Pop-net: A self-growth network for popping out the salient object in videos

Hui Yin\textsuperscript{1} | Ning Chen\textsuperscript{1} | Lin Yang\textsuperscript{2} | Jin Wan\textsuperscript{1}

\textsuperscript{1}Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
\textsuperscript{2}Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University, Beijing, China

Abstract

It is a big challenge for unsupervised video segmentation without any object annotation or prior knowledge. In this article, we formulate a completely unsupervised video object segmentation network which can pop out the most salient object in an input video by self-growth, called Pop-Net. Specifically, in this article, a novel self-growth strategy which helps a base segmentation network to gradually grow to stick out the salient object as the video goes on, is introduced. To solve the sample generation problem for the unsupervised method, the sample generation module which fuses the appearance and motion saliency is proposed. Furthermore, the proposed sample optimization module improves the samples by using contour constraints for each self-growth step. Experimental results on several datasets (DAVIS, DAVSOD, VideoSD, Segtrack-v2) show the effectiveness of the proposed method. In particular, the state-of-the-art methods on completely unfamiliar datasets (no fine-tuned datasets) are performed.

1 | INTRODUCTION

Video object segmentation (VOS) refers to segment the object from the background in a given video sequence. It is of great significance for higher level visual tasks, such as object retrieval, object tracking, video summarisation and video understanding. VOS can be separated into supervised, semi-supervised and unsupervised segmentation according to whether annotations or other information about the object is given. Unsupervised VOS needs to separate the foreground object from background without any prior knowledge, which makes it a very challenging problem. In this article, we focus on unsupervised video object segmentation.

Research on unsupervised VOS can be divided into two categories: non-learning based methods and learning-based methods. For the former, most methods generated object proposals for the whole video sequence, and evaluated each proposal score. The proposal with the highest score is considered as the salient object [1, 2]. In addition, several methods obtain the object through saliency detection [3]. [3] suggest detecting salient objects based on selective contrast. For the latter, machine learning based methods have achieved significant results. For example, [4] proposed a multi-instance learning framework that included low-, mid- and high-level features in the detection process. Generally, the popular methods tend to learn from offline training to combine appearance and motion information to segment objects [5–9]. The characteristics [5–7,9] have in common is that a two-stream network is used to combine motion and appearance information, and [7, 9] respectively fused optical flow and static information to the segmented network. Although these learning-based methods have the potential to take advantage of both static and dynamic cues, there are still some disadvantages. (1) Most methods use groundtruth to train the network during the training phase, which is not completely unsupervised VOS in the true sense. And when testing on a new video sequence (unfamiliar datasets without groundtruth fine-tuned), their segmentation performance deteriorate quickly. (2) The global information of the video sequence is not fully utilised. Some methods propagate the characteristics of the object as the video stream goes on, but the information in the latter frames is not served for the segmentation in the former frames. Also, recent research show that adding global information is effective for improving the segmentation results [10–13].

To tackle above problems, we propose a novel self-growth video salient object segmentation network based on appearance and motion information, called Pop-Net. On the one
hand, unlike the current unsupervised methods which use groundtruth for training network (calculating loss), Pop-Net is completely unsupervised and does not require any groundtruth when training the network (even when calculating loss). On the other hand, different from the general methods which segment a new video directly by offline model or online learning, we focus on the fact that the most salient object appears repeatedly in the whole video stream. So, it is necessary for the network to explore the video content itself effectively, and the information highlighting the object should be captured by the network step by step. Therefore, we divide the automatic salient object segmentation task into two steps. 1. Self growth step — By introducing competition, the salient regions in each video frame are continuously competing. In this way, only the most salient objects will be continuously strengthened and the rest will be gradually suppressed. With the development of the video stream, the segmentation network has gradually grown from focusing on general objects to focusing on the most salient object in the video. 2. Segmentation step— When the video stream is ended, learning is completed. The final ‘mature’ segmentation network is used to segment the salient object in each frame from beginning to end.

As shown in Figure 1, Pop-Net includes three parts: the self-growth segmentation module, the sample generation module, and the sample optimization module. The base network in the self-growth segmentation module could be any segmentation network with any initial parameters. For each frame in the video sequence, the results of the self-growth segmentation module and the sample generation module are combined to obtain the initial training samples. Then the training samples optimised by the sample optimization module are used to train the segmentation network. In this manner, as the video stream advancing, the base network continues to learn the characteristics of the most salient object in the video and shifts attention to it. The proposed self-growth strategy makes full use of the information of the video to aid the base network to gradually grow into a segmentation network for a specific object.

In summary, our contributions are as follows:

1. We formulate a self-growth unsupervised video salient object segmentation network based on appearance and motion saliency (Pop-Net). Specially, we propose a novel self-growth strategy that allows the base network to grow from a general segmentation network to a specific object segmentation network without any manual labelling.

2. To solve the training sample problem, the sample generation module is proposed. Specifically, during the processing of each frame, a mask-level fusion strategy is used to fuse the appearance and motion saliency to generate training samples. In this way, the base network would be continuously modified and improved.

3. Aiming at the problem of sample inaccuracy, the sample optimization module is proposed. In particular, by performing contour constraint and non-principal suppression for each step of self-growth, difficult positive samples are found continuously. Meanwhile, pseudo positive samples are removed.

4. To verify the proposed self-growth strategy completely unsupervised, we conduct comparative experiments on four datasets (one familiar dataset and three unfamiliar datasets) respectively. The experimental results show that on the familiar dataset (fine-tuned) DAVIS 2016 [14], our network is quite close to the state-of-art methods. And on unfamiliar datasets (no fine-tuned): DAVSOD [12], VideoSD [41] and SegTrack-v2 [16], the proposed Pop-Net shows superior performance over recent methods.

The remainder of this article is organised as follows: In Section 2, the classic unsupervised object segmentation methods and the use of motion and appearance in...
unsupervised video object segmentation are reviewed separately. In Section 3, we elaborate on the proposed method. In Section 4, we introduce network implementation and training. Experiments are presented and analysed in Section 5. In Section 6, we present conclusion and suggest aspects of the proposed work that could be developed further.

2 | RELATED WORK

2.1 | Unsupervised video object segmentation

Unsupervised video segmentation does not need manual annotation or user input and could automatically segment the salient object in the video. Early unsupervised VOS methods were based on geometry information [17, 18]. By building the background model, they could find the deviation from the model for each input frame to achieve salient object segmentation. Barnich O et al. [18] proposed a background subtraction algorithm based on samples, which stored a set of values obtained in the same location or adjacent area for each pixel in the past, and then compared them with the current values to determine whether the pixel belongs to the background or not. However, these methods explore motion information in a short period of time, and the performance of the object is poor when it remains static in some frames. To explore the long-term motion cues of the object, long-term point trajectory methods were proposed [19–21]. Long-term point trajectory-based methods mostly used long-range trajectory motion similar for VOS. Lezama J et al. [21] developed an efficient spatio-temporal video segmentation algorithm, which naturally incorporated long-range motion cues from the past and future frames in the form of clusters of point tracks with coherent motion. But these methods do not consider static information. Later, to obtain accurate segmentation results, ‘object-like’ methods were proposed [1, 2, 8, 22–26]. ‘Object-like’ methods generally guide segmentation by generating object proposals or saliency information. Koby J et al. [1] used colour and motion edges to generate candidate regions for the primary object, and used the recursiveness of the primary object to estimate the initial primary object region. [26] introduced a salient object segmentation method by combining a saliency measure with a conditional random field (CRF) model.

In the recent years, due to the rapid development of convolutional neural network in the field of computer vision, video object segmentation based on deep learning has attracted much attention [27–30]. Wang Y et al. [27] established the spatio-temporal consistency model of video sequence by capturing the background correlation of adjacent frames, and used the motion information to select proposals. Song H et al. [28] used ConvLSTM to segment video salient objects. They designed a pyramid expansion convolution (PDC) module to extract spatial features of multiple scales, and then fed these features into the extended deep two-way ConvLSTM to learn spatio-temporal information. Siyang I. et al. [29] proposed a segmentation method by transferring the knowledge encapsulated in image-based instance embedding networks. Most of these methods used groundtruth in the training phase of the network. Distinguishing them, [30] solved the problem of learning object patterns in an unlabelled video, and introduced an unsupervised learning framework by comprehensively capturing the intrinsic properties of video object segmentation.

FIGURE 2  Schematic diagram of pre- and post-optimization object sample based on contour. (a) Segmentation results before contour constraint. (the red pixel $e_4$ is a boundary point of sample region) (b) Superpixel result. ($e_4$ is located inside a superpixel block) (c) Segmentation results after optimization based on superpixel contour constraints.
at multiple granularities. In this article, we focused a completely unsupervised VOS where the segmentation network learns from the video without any manual annotation.

2.2 | Motion and appearance in unsupervised video object segmentation

As there is no idea about the object in unsupervised VOS, motion and appearance information become the main basis for segmentation. Griffin A et al. [31] used motion saliency and visual saliency to generate object proposals. Leey J et al. [2] used static and dynamic prompts to detect persistent ‘object-like’ regions, and then used these regions to estimate the complete video segmentation. All of these methods usually need to operate on the whole video sequence, which not only consumes too much memory but also leads to high computational complexity. Wenguan W et al. [15] incorporated saliency as prior for object via the computation of robust geodesic measurement. They considered spatial edges and temporal motion boundaries as indicators of foreground object locations. Papazoglou A et al. [32] used optical flow to generate moving boundary, and then generated the moving region. However, these methods are very sensitive to the extraction of the moving boundary, and take less consideration of the static information of the object.

In deep learning network, appearance and motion information of the object have been used for unsupervised VOS [5–10, 33, 34]. [5–7, 9] proposed a two branch network combining appearance and motion, to jointly exploit the appearance and motion feature of the object. Tokmakov P et al. [5] trained a network to combine the appearance flow describing the static feature of the object in the video with the time flow capturing the motion clues to generate the final prediction results. Jain D et al. [6] proposed a two-stream FCNN that integrated motion and appearance into a unified framework, to jointly learn object segmentation and optical flow generation. In addition, Zhou T et al. [8] proposed a moving object segmentation framework, by fusing motion saliency and object proposals to improve segmentation results. Xiankai L et al. [10] proposed a co-attention siamese network (COSNet), which used the appearance information of the object to train the network and solved the task of unsupervised VOS from an overall perspective. [34] proposed a bilateral network to estimate the background based on the motion pattern of non-object regions. In this article, instead of considering only static and dynamic information, we try to use the whole information of the video to improve the segmentation ability of the base network by the proposed self-growth strategy.

3 | METHOD

The overall architecture of the Pop-Net is illustrated in Figure 1. There are three modules in our network, including the self-growth segmentation module (S), the sample generation module (G), and the sample optimization module(O).

3.1 | Self-growth segmentation strategy

For learning-based unsupervised VOS, the most difficulty lies in how to obtain effective training samples, especially the samples of the segmented object. Although models with generalisation capability could be trained extensively offline, it is not enough for new videos. Thus, learning and utilising the information of video is particular important. In order to make full use of the rich information of video streams, we propose the self-growth strategy. We emphasize the fact that salient object appears repeatedly in the video stream, so it is possible for the salient object regions in each video frame to be strengthen through continuous competition. In this way, the proposed self-growth base network improves its segmentation ability by continuously learning the information in the video stream.

Specifically, for an input frame, the self-growth segmentation module and the sample generation module simultaneously predict and output the initial training samples, then perform weighted fusion and optimise the initial training samples to obtain the training samples of the salient object. Finally, using the optimised training samples to train the segmentation network, the attention of the network is shifted to the salient object in the whole video.

By self-growth, the most salient object in video stream is continuously strengthened and sticks out, and the initial segmentation network is gradually biased towards the most salient region. Thus, after going through the whole video stream, the basic segmentation network grows into a salient object segmentation network (see Algorithm 1).

Algorithm 1 Pop-Net Segmentation Procedure

**Input:** segmentation net $S$, appearance branch $A$, motion branch $M$, $F_1…F_T$, $N$, $\alpha$, $\beta$, $\gamma$, $r$, $\delta$

**Output:** segmentation result: $M_1…M_T$

1: **Self-growth step:**
2: For $i \in 1…T$
3:    (Initial training samples generation)
4:    $s_i$ ← prediction ($S$, $F_i$)
5:    $a_i$ ← prediction ($A$, $F_i$)
6:    $m_i$ ← prediction ($M$, $F_i$)
7: $Z_i$ ← $\alpha*a_i + \beta*a_i + \gamma*m_i$
8:    (Refined segmentation based on contour constraint)
9:     $P_i$ ← superpixel($F_i$)
10: $E$ ← obtain boundary pixels of $Z_i$
11: For $k \in 1…K$
12:    if $E_k$ with a radius $r$ not fall on the boundary of any superpixel blocks then
13:        count the foreground pixels in the superpixel block which containing $E_k$
14:    if The proportion of foreground pixels $> \delta$ then
15:     region growth from $E_k$ on $Z_i$ according colour similarity
3.2 Sample generation module

The proposed Pop-Net is completely unsupervised, that is, no groundtruth is required during training (even when calculating loss) or testing. Therefore, in order to solve the problem of sample generation in unsupervised video object segmentation, we introduced the sample generation module. For the sample generation module, we adopt the auxiliary segmentation based on appearance saliency and motion saliency respectively.

Initially, the segmentation network has no prior knowledge about the input video and could only segment the general foreground objects in the video. Appearance saliency detection and motion saliency detection detect the saliency of the input video frame according to the appearance and motion information respectively. Optical flow is based on the change of pixels in time domain and the correlation between adjacent frames to determine the relationship between the pixels of continuous video frames, and then calculate the motion information of objects between adjacent frames. Therefore, we use the method in [32] as the motion saliency segmentation module, and improve the motion boundary in [32] according to the method of calculating the motion boundary mentioned in [35]. The appearance saliency module used the method in [36]. In the whole segmentation process, by using the salient region obtained from the detection as a guide, the most of the salient region would be gradually highlighted through continuous correction, and the segmentation network would eventually target the most salient object in the entire video sequence.

The self-growth segmentation network with general prospects is \( S \), the branch of appearance saliency detection is \( A \), and the branch of motion saliency detection is \( M \). Input video frame \( F_1 \ldots F_T \) is the total number of frames. \( S \) estimates the segmentation result of video frame \( F_i \) as \( s_i \), saliency detection result of \( A \) is \( a_i \), motion saliency detection result of \( M \) is \( m_i \). In the self-growth stage, the initial saliency results of the current frame are as follows:

\[
Z_i \leftarrow \alpha * s_i + \beta * a_i + \gamma * m_i \tag{1}
\]

\( \alpha, \beta, \gamma \) are weight factors, which are one-third respectively. \( s_i, a_i \) and \( m_i \) are all binary graphs, and the final \( Z_i \) is also obtained after binarisation. After weighting, \( Z_i \) is more inclined to the most salient region. The final salient candidate region is obtained by modifying the initial candidate region through the sample optimization module which is described next, and is used as training samples to modify \( S \).

3.3 Sample optimization module

Since the proposed Pop-Net does not require any labelling guidance, the choice of samples becomes particularly important. In addition, our goal is to gradually detect salient object, so we need to select the most reliable object area in each frame. However, the initial segmentation network is insufficient, and the results of the sample generation module may contain some background areas, which may seriously interfere with the learning progress of the segmentation network in the subsequent process. Thus, the sample optimization module is proposed to modify the training samples from the sample generation module.

Due to the roughness at the boundary of the object for the pixel-level segmentation, the contour of the object becomes a strong prior knowledge for unsupervised VOS. Therefore, the first step of the sample optimization module is the object correction algorithm based on the contour constraint. We introduce the contour constraint to optimise the significant region by adding the missing region under the contour constraint. The schematic diagram of pre- and post-optimization is shown in Figure 2 and the details are as follows.

The input of the sample optimization module is the superpixel map \( P_i (P_i = \{SE_j | SE_j \text{ is a superpixel block, } j = 1 \ldots J \}) [37] \) and the initial sample region \( Z_i \) of the frame \( F_i \). The contour constraint of the proposed method consists of two steps: Firstly, the boundary of the initial sample region is detected and the sample point set of the boundary is denoted as \( E, E = \{e_k | e_k \text{ is a boundary point of sample region } Z_i, k = 1 \ldots K \} \). Secondly, for each point \( e_k \in E \), if it falls on the boundary of any superpixel blocks, it is considered as an accurate contour point and no contour constrain optimization is required. Otherwise, \( e_k \in E \) is located inside a superpixel block \( SE_j \). If the maximum distance between \( e_k \) and any boundary points of \( SE_j \) is less than \( r \) or the proportion of foreground pixels in \( SE_j \) is less than the threshold \( \delta \), no contour constrain optimization is required (The reason is that the superpixel map may also have some errors for contour location). Otherwise, contour constrain optimization in \( SE_j \) is needed. The colour similarity between the foreground pixels in \( SE_j \) and the other pixels in \( SE_j \) (denoted by \( sp_j \), \( sp_j = \{sp_j^\text{f} | sp_j^\text{f} \text{ is a background pixel in } SE_j, \epsilon = 1 \ldots C \}) \) is computed. Only when the similarity between \( sp_j^\text{f} \) and the foreground pixel sets is bigger than 0.9 (calculated by the Euclidean Distance of colour histogram), that \( sp_j^\text{f} \) is merged into the sample region. By introducing contour constraints, the segmentation results are modified more accurately and some object regions (positive samples) are prevented from being suppressed all the time and not being learnt by the network.

After the contour constrained modification, the results may still have some discrete areas that do not belong to the segmented object, which may affects the learning of the network. In order to obtain more reliable samples, the discrete
point cluster optimization algorithm in [38] is used finally. In this way, the sample optimization module could provide more accurate salient object region to the segmentation network for learning by self-growth strategy.

4 | NETWORK IMPLEMENTATION AND TRAINING

In this section, we describe the implementation details of our approach, including network structure and training.

4.1 | Network implementation

As discussed above, the base network \( S \) could be any segmentation network with any initial network weights, because what does work is the proposed self-growth strategy which allows \( S \) to grow without any groundtruth. In experiments, we adopt the OSVOS [39] as the base network, which uses bilinear interpolation and transposed convolution to make the prediction and the input have the same size. At the same time, we used their pretrained parent network weights, which have the ability to segment common foreground objects, not specific to any specific object. During this pretraining, a random gradient descent algorithm with a momentum parameter of 0.8 was used to train 100,000 times. The initial learning rate was 10−8 and decays with training time. Specifically, for each video frame, we enhance the \( S \) by \( N = 50 \) update steps, and the learning rate is set to \( 10^{-8} \). For the proposal sample optimization stage, \( r = 5, \delta = 0.2 \). For the optimization of discrete point clusters, we use the weights provided by Hui Y et al. [38].

4.2 | Network training

Self-growth video salient object segmentation network consists of two steps: the self-growth step and the segmentation step. The first step aims to enhance the segmentation network by continuously acquiring salient regions, so that it can grow into a segmentation network for the most salient objects in the video. The second step is to segment the whole salient objects in the video sequence.

The description of the procedure is shown in Algorithm 1. The training and testing environment were performed under NVIDIA GTX 1080 Ti, and the segmentation speed for each frame did not exceed 0.9s. We do not conduct any further offline training on any dataset. This is a very obvious difference between this paper and other related work.

5 | EXPERIMENTS

In this section, we describe the evaluations, datasets, and study the quantitative importance of the different components of the proposed methods, and report results comparing the state-of-the-art techniques.

5.1 | Datasets and evaluations

We evaluate the proposed approach on current popular video single object segmentation dataset: DAVIS-2016 [14], which contains 50 high-quality video sequences with all frames annotated with pixel-wise object masks, and includes assorted challenges such as appearance change, occlusion, motion blur and shape deformation. Since DAVIS-2016 focuses on single-object video segmentation, each video has only one foreground object. In fact, each of them is the most salient object in each video sequence. There are 30 training and 20 validation videos. For accuracy, we use three evaluations: region similarity in terms of intersection over union (\( J \)), contour accuracy (\( F \)), and temporal instability (\( T \)) of the masks and we compare with a large set of state-of-the-art methods, including the recent unsupervised techniques: UOVOS [8], LVO [5], FSEG [6], SFL [7], TIS [31], and all of the evaluation results are provided by the DAVIS [14] benchmark.

To illustrate the effectiveness of the self-growth strategy proposed in Section 3.1, we do comparatives experiments on the datasets that have not been fine-tuned by the base network: DAVSOD [12], VideoSD [41] and SegTrack-v2 [16]. DAVSOD covers diverse realistic-scenes, different object appearances and motion patterns, various object categories, and contains 90 test video, including 35 easy subsets. VideoSD contains 10 video sequences under the natural scene, and SegTrack-v2 contains 14 video sequences. All of them have low resolution, which put forward higher requirements for the network segmentation ability.

5.2 | Ablation study On DAVIS

To analyse the importance of the different components mentioned in Section 3, Table 1 illustrates the impact of different components on the DAVIS validation set.

First, for each combination, we do experiments with different update steps \( N \). It can be seen from Table 1 that when the upstate steps \( N \) reach a considerable number of times, the proposed self-growth strategy can ensure that the segmentation network learns enough knowledge. At this time, the change of \( N \) does not have much impact on the segmentation ability of the network. The following description takes \( N = 50 \) as an example.

Then, we study the role of the sample optimization module. The second row of Table 1 shows the results of the lack of

| Method                          | \( N = 0 \) | \( N = 40 \) | \( N = 50 \) | \( N = 60 \) | \( N = 70 \) |
|--------------------------------|------------|------------|------------|------------|------------|
| Pop-net                        | -          | 0.710      | 0.717      | 0.713      | 0.715      |
| No optimization module         | -          | 0.645      | 0.637      | 0.628      | 0.622      |
| No self-growth strategy        | 0.602      | -          | -          | -          | -          |
| No motion saliency             | -          | 0.503      | 0.513      | 0.511      | 0.514      |
| No appearance saliency         | -          | 0.504      | 0.498      | 0.496      | 0.494      |

Notes: The best scores are marked in bold.
the sample optimization module. Compared to the method with the optimization module (Pop-Net), the performance is reduced by 8%. Figure 3 shows the qualitative evaluation of the sample optimization module in the ablation experiment. The first row represents the segmentation results of the complete segmentation network, and the second row represents the segmentation result of the unused sample optimization module. As shown in Figure 3, the sample optimization module could find difficult positive samples and remove pseudo positive samples which improve the reliability of the training samples, and prevent the network segment the regions not belong to salient object.

Next, we study the effectiveness of the self-growth strategy. The third line of Table 1 shows the results of the lack of the self-growth strategy. Compared with the complete network (Pop-Net), the score is reduced by 11.5%. Figure 4 shows the qualitative evaluation of the self-growth strategy in the ablation experiment. The first line represents the segmentation results of the Pop-Net, and the second row represents the segmentation results of the unused self-growth strategy. Comparison results show that the self-growth strategy can help the segmentation network gradually identify the saliency areas and exclude the background areas. Figure 5 shows the growth process of our segmentation network on ‘Scooter-black’ sequence. From the results of the training process, we observe that the initial segmentation network is not aimed at specific object, and segments all foreground objects. But as the video stream progresses, the segmentation network is growing gradually, from general object segmentation to salient object segmentation. In the segmentation step, the salient object in the entire video stream is segmented. The above results also confirm the effectiveness of the proposed self-growth strategy.

Finally, the fourth row and the fifth row of Table 1 represent the segmentation results using only the appearance saliency and only the motion saliency in the sample generation module, respectively. It can be seen from the results that both appearance and motion saliency are essential in the sample generation module. For the results only based on motion saliency, the score is only 49.8%. At this time, the segmentation network would only focussed on the motion areas. For the sequence of intense motion in the background area, the object would be lost. Similarly, only using the appearance saliency would segment too many foreground objects, and the result is 51.3%. The result obtained by the complete sample generation module is significantly better than that of the separate branch, and the segmentation accuracy is improved to 71.7% which shows that by fusing the results of the three parts (as shown in Equation 1), the salient regions learnt by the network will be more accurate.

In addition, to illustrate that the base network S could be any segmentation network with any initial network weights, we conduct experiments on a different initial network. Specifically, we choose SegTrack v2 dataset as the training set in the pre-training stage to obtain a new base network. For the new base network, there are three major challenges: (1) Compared with DAVIS 2016, that has 30 training set, SegTrack v2 has only 14 video sequences. The segmentation ability of the new base network is weak. (2) The video sequences in SegTrack v2 are all low-pixel, which puts higher forward requirements on the segmentation ability of the network. (3) For the new base network, DAVIS dataset is completely unfamiliar. Whether in the pre-training stage or the training stage of the proposed method, there is no information about the groundtruth. Then, we conduct experiments on the DAVIS val set. The score is 67.3% which shows that even using different network parameters and facing completely unfamiliar datasets, our method can still obtain better segmentation results.

5.3 | Comparison to the state of the art on DAVIS

Table 2 shows the overall results with the state-of-the-art unsupervised methods on DAVIS benchmark. In terms of the regional similarity of $J$, our Pop-Net scores 0.717, ranking
third in the state-of-the-art methods. Although there is a gap between this article and the best algorithm, Pop-Net is completely unsupervised, which differs from the methods that use groundtruth to guide the network during the training phases, such as LVO [5]. The advantages of our method is much more obvious on unfamiliar datasets (no fine-tuned datasets), which is reported next. Figure 6 shows the comparison results of regional similarity on each sequence with the methods which have some technical route as ours. It can be seen that our scores are very close to or even better than the state-of-the-art methods for most sequences. In particular, our Pop-Net get a higher score in the Blackswan, Drift-chicane and Horsejump-high sequences. Figure 7 present qualitative comparison of Pop-Net with the state-of-the-art methods in five sequences. For the Blackswan sequence, the water flow would move with the movement of the object which would cause the segmentation network to mistake the water flow as the salient object, such as TIS [31]. However, the proposed sample generation module would not only consider the motion information but also the appearance information of the object, that can effectively remove the background noise. Furthermore, our results is better than the recent methods UOVOS [8] which also used a combination of motion and static information. This is assisted by the fact that Pop-Net consider the most salient object in the entire sequence. Other methods have failed to separate the salient regions. For Breakdance and Car-roundabout sequences, because of the presence of objects of the same appearance as the object in the background, FSEG [6], LVO [5] and LMP [33] mistakenly segment the background area. And the segmentation results of UOVOS lost some areas. Since, the proposed self-growth strategy uses the information of the entire video to highlight the most salient object, our method outperforms other methods. For the Horsejump-high and Soapbox sequences, our results are better than the state-of-the-art methods, such as recent method UOVOS. Different from the consideration of the whole information of the video sequence, UOVOS only considered the information of adjacent frames, which leads to incomplete segmentation results.

5.4 | DAVSOD, VideoSD and SegTrack-v2

In order to prove the completely unsupervised method, we conduct a comparative experiment on DAVSOD [12], VideoSD [41] and SegTrack-v2 [16] with several available methods. On the one hand, they are completely unfamiliar to the
The results on the DAVISOD dataset is presented in Table 3. Based on the data provided in [12], we compared with the state-of-the-art unsupervised algorithms on the 35 simple subsets of the DAVISOD test set. The result index $S$ — measure $S$ proposed in [40] is used to evaluate the segmentation results. The experimental results show that our Pop-Net can also achieve superior performance on DAVISOD. Although the proposed network Pop-Net. On the other hand, VideoSD and Segtrack-v2 are low-resolution datasets compared with high-resolution DAVIS, which is more challenging for unsupervised VOS. The results on the DAVISOD dataset is presented in Table 3. Based on the data provided in [12], we compared with
score of the proposed method is in the second place, it is only 0.9% away from the first one which may be related to the hardware conditions, as the score of the first place is reported in [12] instead of reproducing in our experimental environment. In addition, even working on a completely strange dataset, the proposed method can still exceed the methods using groundtruth in the training process, which prove the completely unsupervised form of the proposed Pop-Net. Table 4 shows the results on the SegTrack-v2 dataset. As shown in Table 4, compared with the recent unsupervised video object segmentation algorithms, we have achieved the best results and our method performs best in most video sequences. This shows that even on a dataset that has not been fine-tuned, our method can still surpass those methods that use groundtruth in the training phase, such as FSEG [6], LVO [5]. Especially, for the recently proposed method UOVOs [8] that also use appearance and motion saliency, we outperform it in most sequences. And for the birdfall and worm sequences that failed in UOVOS due to the failure of static information acquisition, our performance is better in the comparison methods. Unlike UOVOS, which only uses the information of previous frame, our self-growth strategy considers the most significant object in the entire video sequence, and the final segmentation results are determined by the entire video sequence.

Table 5 illustrates the results on the VideoSD dataset. We compared with the methods using a similar technical route (using appearance saliency and motion saliency to segment the object), and our method outperforms all compared methods. In addition, the score of the proposed completely unsupervised method is even better than the methods using groundtruth in the training phase. Furthermore, our method outperforms the baseline by 21.2% which shows that the proposed self-growth strategy has greatly improved the segmentation ability. Figure 8 shows the qualitative results of partial sequences. The results show that even work on the completely unfamiliar datasets, our performance is significantly better than other methods, including the state-of-the-art methods TIS [31] and UOVOS [8]. TIS [31] performs poorly when dealing with unfamiliar datasets. For the BR13OT sequence, TIS lost its primary goal. For the DO_014 sequence, the object area obtained by the TIS not only has holes, but also includes the background area. The above experimental results show that the proposed method is effective. Even if the base network is completely unfamiliar with the datasets, it can still grow into a salient object segmentation network by the self-growth strategy. And low-resolution does not affect the performance of Pop-Net much.
6 | CONCLUSION

In this article, we present a novel completely unsupervised self-growth video salient object segmentation network. Specially, we propose a self-growth strategy to fully use the whole information of the input video, which makes the network gradually to shift its attention to the most significant object as the video stream progresses. To handle the problem of unsupervised training samples, the sample generation module and the sample optimization module are adopted. Extensive evaluations on benchmark datasets illustrate the effectiveness of the proposed method, especially on completely unfamiliar datasets (no fine-tuned datasets). The results encourage us to do more work on self-growth strategy that is in good accordance with human’s cognition, which may be used to improve the performance of other computer vision tasks besides VOS.

ACKNOWLEDGEMENTS

This work is supported by National Key R&D Program of China (2016YFB1200100) and National Nature Science Foundation of China (51827813, 61472029).

REFERENCES

1. Kohy, J., KimC, S.: Primary object segmentation in videos based on region augmentation and reduction. In: IEEE Conference on Computer Vision and Pattern Recognition, 7417–7425. CVPR (2017)
2. Leey, J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: International Conference on Computer Vision, 1995–2002 (2011)
3. Wang, Q., et al.: visual saliency by selective contrast’. IEEE Trans. Circ. Syst. Video Technol. 23(7), 1130–1155 (2013)
4. Wang, Q., et al.: saliency detection by multiple-instance learning’. IEEE Trans. Cybern. 43(2), 660–672 (2013)
5. Tokmakov, P., Alahari, K., Schmid, C.: learning video object segmentation with visual memory. In: Proceedings of the IEEE International Conference on Computer Vision, 4481–4490 (2017)
6. Jains, D., Xiong, B., Grauman, K.: FusionSeg: learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: IEEE Conference on Computer Vision and Pattern Recognition, 2117–2126. CVPR (2017)
7. Cheng, J., et al.: SegFlow: joint learning for video object segmentation and optical flow. In: IEEE International Conference on Computer Vision. 686–695 (2017).
8. Zhuo, T., et al.: Unsupervised online video object segmentation with motion property understanding’. IEEE Trans. Image Process. 29, 237–249 (2019)
9. Jeong, J., YoonT, S., ParkJ, B: MOSnet: moving object segmentation with convolutional networks’. Electron. Lett. 54(3), 136–138 (2017)
10. Xiankai, L., et al.: See more, Know more: unsupervised video object segmentation with co-attention siamese networks. In: Computer Vision and Pattern Recognition (CVPR), 3623–3632 (2019)
11. Faisal, M., et al.: exploiting geometric constraints on dense trajectories for motion saliency, 13258 (2019). arXiv preprint arXiv: 1909.
12. Dengping, F., et al.: Shifting more attention to video salient object detection. In: Conference on Computer Vision and Pattern Recognition, 8546–8556 (2019)
13. Wenguan, W., et al.: Zero-shot video object segmentation via attentive graph neural networks (2019). arXiv preprint arXiv: 2001.06807.
14. Perazzi, F., et al.: A benchmark dataset and evaluation methodology for video object segmentation. In: IEEE International Conference on Computer Vision, 724–732 (2016)
15. Fukuuchi, KenK., Miyazato, Kouji K., Saliency-based video segmentation with graph cuts and sequentially updated priors. Proc. International Conference on Multimedia and Expo (ICME2009) (638–6412009)
16. Fuxin, Li, et al.: Video segmentation by tracking many figure-ground segments. In: IEEE International Conference on Computer Vision. 2192–2199. (2013)
17. Sheikh, Y., Javed, O., Kanade, T: Background subtraction for freely moving cameras. In: International Conference on Computer Vision, 1219–1225 (2009)
18. Barnich, O., MarcV, D.: ViBe: a universal background subtraction algorithm for video sequences’. IEEE Trans. Image Process. 20(6), 1709–1724 (2011)
19. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: European Conference on Computer Vision, 282–295 (2010)

20. Fragkiadaki, K., Zhang, G., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: Computer Vision and Pattern Recognition, 1846–1853 (2012)

21. Lezama, J., et al.: track to the future: spatio-temporal video segmentation with long-range motion cues. In: Computer Vision and Pattern Recognition, 3369–3376 (2011).

22. Rui, Y., et al.: Video object segmentation and tracking: a Survey (2019). arXiv preprint arXiv:1904.09172

23. Faktor, A., Irani, M.: Video segmentation by nonlocal consensus voting. In: Proceedings of the British Machine Vision Conference, BMVC, 2, 8–18 (2014)

24. Taylor, B., Karasev, V., Soatto, S.: Causal video object segmentation from persistence of occlusions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4268–4276 (2015)

25. Yuan, H., JiaB, H., AlexanderG, S.: Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In: Proceedings of the European Conference on Computer Vision, 786–802 (2018)

26. Esa, R., et al.: Segmenting salient objects from images and videos. In: European Conference on Computer Vision, 366–379 (2010).

27. Wang, Y., et al.: Unsupervised video object segmentation with distractor-aware online adaptation (2018). arXiv preprint arXiv:1812.07712

28. Song, H., et al.: Pyramid dilated deeper ConvLSTM for video salient object detection. In: European Conference on Computer Vision, 744–760 (2018)

29. Siyang, L., et al.: Instance embedding transfer to unsupervised video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6526–6535 (2018)

30. Lu, X., et al.: Learning video object segmentation from unlabelled videos (2020). arXiv preprint arXiv:2003.05020

31. Griffin, A., Corso, J.: Tukey-inspired video object segmentation. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1723–1733. WACV) (2019).

32. Wenguan, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3395–3402 (2015)

33. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: IEEE International Conference on Computer Vision. 1777–1784 (2014)

34. Tokmakov, P., Alahari, K., Schmid, C.: learning motion patterns in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3386–3394 (2017).

35. Li, S., et al.: Unsupervised video object segmentation with motion-based bilateral networks. In: European Conference on Computer Vision, 207–223. ECCV (2018)

36. Philippe, W., et al.: Learning to detect motion boundaries. In: IEEE Conference on Computer Vision Pattern Recognition, 2578–2586 (2015)

37. Qinbin, H., et al.: Deeply supervised salient object detection with short connections. IEEE Trans. Pattern Anal. Mach. Intell. 3203–3212 (2017)

38. Achanta, R., et al.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34(11), 2274–2282 (2012)

39. Hui, Y., et al.: Adaptive convolutional neural network for large change in video object segmentation. IET Comput. Vis. 13(5), 452–460 (2019)

40. Caelles, S., et al.: One-shot video object segmentation. In: Computer Vision and Pattern Recognition, 5320–5329. CVPR (2016)

41. Dengping, F., et al.: Structure-measure: a new way to evaluate foreground maps. In: IEEE International Conference on Computer Vision, 4548–4557. (2017). http://dpfan.net/smeasure/

How to cite this article: Yin, H., et al.: Pop-net: A self-growth network for popping out the salient object in videos. IET Comput. Vis. 15(5), 334–345 (2021). https://doi.org/10.1049/cvi2.12032