In our article, we propose a novel method for the segmentation of the motion objects which exist in the video. This method reconciles the model of the pre-trained DCNN with the input film. For the fine-tuning of the model, which is trained as vastly to be special for the video, we created a self-adaptive set that includes multiple frames, which helps to the improvement of the results of the frames of UE. This model is designed for segmentation in the single-shot mode and the zero-shot mode. Also, this model is applied in the YouTube-VOS dataset and the DAVIS-2017 dataset. The tests display that our trained model has a better performance than similar methods. In addition, our model improves the performance of similar methods. For future research, it is suggested to develop a semi-supervised film framework for the accuracy increase. It can also be expected that this efficient self-adaptive method can generate video datasets with accurate labels.

ACKNOWLEDGMENT

This work was supported by “Sponsored by Natural Science Foundation of Xinjiang Uygur Autonomous Region”(2022D01C461), “Research on UAV moving target detection and tracking system based on computer vision”(ZZ202105).

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” Acm computing surveys (CSUR), vol. 38, no. 4, pp. 13-es, 2006.

[2] J. K. Aggarwal and Q. Cai, “Human motion analysis: a review. Nonrigid and Articulated Motion Workshop, 1997,” Proceedings., IEEE, pp. 90–102, 1997.

[3] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, “One-shot video object segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 221–230.

[4] J. Cheng, Y.-H. Tsai, W.-C. Hung, S. Wang, and M.-H. Yang, “Fast and accurate online video object segmentation via tracking parts,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7415–7424.

[5] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, “Segflow: Joint learning for video object segmentation and optical flow,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 686–695.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[7] Y.-T. Hu, J.-B. Huang, and A. Schwing, “Maskrnn: Instance level video object segmentation,” Adv Neural Inf Process Syst, vol. 30, 2017.

[8] H. P. Moravec, “Visual mapping by a robot rover,” in Proceedings of the 6th international joint conference on Artificial Intelligence-Volume 1, 1979, pp. 598–600.

[9] C. Harris and M. Stephens, “A combined corner and edge detector,” in Alvey vision conference, Citeseer, 1988, pp. 10–5244.

[10] J. Shi, “Good features to track,” in 1994 Proceedings of IEEE conference on computer vision and pattern recognition, IEEE, 1994, pp. 593–600.

[11] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int J Comput Vis, vol. 60, pp. 91–110, 2004.

[12] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” IEEE Trans Pattern Anal Mach Intell, vol. 27, no. 10, pp. 1615–1630, 2005.

[13] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” IEEE Trans Pattern Anal Mach Intell, vol. 27, no. 10, pp. 1615–1630, 2005.

[14] C. Stauffer and W. E. L. Grimson, “Learning patterns of activity using real-time tracking,” IEEE Trans Pattern Anal Mach Intell, vol. 22, no. 8, pp. 747–757, 2000.

[15] R. Zhang and J. Ding, “Object tracking and detecting based on adaptive background subtraction,” Procedia Eng, vol. 29, pp. 1351–1355, 2012.

[16] C. Hua, H. Wu, Q. Chen, and T. Wada, “Object tracking with target and background samples,” IEICE Trans Inf Syst, vol. 90, no. 4, pp. 766–774, 2007.

[17] C. Hua, H. Wu, Q. Chen, and T. Wada, “K-means Tracker: A General Algorithm for Tracking People.,” J. Multim., vol. 1, no. 4, pp. 46–53, 2006.

[18] C. Hua, H. Wu, Q. Chen, and T. Wada, “K-means clustering based pixel-wise object tracking,” Information and Media Technologies, vol. 3, no. 4, pp. 820–833, 2008.

[19] P. Ochs, J. Malik, and T. Brox, “Segmentation of moving objects by long term video analysis,” IEEE Trans Pattern Anal Mach Intell, vol. 36, no. 6, pp. 1187–1200, 2013.

[20] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, “Learning video object segmentation from static images,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2663–2672.

[21] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 724–732.

[22] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, “The 2017 davis challenge on video object segmentation,” arXiv preprint arXiv:1704.00675, 2017.

[23] M. Ren and R. S. Zemel, “End-to-end instance segmentation with recurrent attention,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6656–6664.

[24] B. Romera-Paredes and P. H. S. Torr, “Recurrent instance segmentation,” in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14, Springer, 2016, pp. 312–329.

[25] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, “Learning video object segmentation from static images,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2663–2672.

[26] O. Russakovsky et al., “Imagenet large scale visual recognition challenge,” Int J Comput Vis, vol. 115, pp. 211–252, 2015.

[27] A. Salvador et al., “Recurrent neural networks for semantic instance segmentation,” arXiv preprint arXiv:1712.00617, 2017.

[28] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, “Pyramid dilated deeper convlstm for video salient object detection,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 715–731.

[29] P. Tokmakov, K. Alahari, and C. Schmid, “Learning motion patterns in videos,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3386–3394.

[30] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” Adv Neural Inf Process Syst, vol. 24, 2011.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” arXiv preprint arXiv:1412.7062, 2014.

[32] N. Xu et al., “Youtube-vos: A large-scale video object segmentation benchmark,” arXiv preprint arXiv:1809.03327, 2018.

[33] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, “The 2017 davis challenge on video object segmentation,” arXiv preprint arXiv:1704.00675, 2017.

[34] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, “One-shot video object segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 221–230.

[35] P. Voigtlaender and B. Leibe, “Online adaptation of convolutional neural networks for video object segmentation,” arXiv preprint arXiv:1706.09364, 2017.

[36] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, “Efficient video object segmentation via network modulation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6499–6507.

[37] K.-K. Maninis et al., “Video object segmentation without temporal information,” IEEE Trans Pattern Anal Mach Intell, vol. 41, no. 6, pp. 1515–1530, 2018.

[38] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5977–5986.

[39] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In Proc. CVPR, 2020.

[40] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. Proc. ACCV, 2020.

[41] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In AAAI, volume 34, pages 13066–13073, 2020.

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Proc. ECCV, 2020.

[43] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. International Journal of Computer Vision, 2019.

[44] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In Proc. CVPR, 2019.

[45] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In Proc. ICCV, 2019.