Attention-based video object segmentation algorithm

Ying Cao \(^1\)  |  Lijuan Sun \(^2,3\)  |  Chong Han \(^2,3\)  |  Jian Guo \(^2,3\)

\(^{1}\) Henan University, Kaifeng, Henan, China  
\(^{2}\) Nanjing University of Posts and Telecommunications, Nanjing, China  
\(^{3}\) Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications, Nanjing, China

Abstract

To improve the segmentation performance on videos with large object motion or deformation, a novel scheme is proposed which has two branches. In one branch, the attention mechanism is first utilized to highlight objects-related features. Then, to well consider the temporal coherence of videos, Conv3D is integrated to capture short-term temporal features, and the designed attention residual convolutional long–short-term memory is adopted to capture the long–short-term temporal information of objects under the interference of redundant video frames. Meanwhile, considering the negative effect of background motion, in another branch, the optical flow-based prediction model is introduced to predict objects regions in subsequent video frames with the annotated initial frame. At last, based on the fused results of two branches, the global thresholds and noising area clean method are employed to obtain segmented objects. The experiments on DAVIS2016 and CDnet2014 exhibit the competitive performance of the proposed scheme.

1 | INTRODUCTION

Video object segmentation (VOS) is an important problem in computer vision, and it has been used widely in vision tasks like object tracking [1], event recognition [2] and video indexing [3]. Different with image segmentation, which groups similar pixels into regions based on certain features in spatial domain, VOS needs to consider information in temporal domain due to the strong temporal correlation between consecutive video frames. The temporal features consideration in recent researches mainly revealed in motion estimations like optical flow [4, 5] or pixel trajectory [6]. However, motion estimation like optical flow and pixel tracking has the difficulty in obtaining accurate segmentation of videos with noise, blurring, deformation, occlusion and large motion.

With the development of deep learning, recent researches attempt to handle VOS with neural network (NN) framework. With the powerful learning ability and amounts of training data, methods based on deep learning [7, 8, 10, 11] can harvest pretty quality segmentation results. In recent methods, there exist schemes simply treat a video as static images, and segment the objects individually, hence, when video frames increase or video objects exist violent motion or deformation [7, 9], the temporal coherence would be weaken, and the segmentation quality would be affected. Therefore, scholars attempt to integrate motion estimation into NN architecture [8], which can improve the quality of segmented objects and is more suitable for VOS problem. However, the NN-based motion estimation can also be affected by large motion or deformation in videos, therefore, in order to reduce the negative impact of large objects motion or deformation on temporal information, this paper focuses on short- and long-term temporal features simultaneously, and designs to adapt attention mechanism as well as optical flow information to exclude the negative effect of background information.

Specifically, a novel scheme for VOS problem is proposed named as attention-based spatiotemporal convolutional neural network–long-short-term memory video object segmentation (ASCNN-LSTMVOS), as shown in Figure 1. The proposed model has two branches: neural network-based prediction (NNbP) model and optical flow-based prediction (OFbP) model. In NNbP, attention mechanism is first employed to highlight objects-related features. Then, to well consider temporal coherence of video frames, Conv3D is adopted to capture...
short-term temporal features, and the designed attention residual convolutional long–short-term memory (AR-ConvLSTM) is integrated to capture long–short-term temporal features of objects under the interference of redundant video frames. In this way, NNbP can harvest objects confidence maps based on obtained spatiotemporal features. Besides, considering the negative effect of background motion, in OFbP, the optical flow between frames is employed to predict objects regions in subsequent video frames with annotated initial frame. At last, based on the merged results of two branches, global thresholds and noising area clean algorithm are utilized to obtain the segmented objects of videos. The proposed model is validated with DAVIS2016 [15] and CDnet2014 [16], the experimental results indicate that ASCNN-LSTMOVOS exhibits competitive performance. Our main contributions are listed as follows:

1. Based on the objects-related short-term motion information obtained by Conv3D, the AR-ConvLSTM is designed to properly capture the long–short-term temporal features under the interference of redundant video frames.
2. The optical flow information of video sequence is analysed to alleviate the negative effect of background motion on segmentation performance.

2 | RELATED WORK

Over the past few decades, diverse approaches have been developed for VOS problem, and these existing methods can be mainly divided into semi-supervised methods and unsupervised methods.

2.1 | Unsupervised VOS

Unsupervised VOS methods have received widely attraction for the needless of human annotation, and there exists pretty work like Hu et al. [19] proposed a motion-guided cascaded refinement network (CRN) for VOS. In their work, they applied an active contour model on optical flow to coarsely segment the foreground by assuming that the target object had different motion patterns with background. Then, the coarse segmentation was taken as the guidance to generate the accurate segmentation for each frame. In this way, the motion information and deep CNNs could well complement each other to accurately segment the foreground objects from video frames. Besides, the experiments on benchmarks indicated that their method achieved state-of-the-art performance at a much faster speed.

In addition, Tokmakov et al. [20] proposed MP-Net network to determine whether an object is in motion, irrespective of camera motion. And in order to solve the limitations of appearance features of objects in MP-Net, in [11], Tokmakov et al. integrated one stream with appearance information and a visual memory module based on convolutional gated recurrent units (GRU). Li et al. [12] introduced an unsupervised VOS method by transferring the knowledge encapsulated in image-based instance embedding networks.

In [13], Song et al. proposed a fast video salient object detection model, based on a novel recurrent network architecture, named pyramid dilated bidirectional ConvLSTM (PDB-ConvLSTM). In PDB-ConvLSTM, a pyramid dilated convolution (PDC) module is first designed for simultaneously extracting spatial features at multiple scales. And then, these spatial features are concatenated and fed into an extended deeper bidirectional ConvLSTM (DB-ConvLSTM) to learn spatiotemporal information. Afterwards, DB-ConvLSTM was further augmented with a PDC-like structure by adopting several dilated DB-ConvLSTMs to extract multi-scale spatiotemporal information. Extensive experiments indicated the superior performance of PDB-ConvLSTM.

Lu et al. introduced a unified unsupervised/weakly supervised learning scheme called MuG in [37] for VOS that
addressed object pattern learning from unlabelled videos. The proposed approach could help advance understanding of visual pattern in VOS and significantly reduced annotation burden. Experiments demonstrated the promising potential of MuG in leveraging unlabelled data to further improve the segmentation accuracy.

2.2 Semi-supervised VOS

Different with unsupervised methods, based on the given first frame or key frames, semi-supervised methods segment videos objects by transferring objects mark information between video sequences. For example, In [9], Caelles et al. proposed one-shot video object segmentation (OSVOS) based on a fully convolutional neural network (FCN) architecture to obtain the segmented objects. The OSVOS was able to successfully transfer generic semantic information, which learned on ImageNet, to the task of foreground segmentation, and through the appearance learning of a single annotated of test sequence, the OSVOS could obtain the object in each video frame individually. The experiments on two annotated video segmentation data sets demonstrated that OSVOS was fast and improved the state-of-the-art by a significant margin (79.8% versus 68.0%). On the basis of OSVOS, Voigtlaender and Leibe [23] introduced online adaptive video object segmentation (OnAVOS) to update the network online using training examples selected based on the confidence of the network and spatial configuration. The validated experiments showed that both extensions were highly effective and improved the state-of-the-art on DAVIS to an intersection-over-union score of 85.7%.

In SegFlow, proposed by Cheng et al. [8], the objects-related information and optical flow were propagated bidirectionally. Through this way, the SegFlow can predict pixel-wise object segmentation and optical flow in videos simultaneously. Extensive experiments on VOS and optical flow data sets demonstrated that the introduced optical flow could improve the performance of segmentation and vice versa.

Jain et al. proposed an end-to-end learning framework for segmenting generic objects in videos in [21]. In their paper, Jain et al. formulated this task as a structured prediction problem and designed a two-stream FCN which fused motion, appearance in a unified framework. The experiments on three challenging video segmentation benchmarks indicated that the proposed method substantially improved the state-of-the-art results for segmenting (unseen) objects.

In [22], Jampani et al. proposed a video propagation network (VPN) to process frames in an adaptive manner. In their paper, two components were combined: a temporal bilateral network for video adaptive filtering, followed by a spatial network to refine features and increased flexibility. Extensive experiments on VOS and semantic video segmentation depicted the increased performance compared to the best previous task-specific methods, while had favourable runtime.

Lin et al. introduced a deep learning-based approach with a novel online adaptation technique using optical flow in [29] named Flow Adaptive Video Object Segmentation (FLAVOS). Comparing with other deep learning-based schemes, the authors provided extensive complexity analysis which additionally demonstrated that FLAVOS was natural for real-world applications by introducing an interactive pipeline that enabled the user to provide feedback for online training. The experiments on three challenging benchmark data sets and nearly ground-truth level segmentation results with interactive user feedback demonstrated the superiority of FLAVOS.

In [36], for the rough and inflexible problem of the existing approaches that relied on selecting the region of interest for model update, Sun et al. proposed a novel approach which utilized reinforcement learning to select optimal adaptation areas for each frame based on historical segmentation information. Furthermore, to speed up the model adaption, a novel multi-branch tree-based exploration method was designed to quickly select the best state action pairs. Evaluated results on common VOS data sets demonstrated the competitive performance of the introduced scheme.

The previous work indicates the importance of temporal information for VOS, however, the large deformation or motion of objects in videos would affect the long-term temporal features. Therefore, in order to reduce the negative impact of large object deformation or motion on temporal information, this paper focuses on short- and long-term features simultaneously, and designs to employ optical flow to exclude the negative effect of background motion.

3 THE PROPOSED SCHEME

To improve the segmentation performance on videos with large motion or deformation, ASCNN-LSTMVOS is proposed by constructing two branches: NNoP and OFbP. In NNoP, the attention mechanism is employed to highlight the features of objects. Then, to well capture temporal information, Conv3D is adopted to obtain the short-term temporal features, and AR-ConvLSTM is designed to capture the long–short-term temporal features under the interference of redundant frames. Through the captured features, NNoP harvests confidence maps. Meanwhile, considering negative influence of background motion on segmentation, in OFbP, optical flow between video frames and annotated initial frame are employed to predict the objects regions in subsequent video frames. At last, based on the fused results of two branches, global thresholds and noising area clean method are integrated to generate segmentation of videos.

3.1 The NN-based branch

In proposed model, the NN-based branch is constructed by encoder, decoder and classifier modules. The encoder module mainly includes Conv3D, attention and mini-encoder block. In this stage, attention mechanism is used to highlight objects-related features, and Conv3D is adopted to acquire
short-term objects’ temporal features. The decoder module is mainly constructed by concatenation and batch normalization (BN). At the end of encoder and decoder, the designed AR-ConvLSTM is integrated to retain long-term temporal coherence of objects’ features under the interference of redundant video frames. Finally, the sigmoid function is adopted as the classifier to generate confidence maps.

### 3.1.1 Conv3D

The temporal consistency of objects is one of the key to solve the VOS problem. Based on this, as described in Figure 2, the structure of Conv3D [34, 35] indicates that its kernel is a cube, which is formed by stacking several contiguous frames together. Consequently, Conv3D can capture the short-term or local temporal features by considering the temporal motion relationship of multiple continuous video frames. Therefore, Conv3D [34, 35] is integrated in the proposed scheme to capture local (short-term) temporal features. Formally, the feature map of a standard Conv3D is defined as:

$$C_{3D}(x_t, z) = \sum_{t=1}^{T-1} \sum_{k=0}^{K-1} \sum_{l=0}^{K-1} w(t, k, l) \ast x(z + t, x + k, y + l),$$  

(1)

where $w \ast$, $T$, $K$, $x$, $(x_t, y, z)$ and $(t, k, l)$ represent the kernel, the convolution operation, temporal length of the data, size of the kernel, input data, first coordinate of the video frame and element index of the kernel, respectively.

### 3.1.2 Attention mechanism

In convolution operation, the elements in kernel have different reflection to background and foreground objects, but in terms of the weight sharing characteristic of convolution operation, the features of background would affect the accuracy of segmentation. For this problem, the spatial attention mechanism is adopted to highlight objects-related features. Taking a single video frame as the example, attention scheme can be described as Figure 3, where different colour represents different weights, and the objects-related features can be highlighted through these different attention weights.

### 3.1.3 AR-ConvLSTM

In order to well capture the long-term temporal features of videos objects, AR-ConvLSTM is designed on the basis of [18, 33].

Convolutional long–short-term memory (ConvLSTM), first introduced in [17], has the advantage of obtaining motion trajectory in contiguous video frames [13]. Formally, ConvLSTM can be defined as:

$$i_t = \sigma(W_{ix} \ast x_t + W_{ih} \ast h_{t-1} + b_i),$$  

(2)

$$f_t = \sigma(W_{fx} \ast x_t + W_{fh} \ast h_{t-1} + b_f),$$  

(3)

$$o_t = \sigma(W_{ox} \ast x_t + W_{oh} \ast h_{t-1} + b_o),$$  

(4)

$$C_t = \tanh(W_{cx} \ast x_t + W_{ch} \ast h_{t-1} + b_c),$$  

(5)

$$h_t = o_t \circ C_t + i_t \circ C_t,$$  

(6)

$$h_t = o_t \circ \tanh(C_t),$$  

(7)

where $x_t$ is the input. $W_{x}$ and $W_{h}$ are convolutional kernel. $\sigma$ is the sigmoid function, $i_t$, $f_t$ are cell state, hidden state, respectively. $C_t$ is the candidate memory, $i_t$, $f_t$, $o_t$ are input gate, forget gate and output gate. The symbol ‘$\ast$’ is the convolution operation, ‘$\circ$’ denotes the Hadamard product.

In ConvLSTM structure, the input $x$, is the spatiotemporal features obtained based on adjacent several frames. However, there exist redundant video frames, and this redundant information in $x$ would affect the accuracy of segmented objects. Therefore, in this section, the spatial attention mechanism is integrated in ConvLSTM as [18], based on the hidden state $h_{t-1}$, to alleviate the interference of redundant information and enhance the temporal coherence of features, thus properly obtaining the objects-related long–short-term temporal information. The adopted attention mechanism is described as:

$$Z_t = W_{z} \ast \tanh(W_{za} \ast x_t + W_{zh} \ast h_{t-1} + b_z),$$  

(8)
An overview of AR-ConvLSTM

where $x_t$ is the input, $W_z$ are convolutional kernel. $L$ is the spatial index set, $A_t$ represents the attention weight corresponding to the element at $l$ position in $x_t$. With $A_t$, $x_t$ is adjusted to $\tilde{x}_t$ to alleviate the interference of redundant frames. Then $\tilde{x}_t$ is integrated to obtain valuable long–short-term temporal information of objects.

After that, based on $\tilde{x}_t$ which represents the valuable spatiotemporal information, as described in [18], ConvLSTM can only focus on the spatiotemporal features fusion along with the recurrent steps, and the gate value need not to be calculated for each element, just for each feature map of the states. Hence, the global average pooling can be utilized to reduce the spatial dimension of input features $\tilde{x}_t$ and hidden state $h_{t-1}$. Then the acquired spatiotemporal features are fed into sigmoid module to generate probability distribution maps.

### 3.2 OFbP model

In ASCNN-LSTMVOS, Conv3D and AR-ConvLSTM are used to acquire short- and long-term temporal features of objects. However, the background motion in videos would be noising information in segmenting objects. Therefore, another branch is designed in the proposed model to exclude the noising background motion. In OFbP, the optical flow between video frames is first calculated, then the results and annotated initial frame are used to predict the objects regions through interpolation operation. Finally, the predictions are combined with the first branch results, the combination is defined as:

$$CombineObject = \alpha \times NN_{seg} + (1 - \alpha) \times OF_{pred},$$

where $NN_{seg}$ is the results generated by first branch, $OF_{pred}$ is the predicted frames on the basis of optical flow information, $\alpha$ is the parameter to balance the $NN_{seg}$ and $OF_{pred}$, the value is

**Algorithm 1** Optical Flow-Based Prediction Model

**Input:** Video frame number $frame\_number$, annotated initial frame: $I_{initial}$

**Output:** Predictions set: $OF_{preds}$

1) calculate the optical flow and save in $OF_{set}$
2) for $img := 2; img <= frame\_number; do$
3) $[vx, vy] = OF_{set}(img)$
4) $OF_{pred} = interp2(I_{initial}, vx, vy)$
5) put $OF_{pred}$ in $OF_{preds}$
6) $I_{initial} = OF_{pred}$
7) endfor
8) return Predictions set $OF_{preds}$
in range [0,1]. In this way, the proposed model can handle the background motion properly.

3.3 The noising area clean algorithm

With the merged results, the proposed model employs empirical global thresholds to generate initial segmentation. Then the noising area clean algorithm is employed to obtain final segmentation. In this algorithm, the initial segmented area are calculated, and the noising clean algorithm deletes the region when its computed area is lower than threshold \( T_c \).

4 EXPERIMENT

In this section, the proposed model is evaluated on DAVIS2016 [15] and CDnet2014 [16] data sets, with total 25 videos that contain dynamic background, camera jitter, shadow, deformation, large motion and occlusion.

In order to satisfy the input requirement of the proposed scheme, the experimental data in \( 4D \in \mathbb{R}^{b \times t \times H \times W} \) should be transformed into \( 5D \in \mathbb{R}^{b \times c \times t \times H \times W} \) as [14], where \( b \), \( t \), \( H \), \( W \) and \( C \) are batch size, the number of look back frames, frame height, frame width and channel, respectively. Furthermore, the proposed model is carried on the platform with GPU, which is an NVIDIA GeForce GTX 1080 Ti with 11 GB memory.

Besides, to appropriately measure the performance of the proposed method, FoM is utilized as the main metric, and it is defined in Equation (20):

\[
\text{FoM} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \times 100\%, \tag{20}
\]

where \( \text{Precision} \) is the percentage of correct prediction compared to the total number of positive, and \( \text{Recall} \) describes how many positive examples are correctly predicted.

4.1 Experiments on DAVIS2016

In this section, ASCNN-LSTMVOS is evaluated on DAVIS2016 with 20 videos, namely, Blackswan (BS), Bmx-trees (BT), Breakdance (BD), Camel (CA), Car-roundabout (CR), Car-shadow (CS), Cows (CO), Dance-twirl (DT), Dog (DO), Drift-chicane (DC), Drift-straight (DS), Goat (GO), Horsecross-jump (HH), Kite-surf (KS), Libby (LI), Motocross-jump (MJ), Paragliding-launch (PL), Parkour (PA), Scooter-black (SB) and Soapbox (SO). In training strategy, the supervised training method is adopted, and based on the model trained with 30 video sequences, the fine-tuning strategy is utilized for validation data set, moreover, the samples satisfy Equation (21), where \( N \) is a constant, in this experiment, \( N = 1 \), and the rest as the test set.

\[
\text{Number}_{\text{samples}} = b \times N + (t - 1). \tag{21}
\]

TABLE 1  Average performance comparison on DAVIS2016 (FoM, 0 ≤ FoM ≤ 100)

| Method                          | FoM       |
|--------------------------------|-----------|
| 3D CNN-LSTM [24]              | 87.20     |
| 3D CNN-LSTM-2                 | 84.24     |
| Ours                           | 90.29     |

TABLE 2  Average performance comparison on DAVIS2016 (%)

| Method       | J_w | F_w |
|--------------|-----|-----|
| BVS [25]     | 57.1(60.0) | 65.7(68.0) |
| OFL [24]     | 67.8(79.7) | 78.9(80.6) |
| MSK [27]     | 78.4(79.8) | 82.5   |
| OSVOS [9]    | 80.6 |      |
| Ours         |      |      |

In this simulation, \( \text{FoM} \) and public metrics of DAVIS2016 are utilized to achieve the performance comparisons with 3D CNN-LSTM [28] and OFL [24], BVS [25], MSK [27], OSVOS [9]. The employed public metrics of DAVIS2016 are region similarity (\( J \)) and contour accuracy (\( F \)) that defined in Equations (22) and (23):

\[
J = \frac{M \cap G}{M \cup G}, \tag{22}
\]

\[
F = \frac{2P_R}{P_R + R_c}, \tag{23}
\]

where \( M \) is the segmentation object region, \( G \) is the ground truth. \( P_c \) is the precision of segmented object contour and \( R_c \) is the recall of segmented object contour.

The experimental results are displayed in Tables 1 and 2. Table 1 shows the compared results of the proposed with 3D CNN-LSTM using FoM, and the exhibited results indicate that ASCNN-LSTMVOS has \( \approx 3\% \) improvement. In addition, the same experiment condition is applied to 3D CNN-LSTM which is represented as 3D CNN-LSTM-2 in Table 1, and the result also demonstrates the superior performance of ASCNN-LSTMVOS.

Meanwhile, the proposed scheme is validated with public metrics of DAVIS2016, and the compared results are described in Table 2. In this table, the values outside the parenthesis represent the results with the test set used in our experiments, and the experimental results displayed in this table verify a certain improvement of ASCNN-LSTMVOS.

Besides, Figure 8 displays the segmentations of ASCNN-LSTMVOS. From the segmented results, it can be found that the proposed scheme can completely segment objects from background. Specifically, ASCNN-LSTMVOS can accurately segments objects of videos with stable motion like ‘Blackswan’, ‘Camel’ and ‘Dog’. For videos with violent object motion and object deformation, the proposed model can properly segment the objects like ‘Dance-twirl’ and ‘Parkour’. However, ASCNN-LSTM needs to be further optimized to improve the performance on videos with occlusion like ‘Bmx-tree’.

Number of samples = \( b \times N + (t - 1) \).
4.2 Experiments on CDnet2014

In this section, in order to further evaluate the performance of ASCNN-LSTMVOS, the proposed model is then validated on CDnet2014 with five videos (highway, office, Pedestrians, PETS2006 and Sofa), and made comparisons with DBFCN [30], DeepBS [31], MBS [32] and 3D CNN-LSTM [24].

In training strategy of CDnet2014, the videos are trained individually, and in supervised training strategy, considering the temporal connectivity of video objects, instead of random selection, the first 50% frames in each video are selected as the training set, the rest is employed as the test set. The evaluated results are shown in Table 3, the exhibited results demonstrate that ASCNN-LSTMVOS has better performance than algorithms DBFCN [30], DeepBS [31] and MBS [32], but still exists a gap with 3D CNN-LSTM [24]. However, opposed to 3D CNN-LSTM, which utilizes 70% frames as the training set, the training samples of proposed model in CDnet2014 is 50%. Furthermore, the same experiment situation is applied to 3D CNN-LSTM, and the results are represented as 3D CNN-LSTM-2 exhibited in Table 3. Compared with 3D CNN-LSTM-2, the proposed model presents superior performance.

In addition, Figure 6 presents a part of FoM results of each videos in CDnet2014. From the displayed FoM results, it is obviously to find that ASCNN-LSTMVOS can completely segment videos objects and achieve >90% FoM results in most videos.

The FoM results of highway test set are described in Figure 6(a). The FoM results indicate that the FoM of most video frames in highway are above 90%, but lower than 90% in range of 1487–1490 frames. Through the analysis of experimental
results, as described in Figure 7, it can be found that the shadow of trees branches leads to the loss of segmented vehicles regions in 1487–1490 frames.

Figure 6(b) presents a part of FoM results of office test set, and the exhibited results show that ASCNN-LSTMVOS has a pretty segmentation performance, but still exists a significant performance drop in last consecutive 10 frames. With the segmentation presented in Figure 7, the experimental analysis indicates that the continue decreasing of objects’ area in subsequent consecutive frames affect the segmentation performance of the proposed scheme. Therefore, the proposed model needs to be further improved to properly process small VOS problem.

The FoM results of pedestrians test set are partly described in Figure 6(c). The shown results depict that the proposed model achieves ≥90% FoM in most frames. However, ASCNN-LSTMVOS has relatively poor performance in 830–833 frames. The experiments analysis indicates that the poor performance is caused by gradually smaller objects.

Figure 6(d) describes a part of FoM results of PETS2006, from the experimental results, it can be found that ASCNN-LSTMVOS has the value decrease problem of FoM in 1157–1184 frames, and the analysis of segmented objects depict that, same with videos above, the proposed model needs to be optimized to improve the segmentation performance of small video objects.

Compared with other video sequences, Sofa video sequence has relatively complex objects motion, and its experimental results are exhibited in Figures 6(e) and 7. The results evaluate that the proposed scheme is able to handle the complex...
movement of video objects and segment the objects properly from video background. However, there exist relatively more frames which $\text{FoM} \leq 90\%$ in Sofa video, the experimental analysis indicates that the performance of ASCNN-LSTMVOS is partly affected by video background.

In brief, the experimental results depict that the proposed model exhibits the competitive performance. And compared with DAVIS2016 in terms of FoM, there exists $\approx 4\%$ increase in CDnet2014, the experimental analysis indicates that the training set in CDnet2014 is more than that in DAVIS2016, thereby improving the performance. In a way, the higher FoM results can demonstrate the superiority of ASCNN-LSTMVOS in long videos.

### 4.3 The validation of attention and AR-ConvLSTM

Furthermore, based on 3D CNN-LSTM [24], the validation of attention, AR-ConvLSTM are carried out in this section, and the results are described in Table 3, where $+\text{AR-ConvLSTM}$ represents the experimental results of the model constructed by only adding AR-ConvLSTM to 3D CNN-LSTM.

Compared with 3D CNN-LSTM-2 displayed in Table 3, the experimental results depict that the employment of AR-ConvLSTM has a certain improvement of performance especially in videos with relatively complex motion like Sofa. And faced with the ASCNN-LSTMVOS, the compared results demonstrate the utilization of attention mechanism improves the overall performance, but not always helpful like in highway sequence which needs to be further analysed.

### 4.4 The validation of OFbP and noising area clean algorithm

Taking Car-roundabout (CR) as the example, in this section, the fused results of two branches is displayed in Figure 8, and from the results, it is obviously to find that the employment of optical flow helps identify the objects from background which has similar features.

In addition, Figure 9 represents segmentation generated by the integrated noising area clean algorithm, and the presented results prove the significant of noising area clean algorithm.

### 4.5 Runtime analysis

The runtime of ASCNN-LSTMVOS is validated in this section. Taking CDnet2014 as the example, the consumed time of the proposed model on CDnet2014 is shown in Table 4. In Table 4, NNbP represents the runtime of first branch, OFbP is the consumed time of optical flow prediction branch, Global

**TABLE 4** The runtime of ASCNN-LSTMVOS on CDnet2014 (second/s)

|                  | Highway | Office | Pedestrians | PETS2006 | Sofa |
|------------------|---------|--------|-------------|----------|------|
| NNbP             | 0.230   | 0.260  | 0.170       | 0.200    | 0.370|
| OFbP             | 1.109   | 1.107  | 1.107       | 1.103    | 1.109|
| Global           | 0.005   | 0.005  | 0.005       | 0.005    | 0.004|
| Clean            | 0.003   | 0.003  | 0.003       | 0.003    | 0.003|
means the cost time of classify using global thresholds and clean stands for the time expended by noising area clean algorithm. Then, under the same experiment condition, compared with 3D CNN-LSTM which is 28 frames-per-second (FPS), it is obviously to find that there exists a gap of the proposed model on runtime. The experimental analysis indicate that the heavy time overhead of the proposed scheme is mainly caused by optical flow prediction model. Besides, the computation speed of NN-based part in the proposed model is only about 15 FPS. Therefore, ASCNN-LSTMVOS needs to be optimized to decrease the time consumption in VOS problem.

4.6 Limitations

The failure cases of ASCNN-LSTMVOS are presented in Figure 10. Combining the segmentations displayed in Figures 5 and 7, it can be found that ASCNN-LSTMVOS has the difficulty to segment the target from the video with occlusion, such as ‘Bmx-tree’ and ‘Highway’. And the proposed model cannot handle properly in videos with small videos objects.

In addition, on the basis of DAVIS2016 and CDnet2014, the comparisons are also made with [29, 36] and [33]. As described in Table 5, compared with related approaches [29, 36], disparity obviously exists in segmentation performance. In Table 6, compared with [33] which employs 50% frames in training stage, the exhibited results on CDnet2014 depict the superior performance of the proposed in PET2006, but needed to be further improved considering the overall results.

In summary, the experimental results depict that the proposed model exhibits competitive performance and has a certain superiority in long video segmentation tasks. Detailedly, ASCNN-LSTMVOS can well segment targets from video sequences with violent object motion or deformation, the proposed OFbP model can help identify the target from the background which has similar features with objects, and the integrated noising area clean algorithm can effectively remove non-objects’ regions in segmentation.

However, there still exist limitations of proposed scheme. First, ASCNN-LSTMVOS has difficulties in segmenting targets from videos with small objects or severe environmental occlusion. Secondly, compared with the existing segmentation algorithms, the segmentation performance and speed need to be further optimized, as also the adopted supervised training method which is relatively complicated and inconvenient.

5 CONCLUSION

To improve the segmentation performance on videos with large object motion or deformation, ASCNN-LSTMVOS is proposed for VOS problem. In proposed model, two branches are constructed: NNbP and OFbP. In NNbP, attention mechanism is first employed to highlight objects-related features. Then, to well consider temporal coherence of video frames, Conv3D is adopted to capture short-term temporal features, and the designed AR-ConvLSTM is integrated to capture long–short-term temporal information. In this way, NNbP can harvest confidence maps based on obtained spatiotemporal features. Besides, considering the negative effect of background motion, in OFbP, the optical flow between frames is employed to predict objects regions in subsequent video frames with annotated initial frame. At last, based on the merged results of two branches, global thresholds and noising area clean algorithm are utilized to obtain the segmented objects of videos. The proposed model is validated with DAVIS2016 and CDnet2014, and the experimental results indicate competitive performance of ASCNN-LSTMVOS.

However, ASCNN-LSTMVOS has difficulties in segmenting targets from videos with small objects or severe environmental occlusion. Secondly, there still exist gaps in segmentation performance and speed. Thirdly, the adopted supervised training method is relatively complicated. Therefore, in next stage, the weaknesses of the proposed model will be analysed and further optimized by taking advantage of short- and long-term temporal information.
ACKNOWLEDGEMENTS
This work is supported by the National Natural Science Foundation of China under Grant No. 61572261, 61702284 and 61672209, the Natural Science Foundation of Jiangsu Province under Grant No. BK20150868, the Postdoctoral Found of Jiangsu Province under Grant No. 1701165C.

ORCID
Ying Cao https://orcid.org/0000-0002-8256-8497

REFERENCES
1. Yeo, D., Son, J., Han, B.: Superpixel-based tracking-by-segmentation using Markov chains. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 21–26 (2017)
2. Hu, W., et al.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 41(6), 797–819 (2011)
3. Baraldi, L., Grana, C., Cucchiara, R.: NeuralStory: An interactive multimedia system for video indexing and re-use. In: Proc. of the International Workshop on Content-Based Multimedia Indexing, pp. 1–6 (2017)
4. Ilg, E., Mayer, N., Saikia, T.: FlowNet2.0: Evolution of optical flow estimation with deep networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
5. Hu, P., et al.: Recurrent spatial pyramid CNN for optical flow estimation. IEEE Trans. Multimedia 20(10), 2814–2823 (2018)
6. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by GPU-accelerated large displacement optical flow. In: Proc. of European Conference on Computer Vision (ECCV) (2016)
7. Paul, V., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: Proc. of the Int. Conf. on British Machine Vision Conference (2017)
8. Cheng, J., Tsai, Y.-H., Wang, S.: Segflow: Joint learning for video object segmentation and optical flow. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2017)
9. Caelles, S., Maninis, K.-K., Pont-Tuset, J.: One-shot video object segmentation. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
10. Shin Yoon, J., Rameau, F., Kim, J.: Pixel-level matching for video object segmentation using convolutional neural networks. In: Proc. of IEEE International Conference on Computer Vision (ICCV) (2017)
11. Perazzi, F., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2017)
12. Li, S., et al.: Instance embedding transfer to unsupervised video object segmentation. In: Proc. of Int. Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
13. Song, H., et al.: Pyramid dilated deeper ConvLSTM for video salient object detection. In: Proc. of the European Conference on Computer Vision (ECCV) (2018)
14. Akilan, T., et al.: A 3D CNN-LSTM based image-to-image foreground segmentation. IEEE Trans. Intell. Transp. Syst. 21(3), 959–971 (2020)
15. Perazzi, F., Pont-Tuset, J., McWilliams, B.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
16. Wang, Y., Jodoin, P. M., Porikli, F.: CDnet 2014: An expanded change detection benchmark dataset. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2014)
17. Shi, X., et al.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS’15), pp. 802–810 (2015)
18. Zhang, L., et al.: Attention in convolutional LSTM for gesture recognition. In: Proc. Int. Conf. Neural Information Processing Systems (NIPS) (2018)
19. Hu, P., et al.: Motion-guided cascaded refinement network for video object segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 42(8), 1957–1967 (2020)
20. Tokmakov, P., et al.: Learning motion patterns in videos. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
21. Jain, S., et al.: FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2126 (2017)
22. Jampani, V., et al.: Video propagation networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 451–461 (2017)
23. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
24. Tsai, Y.H., et al.: Video segmentation via object flow. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3899–3908 (2016)
25. Maerki, N., et al.: Bilateral space video segmentation. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 743–751 (2016)
26. Bianco, S., Ciocca, G., Schettini, R.: Combination of video change detection algorithms by genetic programming. IEEE Trans. Evol. Comput. 21(6), 914–928 (2017)
27. Perazzi, F., et al.: Learning video object segmentation from static images. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2663–2672 (2017)
28. Thangarajah, A., et al.: A 3D CNN-LSTM based image-to-image foreground segmentation. IEEE Trans. Intell. Transp. Syst. 99, 1–13 (2019)
29. Lin, F., Chou, Y., Martínez, T.: Flow adaptive video object segmentation. Image Vis. Comput. 94, 103864 (2020)
30. Lu, Y., et al.: Deep background modeling using fully convolutional network. IEEE Trans. Intell. Transp. Syst. 19(1), 254–262 (2017)
31. Babaei, M., et al.: A deep convolutional neural network for background subtraction. Pattern Recognit. 76, 635–649 (2018)
32. Sajid, H., Cheung, S.C.S.: Universal multimode background subtraction. IEEE Trans. Image Process. 26(7), 3249–3260 (2017)
33. Chen, Y., et al.: Pixelwise deep sequence learning for moving object detection. IEEE Trans. Circuits Syst. Video Technol. 29(9), 2567–2579 (2019)
34. Ji, S., et al.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35(1), 221–231 (2013)
35. Liu, H., et al.: Two-stream 3D convolutional neural network for skeleton-based action recognition. arXiv:1705.08106 (2017)
36. Sun, M., et al.: Adaptive ROI generation for video object segmentation using reinforcement learning. Pattern Recognit. 106, 107465 (2020)
37. Lu, X., et al.: Learning video object segmentation from unlabeled videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8960–8970 (2020)

How to cite this article: Cao Y, Sun I, Han C, Guo J. Attention-based video object segmentation algorithm. IET Image Process. 2021;15:1668–1678. https://doi.org/10.1049/ipr2.12135