Video salient object detection via spatiotemporal attention neural networks

Video salient object detection via spatiotemporal attention neural networks

Yi Tang[a], Wenbin Zou[a,*], Yang Hua[b], Zhi Jin[a] and Xia Li[a]

[a] *Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Guangdong Laboratory of Artificial Intelligence and Digital Economy(SZ), Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China*

[b] *EEECS/ECIT Queen's University Belfast, United Kingdom*

ARTICLE INFO

Keywords:
Video salient objects detection
motion information
spatiotemporal feature fusion
coarse pixel-wise labels

ABSTRACT

Recently, deep convolutional neural networks have been widely introduced into image salient object detection and achieve good performance in this community. However, as the complexity of video scenes, video salient object detection with deep learning models is still a challenge. The specific difficulties come from two aspects. First of all, the deep networks on image saliency detection cannot capture robust motion cues in video sequences. Secondly, as for the spatiotemporal fusing features, the existing methods simply exploit element-wise addition or concatenation, which not fully explores the contextual information and complementary correlation, thus they cannot produce more robust spatiotemporal features. To address these issues, we propose a two-stream based spatiotemporal attention neural network (STAN) for video salient object detection. We amply extract the motion information in terms of long short term memory (LSTM) network and 3D convolutional operation from optical flow-based prior and video sequences. Moreover, an attentive module is designed to integrate the different types of spatiotemporal feature maps by learning the corresponding weights. Meanwhile, in order to generate sufficient pixel-wise annotated video frames, we manually generate lots of coarse labels, which are well utilized to train a robust saliency prediction network. Experiments on the widely used challenging datasets (e.g., FBMS and DAVIS) prove that the proposed STAN has competitive performances among salient object detection methods.

1. Introduction

The purpose of salient object detection is to precisely and uniformly identify the most visually distinctive regions in an image or video. It has become a very active research topic in computer vision, because it is able to support high-level visual tasks, such as segmentation [65, 44, 60, 43, 63], visual tracking [66, 8, 9], thumbnail creation [64] and photo cropping [58]. In the community of saliency detection, we can categorize it into image salient object detection and video salient object detection by the input of saliency models. In the past decades, image salient object has been widely studied, while video salient object detection attracts less attention due to the difficulty of extracting robust motion information and the shortage of large-scale annotated datasets.

Recently, many saliency detection approaches have introduced the deep learning models, which can substantially improve their performance in static images. Especially, with the proposal of the fully convolutional network (FCN), more and more end-to-end deep learning based saliency methods have been proposed. The efficiencies of the approaches also have enormously boosted. However, these deep learning models cannot adapt directly to the video saliency detection, even if transfer learning is introduced into this community. The difficulties behind this phenomenon are three-fold:

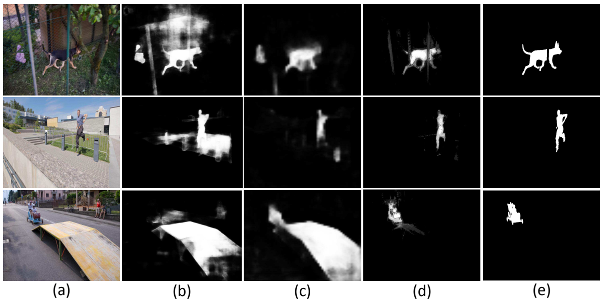


Figure 1: Salient object detection by using different deep learning approaches. (a) Video frames. (b) Results by the approach in still images [71]. (c) Results by the approach in video sequences [62]. (d) Results of the proposed network. (e) Ground truths.

The first one is that the traditional saliency networks in static images cannot capture motion cues in video frames. These networks are only able to learn the spatial deep features in an image space. Due to the lack of motion information, these deep features cannot completely distinguish the salient and non-salient regions in the complex video scenes (e.g., Figure 1 (b)). Secondly, there is not sufficient exploration in the aspect of spatiotemporal fusing features. Recent work [62] tries to build a suitable deep network to extract robust spatiotemporal features and fuse them in an efficient way. However, their stepwise structure cannot entirely eliminate non-salient regions in some videos (e.g., Figure 1 (c)). Last but not least, unlike the deep networks in image saliency inference, the deep network of video saliency is the lack of sufficient pixel-wise labeled training data. Compared

Email address: yitang@szu.edu.cn(Y. Tang).
zouszu@sina.com(W. Zou).
Y.Hua@qub.ac.uk(Y.Hua).
jinzhi_126@163.com(Z. Jin).
lixia@szu.edu.cn(X. Li).

*Corresponding author

ORCID(s):

with the image saliency datasets, current datasets for video saliency estimation have very limited pixel-wise ground truths. The total number of labeled data in the widely used video datasets (e.g., SegTrackV2 [24], FBMS [1], DAVIS [40]) is less than 5000 frames (including the data for testing). Moreover, the pixel-wise labels of some video sequences are discontinuous.

Based on the aforementioned issues, on the one hand, we try to build a deep neural architecture, which is not only able to fully extract robust spatial and temporal deep feature maps, but also fuses them in an efficient way. On the other hand, the sufficient pixel-wise labeled video frames are obtained to support the network training. Therefore, we propose a spatiotemporal attention neural network (STAN) for video salient object detection and produce an amount of coarse pixel-wise labels for network training. In our framework, we extract the motion information from two aspects. They are a motion prior based on the optical flow and the deep motion features learned from the LSTM structure and 3D convolutional operation. Meanwhile, an attentive module is employed to fuse the spatial and temporal cues by learning the weights from different types of features. As for the pixel-wise labeled training samples, we introduce a coarse labeling strategy, which fuses saliency maps from different saliency models, and then manually erases the error detection regions. Through this strategy, we can produce rapidly a number of coarse pixel-wise labels to support the training of the proposed network.

In summary, we can conclude three contributions of this paper as follows:

- We propose a spatiotemporal attention neural network (STAN) for saliency estimation. This architecture not only retains original spatial cue, but also effectively extract temporal deep features in terms of optical flow and the video sequences.
- In STAN framework, we introduce an attentive module to learn the weights of spatial and temporal features. Further, it can help the network to learn the complementary correlation between spatial and temporal information and generate more robust fusing spatiotemporal features.
- Following the labeling criteria of [1], we choose some videos in FBMS dataset and label the frames in a coarse pixel-wise way. These labels are able to effectively support the network training, thus learn robust salient features.

This work extends from a conference paper [49] by further exploiting the motion information and attentive module in the proposed framework. More detailed experiments are conducted to validate the effectiveness of the different components and some extra amplified instructions are provided in this paper. The remaining sections can be organized as follow. Firstly, we briefly review the related works in Section 2. Secondly, we describe detailedly the proposed STAN framework in Section 3. Then, the experiments and comparisons are presented in Section 4. Finally, we conclude the proposed approach in Section 6.

2. Related works

2.1. Salient object detection in still images

Over recent decades, a variety of techniques and theories have been exploited to detect salient objects in still images. The traditional approaches mainly employ handcrafted features and space constraint models to estimate salient regions. These approaches contain regional contrast [14, 19, 41, 51], low-rank matrix recovery [77], graphical modeling [31], spatial prior [39] and so on. With the resurgence of convolutional neural network, deep learning based saliency methods have gradually become the mainstream. At the beginning, the pre-trained deep models are employed to extract the deep features of the superpixels in an image. In [73], Zhao et al. propose a multi-context framework to estimate saliency by using the deep features of the image patches. Meanwhile, Wang et al. [53] introduce a local estimation and global search approach, which combines the region proposal and the deep feature of superpixels. After that, with the wide employment of the end-to-end framework FCN [34], the performance and efficiency of saliency detection have dramatically been improved. Liu et al. [32] modify the original FCN and propose a novel deep hierarchical structure (DHSN) to conduct saliency prediction. Li et al. [27] try to introduce the dilated convolution to ensure the suitable receptive field in their saliency network. In the network optimization, DSMT [30] proposes a multi-task network, which integrates the cross-entropy loss and Euclidean loss. RFCN [55] introduces a recurrent-based module, which incorporates a salient prior map to detect salient objects. Hou et al. [16] propose a new structure called short connection. This approach can efficient condense the multi-scale features and achieve good performance in saliency detection. ASNet [56] exploits the fixation prediction as a guidance to complete the saliency inference.

2.2. Salient object detection in video scenes

Due to the lack of motion information, the approaches of image salient object detection cannot be directly adapted to video scenes. Moreover, motion information plays a key role in video saliency estimation and directly affects the quality of the final saliency maps. Some traditional video saliency models [59, 21, 67, 15] mainly employ optical flow to extract the motion information. Besides, some novel spatiotemporal fusing models [33, 61, 3, 74] are progressively proposed. The combination of the motion information and the energy function achieves substantial development of the video saliency inference. However, the employment of optical flow and complex optimization model makes these methods suffer from heavy computational burden. Additionally, the traditional handcrafted low-level features cannot completely handle some complicated video scenes. It is similar to image salient detection that the exploitation of deep learning brings a breakthrough in video saliency inference. In [23], instead of handcrafted features, deep features are combined with spatiotemporal conditional random field to highlight the salient objects. Meanwhile, a FCN-based deep network [62] is proposed to generate saliency maps. To get rid of time-consuming motion computation, without using the optical flow, this network directly takes two successive frames as input to extract the motion information. Due to the shortage of pixel-wise labeled ground truths for network training, Wang et al.[62] adopt a synthetical strategy, which leverages optical flow to generate large-scale labeled video data. Li et al. [29] build a new video-based salient object detection dataset, which contains 200 video sequences and 7,650 pixel-wise ground truths. However, these labeled video sequences are still discontinuous and not suitable for the sequential module of the neural network. In [26], an end-to-end recurrent neural network is proposed. It achieves the competitive performance by extracting motion information from the optical flow and sequential features using LSTM framework. In order to solve the shortage of labeled video data, Wang et al. not only collect a scalable and diverse dataset DHF1K, but also build a large-scale benchmark in [57]. PDB [45] tries to find a more suitable dilated convolutional module and then propose the pyramid dilated deeper ConvLSTM in video saliency community.

2.3. Attention mechanisms

Attention mechanism is employed to guide human's gazes on relevant parts and predict a weighting of CNN output. Recently, combined with convolutional neural network, attention mechanism has been widely applied for many fields of computer vision, such as image captioning [35, 4], visual question answering [35, 70], fine-grained image recognition [12], etc. In [7], an LSTM-based saliency attentive model is proposed to iteratively focus on relevant locations of the image to refine salient features. In [25], an attentive module is incorporated with CNN to learn the weights of multi-scale feature maps, then the element-wise production is introduced to fuse these feature maps to generate the final prediction. Pei et al. [37] modify the LSTM unit. An attention gate is embedded in LSTM to learn the hidden representation for the video classification.

In addition, spatiotemporal attention is also employed for different kinds of visual tasks [46, 13, 76, 69, 38]. In [46], Song et al. propose a joint spatial and temporal attention model to conduct human action recognition. In person re-identification field, STA [13] proposes a simple yet effective spatiotemporal attentive model to achieve discriminative parts mining and frame selection. Xu et al. [68] propose a spatiotemporal attention pooling to learn the representation of video sequences. In total, spatiotemporal attention is usually introduced to learn weights of spatiotemporal features and fuse them in a suitable way. In this paper, we also try to exploit this characteristic in video saliency estimation.

2.4. Video object segmentation

Video object segmentation can be divided into two categories: semi-supervised video object segmentation [60, 2, 52] and unsupervised video object segmentation [18, 50, 45]. The semi-supervised video segmentation is able to use the ground truth of the first frame to obtain the specific object information, and then segments out the objects in a video sequence. However, the object information of the first frame cannot be given to unsupervised video object segmentation. Therefore, the target of video salient object detection is very similar to unsupervised video object segmentation. The former is to estimate the probability value of each pixel and the latter is to conduct a binary classification of them.

The development of unsupervised video segmentation is also similar with video saliency. Before the introduction of deep learning, unsupervised video segmentation has usually employed heuristic methods. For example, SAGE [36] proposes to exploit an estimation based on motion boundaries to bootstrap an appearance saliency model. Based on the geodesic distance, SA [59] builds a spatiotemporal boundary constraint to detect salient objects. With the employment of neural network, this community has further developed. In [50], Tokmakov et al. firstly exploit FCN and optical flow to generate a coarse saliency estimation, and then objectness map and a conditional random field are combined to further improve the labeling. FusionSeg [18] proposes an end-to-end convolutional network, which extracts robust features from optical flow and video sequence and fuses them together at the top of the network.

3. The proposed approach

Different from the cascade structure in [48], we propose a two-stream deep network to estimate salient regions. The specific framework are shown in Figure 2, which consists of four components, two network streams without shared parameters, refinement of motion information and attentive module. Both of streams are the modified VGG-based FCN, whose convolutional layers are replaced by the dilated ones in the last two convolutional blocks. The only difference between the two streams is the network input. In the spatial stream, the RGB video sequence is fed into the network. In the temporal stream, we treat the RGB frames and corresponding motion priors as the network input. The temporal prior is generated by [48]. Specifically, the superpixels of the optical flow map are firstly obtained by using [10]. Then, the deep features of these superpixels are extracted by AlexNet. We use these deep features to implement a three fully connected layers neural network to estimates the salient value of superpixels. After the feature extraction from the two streams, the spatial and temporal feature maps are concatenated and fed into the ConvLSTM units and a 3D convolutional layer to further refine motion information. After that, an attentive module is exploited to learn the weights of spatiotemporal feature maps and a multiplication operation is employed to fuse them. At last, by means of a pixel-wise weighted sum, the final saliency map can be generated from these spatiotemporal feature maps at the top of the proposed network.

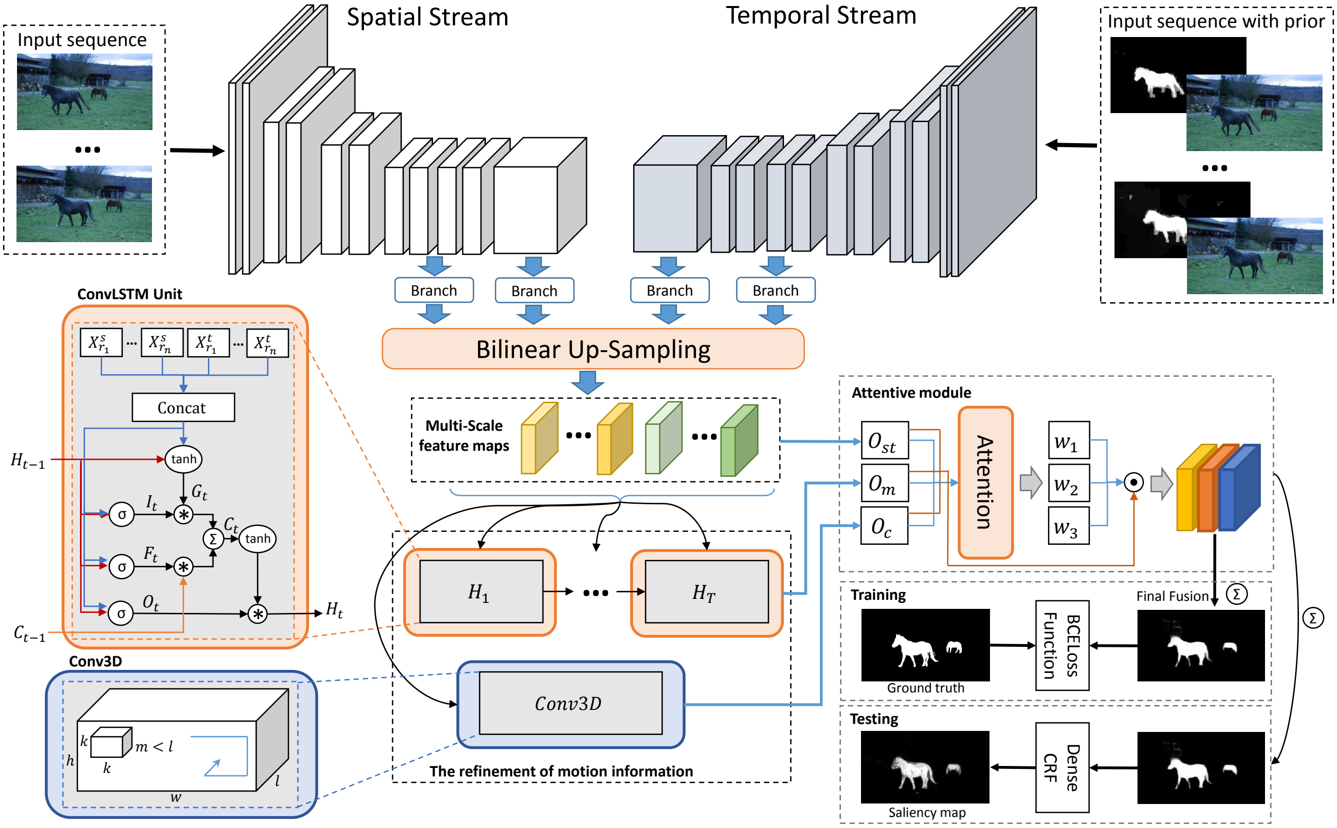


Figure 2: The framework of the proposed spatiotemporal attention neural network. The multi-level feature maps are generated from spatial stream and temporal stream, which are trained by video frames and video frames with the corresponding motion prior, respectively. With the refinement of ConLSTM units and 3D convolutional operation, different kinds of feature maps are fed into a attentive module to learn the weights of spatiotemporal feature maps. Through a average sum of weighted feature maps in a pixel-wise manner, the final saliency map can be obtained.

3.1. The spatiotemporal attention neural network

As we known, the original VGG network contains six convolutional blocks. Moreover, with the increase of the pooling layers, the scale of the generated feature maps are also gradually decreasing, which is not suitable to obtain dense features in fully convolutional networks. In this paper, by following the [5], we modify the stride to 1 at the last two pooling layers. Additionally, the dilated convolutional layers are embedded into the last two convolutional blocks to generate dense features maps and retain the original receptive field. After that, these dense feature maps are resized with the same size of the input and fed into the module of the refinement of motion information.

In the bottom-left corner of Figure 2, we can see the specific module of the refinement of motion information. It consists of ConvLSTM units and 3D convolutional operation (Conv3D). Given the input sequences of both streams I^s and I^t, we can obtain the feature maps of different levels:

$$\begin{aligned} X^s_{r_i} &= \mathcal{U}^s_{r_i}(\mathcal{F}^s_{r_i}(I^s;\theta^s_{r_i})) \\ X^t_{r_i} &= \mathcal{U}^t_{r_i}(\mathcal{F}^t_{r_i}(I^t;\theta^t_{r_i})), i \in \{1,2,...,n\} \end{aligned} \tag{1}$$

where $X^s_{r_i}$ and $X^t_{r_i}$ represent the feature maps at different level r_i from spatial and temporal stream, respectively. $\mathcal{F}^s_{r_i}(\cdot)$, $\mathcal{F}^t_{r_i}(\cdot)$ are the convolutional operations and $\mathcal{U}^s_{r_i}(\cdot)$, $\mathcal{U}^s_{r_i}(\cdot)$ denote the bilinear up-sampling operations in two streams. After the extraction of feature maps, all of them are stacked to X^τ ($\tau \in \{1,2,...,T\}$) by a concatenation operation **Con**:

$$X^\tau = \mathbf{Con}(X^s_{r_1},...,X^s_{r_n},X^t_{r_1},...,X^t_{r_n}) \tag{2}$$

The concatenated feature maps $X^1,...,X^T$ (T is the timestep) are the input of the ConvLSTM $\mathcal{M}(\cdot)$ and Conv3D $\mathcal{C}(\cdot)$. The overall sequential operation can be listed in Eq.3, where O_m denotes the many-to-one output of ConvLSTM, O_c denotes the output of Conv3D operation, θ_m and θ_c are the parameters of ConvLSTM and Conv3D, respectively:

$$\begin{aligned} O_m &= \mathcal{M}(X^1,...,X^T;\theta_m) \\ O_c &= \mathcal{C}(X^1,...,X^T;\theta_c) \end{aligned} \tag{3}$$

The ConvLSTM structure is composed of three gate operations, which can be formulated as follow:

$$\begin{aligned} I_t &= \sigma(W_{xt} * X_t + W_{ht} * H_{t-1} + b_i) \\ F_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f) \\ O_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o) \\ C_t &= F_t \circ C_{t-1} + I_t \circ tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ H_t &= O_t \circ tanh(C_t) \end{aligned}$$

$$(4)$$

where the first three equations are the input gate, forget gate and output gate. $W_{(\cdot)}$, $b_{(\cdot)}$ and $\sigma(\cdot)$ are the corresponding parameters, biases and sigmoid activation function, respectively. X_t denotes the input of ConvLSTM structure, H_{t-1} is the previous hidden state. C_t and H_t represent the current memory cell and hidden state. Different from the original LSTM, ConvLSTM exploits the convolutional operation '$*$' to compute the value of the gates. Then, the current state can be obtained by the Hadamard production '$\circ$'.

At the fusing phase of deep features, an 1-channel convolutional layers $\mathcal{W}^{st}_{r_i}$ with kernel size of 1 × 1 are firstly used to fuse the multi-level feature maps from spatial and temporal stream, respectively. Then, we sum them up element-wisely to obtain the sub-spatiotemporal feature maps O_{st}. This operation can be formulated as below:

$$O_{st} = \sum_{i}^{n} \mathcal{W}^{st}_{r_i} * \mathbf{Con}(X^{s}_{r_i}\ X^{t}_{r_i}) \tag{5}$$

At the top of the proposed STAN, an attentive module is exploited to generate the final saliency estimation $\mathcal{S}$ with the three kinds of the feature maps O_{st}, O_m, O_c. This module is presented in detail in Section 3.2.

To obtain the gradients and train the parameters of the entire network, we employ a binary cross entropy loss function $\mathcal{L}$ to optimize the proposed STAN:

$$\begin{aligned}\mathcal{L}(\mathcal{S},\mathcal{G}) = &- \sum_{i=1} g_i \log P(s_i = 1|I^s_i, I^t_i;\Theta) \\ &- \sum_{i=1} (1-g_i) \log P(s_i = 0|I^s_i, I^t_i;\Theta)\end{aligned} \tag{6}$$

where $\mathcal{S}$ and $\mathcal{G}$ denote the saliency prediction from the proposed STAN and ground truth, respectively. s_i and g_i demonstrate the saliency value and the label for a pixel in $\mathcal{S}$ and $\mathcal{G}$; Θ is the parameter of the proposed network; $P(\cdot|\cdot)$ represents the confidence probability of the prediction.

3.2. The spatiotemporal attentive module

At the bottom of the proposed network, three kinds of feature maps (O_{st}, O_m, O_c) are extracted from two stream, ConvLSTM and 3D convolutional operation, respectively. We hope to find a suitable fusing method and obtain robust features to estimate the final saliency map. A straightforward way is to use element-wise addition [27], but it cannot fully utilize the contextual information and complementary correlation. Hence, as shown in Figure 3, we introduce a spatiotemporal attentive module to integrate all kinds of feature maps.

Given the feature maps (O_{st}, O_m, O_c) as an input, the whole operation of the spatiotemporal attentive module can be formulated as below:

$$\mathcal{S} = \sum_{a\in\{st,m,c\}} w_a \odot O_a \tag{7}$$

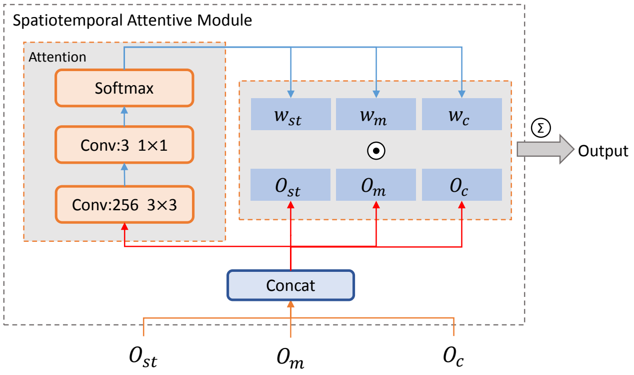


Figure 3: The architecture of the spatiotemporal attentive module.

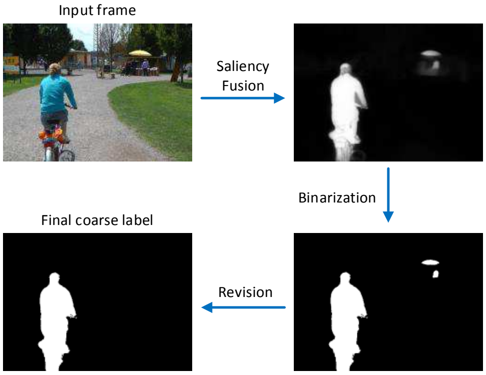


Figure 4: The generation of coarse pixel-wise labels.

where O_a represents the feature maps from different spatiotemporal modules and w_a denotes the corresponding learned weights, which indicate the attentive extent at different kinds of feature maps.

Specifically, to learn the attentive weights, we firstly concatenate all kinds of feature maps and feed them into two convolutional layers, whose parameters are 256 channels with 3 × 3 kernels and three channels with 1 × 1 kernels (shown in Figure 3). Then, the learned weights are obtained through a softmax activation function. At last, an element-wise multiplication '$\odot$' and an element-wise addition '$\sum$' are employed to generate the final saliency map.

3.3. The coarse pixel-wise labeling

The data-driven deep learning approaches require large-scale labeled data to train the networks. However, in the field of video salient object detection, the current datasets have very limited pixel-wise ground truths. However, it is very time-consuming to label new pixel-wise labeled video sequences. Therefore, to obtain sufficient pixel-wise labeled video sequences, we introduce a coarse pixel-wise labeling strategy, which needs little manual interference and is able to quickly obtain an amount of consecutive coarse labels. Together with the original ground truths of video frames, we can train a more robust deep network.

As shown in Figure 4, we firstly produce the fused saliency maps of video frames by a weighted linear combination [20] of three competitive saliency approaches in still images (i.e., DCL [27], RFCN [55], DSMT [30]). These fusing saliency maps exist some background noise in the complex scenes, but the main salient regions can be highlighted. Secondly, an Ostu thresholding method is exploited to generate the binary pixel-wise labels. At last, the existing background noise is manually erased to generate the final coarse pixel-wise labels. Through this strategy, we can quickly obtain a number of pixel-wise labeled data (totally 3,326 frames) to support the network training.

4. Experiments

In this section, the brief instruction of the used video saliency datasets and evaluation criteria are firstly presented. Then, we give the implementation details of the proposed STAN. After that, we give the specific comparison between the proposed STAN and the state-of-the-arts. Besides, the effectiveness of each module is also reported. In the end, we have an analysis of the failure cases and runtime complexity.

4.1. Datasets and performance evaluation criteria

In our experiments, three widely used datasets are exploited to train and validate the proposed STAN. The datasets are SegtrackV2 [24], FBMS [1] and DAVIS [40].

SegtrackV2 consists of 14 video sequences and totally contains 1,066 frames. Though each frame is manually pixel-wisely labeled for salient objects, most of the sequences are very short, i.e., appropriately 100 frames per sequence.

FBMS contains 59 video sequences, 13,960 frames in total. In this dataset, 29 videos are used for traning and the remains are used for testing. Although FBMS have sufficient video sequences, piexl-wise ground truths are discontinuous.

DAVIS has 50 video sequences and each frame has the corresponding pixel-wise ground truth. This dataset includes a lot of complex scenes, which makes it challenging for saliency detection.

In our experiments, our training data includes the training set of DAVIS and FBMS (with generated coarse pixel-wise labels) and all of SegTrackV2 videos. The testing set is the remaining video sequences of DAVIS and the testing set of FBMS (with ground truths).

As for quantitative evaluation, we adapt precision-recall (PR) curve, mean absolute error (MAE) and F-measure to validate the proposed STAN on the above datasets. The detailed experimental results are shown in the latter sections.

4.2. Implementation

To train an efficient neural network and obtain robust spatiotemporal features, we implement the proposed STAN following next several settings:

- The pre-training VGG-based FCN from [27] is introduced to initialize the two streams in our STAN. However, as the 4-channel input of the temporal stream, a Gaussian distribution is used to re-initialized the first convolutional layer.
- In our STAN network, several branches are embedded to extract the multi-level feature maps. Specifically, we set two branches in each stream. The first one is embedded after the fourth pooling layers, the other is after the last convolutional layer.
- In the training phase, due to the limitation of GPU memory, the resolution of input images is resized at 512 × 512. Besides, the time span of the training frames is set to 4.
- In the testing stage, we exploit a dense CRF [5] as a post-processing method to improve spatial coherence and refine the generated saliency maps.
- We employ the Adaptive Moment Estimation (Adam) to optimize the network in the entire training process. The initial learning rate is set to 10^{-3} to train the proposed STAN.

4.3. Comparison to the state-of-the-art saliency models

In this section, 19 recent state-of-the-art approaches are compared against the proposed STAN. Among them, there are ten state-of-the-arts in video scenes and nine approaches in still images. The video salient object detection includes cluster-based co-saliency method (CS) [11], space-time saliency detection (ST) [75], segmenting saliency detection (SS) [42], saliency-aware method (SA) [59], consistent gradient based saliency (CG) [61], video salient object detection via fully convolutional networks (SFCN) [62], multi-scale spatiotemporal network (MSST) for video saliency [49], spatiotemporal cascade network (SCNN) [48], flow guided recurrent neural encoder (FGRNE) [26] and pyramid dilated deeper convlstm (PDB) [45]. The image saliency models include recurrent fully convolutional network (RFCN) [55], deep contrast learning (DCL) [27], visual saliency on multi-scale deep features (MDF) [28], deep saliency multi-task (DSMT) [30], deeply supervised salient object detection (DSS) [16], deep hierarchical saliency network (DHSN) [32], aggregating multi-level convolutional features (Amulet) [71], saliency detection with image-level supervision (WSS) [54] and learning uncertain convolutional features (UCF) [72].

The results on the PR curves, MAEs and F-measures of the comparative approaches and the proposed one are shown in Figure 5 and Table 1. As we can see, although PDB is better, our approach achieves competitive performance on both datasets in terms of PR curves, MAEs and F-measures. Figure 8 shows the qualitative comparison of the saliency maps generated by the compared 19 models on the two datasets. The first three sequences (*soapbox, dancetwirl, horsejump*) are from DAVIS and last three sequences (*horse04, horse05, cats06*) are from FBMS. Based on these examples, we can have three statements as follows:

Multiple salient objects: The *horse04* and *horse05* sequences contain multiple salient objects. The baseline ap-

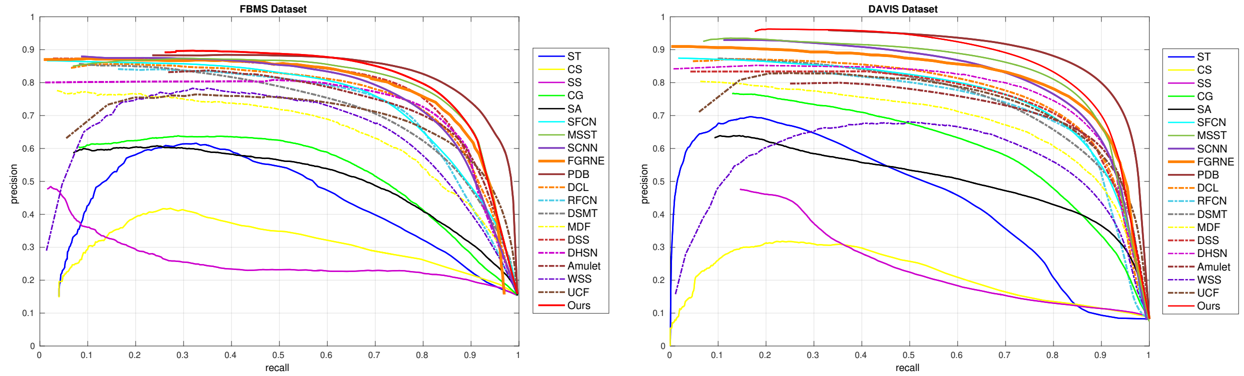


Figure 5: The quantitative comparison between the state-of-the-art saliency detection approaches and the proposed one. The comparative approaches contain 10 video saliency methods (solid lines) and 9 image saliency methods (dashed lines). The PR curves on FBMS and DAVIS are shown at left and right, respectively.

Table 1
Comparison of F-measure and MAE from the different video and image salient object detection methods on FBMS and DAVIS datasets.

Method		ST	CS	SS	CG	SA	SFCN	MSST	SCNN	FGRNE	PDB
FBMS	F-measure ↑	0.581	0.426	0.292	0.606	0.567	0.756	0.797	0.780	0.779	0.815
	MAE ↓	0.180	0.176	0.378	0.172	0.182	0.102	0.088	0.105	0.083	0.069
DAVIS	F-measure ↑	0.519	0.387	0.376	0.627	0.528	0.749	0.817	0.774	0.786	0.849
	MAE ↓	0.140	0.115	0.432	0.095	0.106	0.055	0.045	0.074	0.043	0.030
Method		DCL	RFCN	DSMT	MDF	DSS	DHSN	Amulet	WSS	UCF	Ours
FBMS	F-measure ↑	0.774	0.757	0.727	0.674	0.788	0.760	0.747	0.707	0.716	0.812
	MAE ↓	0.153	0.108	0.123	0.134	0.082	0.086	0.110	0.117	0.150	0.087
DAVIS	F-measure ↑	0.755	0.731	0.734	0.684	0.748	0.785	0.723	0.675	0.741	0.834
	MAE ↓	0.132	0.068	0.087	0.102	0.066	0.042	0.0837	0.0733	0.108	0.041

proaches are only able to detect part of them. By fully refining the spatiotemporal features, the proposed STAN can completely highlight all of the salient objects from these kinds of videos.

Tiny salient objects: In some video scenes (e.g., the *cats06* sequence), the salient objects are sometimes very small, which often leads to the failure of saliency detection. Due to the suitable architecture and sufficient training data, the proposed STAN can highlight the tiny objects.

Motion features and attentive module: The proposed STAN effectively extracts motion information from optical flow, ConvLSTM units and 3D convolutional operation. Besides, an attentive module is introduced to learn the weights of different kinds of features and fuse them in a practical way, which can highlight precisely the moving objects and eliminate background noise (e.g., the fence in *horsejump* sequence) as much as possible.

4.4. Ablation studies

As described in Section 3, the proposed deep learning architecture contains five modules such as spatial stream, temporal stream, ConvLSTM module, 3D convolutional operation and spatiotemporal attentive module. In order to validate the availability of each module, we conduct the experiments with different configurations of the network.

- *OS*: The spatial stream is only used to complete the saliency prediction. We train this model only with video frames and do not encode any motion information.
- *ST*: Both of two streams are exploited to extract the multi-level feature maps. Moreover, these feature maps are exploited to estimate the saliency value by element-wise addition at the top of the network.
- *STC*: As a comparison, the motion refinement module only employs the 3D convolutional operation to extract temporal features. The fusing method of different type feature is element-wise addition as well.
- *STL*: In this scenario, the multi-level feature maps from two streams are fed into the ConvLSTM units to further refine the motion information. Then, the feature maps from two streams and ConvLSTM units are fused to generate the final saliency map by element-wise addition.
- *STLC*: Along with STL scenario, a 3D convolutional operation is jointed to produce motion features. Furthermore, all kinds of feature maps are element-wisely sum up at the top of the network.

As shown in Figure 6(a), we validate the significance of different components step by step. At first, we exploit the

Table 2
Comparison with the unsupervised video object segmentation approaches on the DAVIS test set.

Dataset	Metric	CUT	FST	SFL	LMP	FSEG	STAN
DAVIS	$\mathcal{J}$ mean	55.2	55.8	67.4	70.0	70.7	71.1
	$\mathcal{J}$ recall	57.5	64.9	81.4	85.0	83.5	81.6
	$\mathcal{J}$ decay	2.2	0.0	6.2	1.3	1.5	0.0
	$\mathcal{F}$ mean	55.2	51.1	66.7	65.9	65.3	67.0
	$\mathcal{F}$ recall	61.0	51.6	77.1	79.2	73.8	77.6
	$\mathcal{F}$ decay	3.4	2.9	5.1	2.5	1.8	0.1

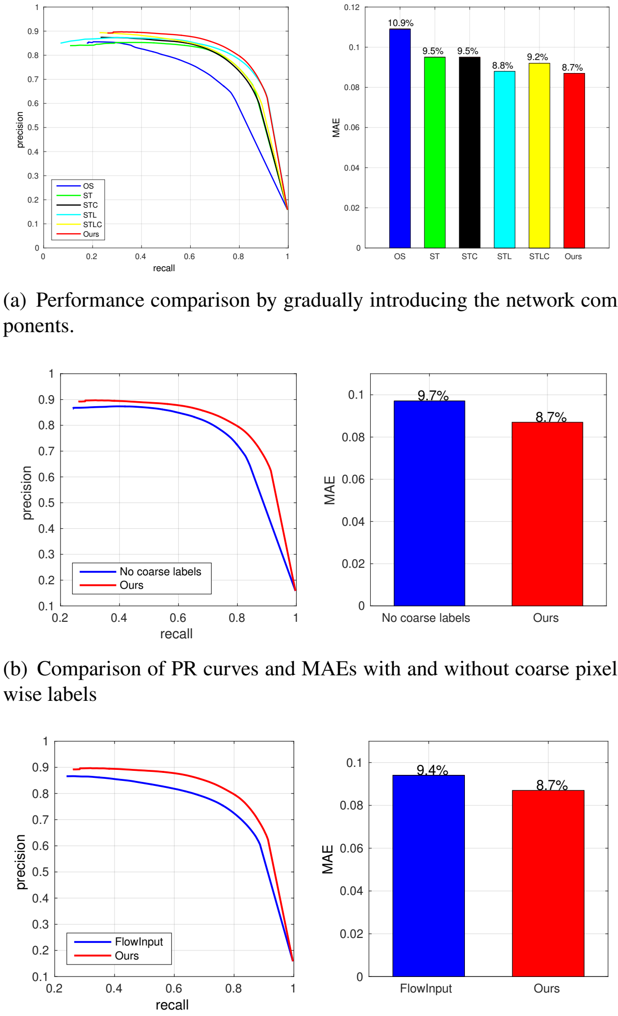
spatial stream to estimate saliency. Due to the shortage of motion information, the saliency result exists serious shortcoming. After that, we combine spatial and temporal stream to conduct saliency prediction, which makes much improvement on the performance. With the employment of 3D convolution or ConvLSTM units, the saliency result is still enhanced to some extent. However, without suitable module to fuse the features by ConvLSTM and 3D convolution, the PR curve is slightly increased in the precision side. The recall of the PR curve and MAE are worse than the previous scenario. The reason is that the different types of feature maps cannot be fully fused by the simple element-wise addition. Further, the contextual information and complementary correlation are also inadequately explored. Therefore, we introduce a spatiotemporal attentive module to solve the issue. The final results in the PR curve and MAE prove the efficiency of the proposed attentive module.

To prove the effectiveness of the coarse pixel-wise labels, we design a contrast experiment by training the STAN with and without the coarse labels. The experimental results are shown in Figure 6(b). The performance of PR curve and MAE proves that the generated coarse pixel-wise labels are very helpful to the proposed STAN. Though the exploitation of the coarse labels, the network training data are rising. Additionally, many complex video scenes are added into the training set, which can make the network learn more abundant salient information and then improve the performance of saliency inference.

In our network, we use a four-channel image, that stacks a video frame and an optical flow-based motion prior map, as the input of temporal stream rather than the original optical flow map. In order to prove that this scenario is more suitable to the proposed network, we set an experiment to train the whole network by using the original optical flow map and the four-channel image, respectively. The PR curves and MAEs are displayed in Figure 6(c). As we can see, the proposed scenario by stacking video frame and an optical flow-based motion prior achieves better performance.

4.5. Comparison with unsupervised video object segmentation approaches

The purpose of video salient object detection is very similar to that of video object segmentation. The former is to estimate the salient values of the corresponding pixel in an video frame. The latter is to obtain the results of binary classification. Therefore, we provide the comparison experiments with the evaluation criteria of video segmentation in this section. The criteria includes the mean, recall and decay of both the intersection-over-union metric ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$) metrics.



(a) Performance comparison by gradually introducing the network components.

(b) Comparison of PR curves and MAEs with and without coarse pixel-wise labels

(c) Performance validation between the usage of original optical flow map and motion prior-based four-channel image

Figure 6: The quantitative comparison with different configurations on FBMS dataset.

Table 2 shows the results of the proposed STAN and the comparative methods. These methods include CUT [22], FST [36], SFL [6], LMP [50] and FSEG [18]. As shown in Table 2, our approach has competitive performance in segmentation evaluation, which proves that the proposed refinement of motion information and the attentive module are effective.

4.6. Runtime analysis

We run all of the approaches on a GPU workstation with an Intel(R) i7-5820 CPU (3.3 GHz), a Nvidia Geforce TITAN X GPU (12 GB memory), and 64G RAM. Table 3 and Table 4 present the average run time per frame of different approaches and the run time of each module of the proposed STAN on DAVIS dataset, respectively. As we can see, due to the exploitation of optical flow, SS, SA and CG are very

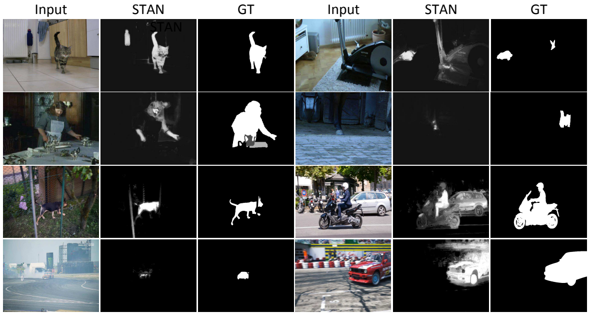


Figure 7: Some failure detection by our approach on FBMS and DAVIS datasets.

time-consuming. In the proposed STAN, we also use the optical flow to produce the motion prior. To decrease the computation of optical flow, instead of the traditional method of optical flow extraction [47], we employ the FlowNet2.0 [17], a deep learning-based model, to obtain optical flow maps. The generated optical flow maps are not only accurate but also fast. Quantitatively, FlowNet2.0 can directly decreases the computation time of optical flow from over 36s to 0.739s. Therefore, the speed of the proposed STAN can reach 2.144s per frame.

4.7. Analysis of failure detection

Although the proposed STAN can handle most of video sequences, there are still some failure cases on both two datasets. Figure 7 shows the failure examples by the proposed STAN. The reasons of these failure detection are two aspects. Firstly, in some complex scenes, as the problems of illumination and color contrast, the salient objects are very similar to some background regions, which affects directly the network to extract spatial features. Secondly, the motion blur has negative effect on the extraction of optical flow. The inaccurate optical flow leads to the robustness of temporal features. As the impact of spatiotemporal deep features, the proposed STAN fails to detect the entire salient regions in some video frames. In the future, we will try to overcome these difficulties by extracting more robust spatiotemporal features.

5. Conclusion

In this paper, we propose a novel spatiotemporal attention neural network for video salient objects detection. Our network is composed of the complementary components, two network streams, ConvLSTM units, 3D convolutional operation and spatiotemporal attentive module. The network can effectively extract spatial features and robust motion information. Besides, through the spatiotemporal attentive module, the proposed STAN is able to further integrate the spatial and temporal cues to generate high-quality saliency maps. Meanwhile, to ensure the network training, a number of coarse pixel-wise labels are generated to improve the robustness of the proposed network. In the end, the experiments on the FBMS and DAVIS indicate that our STAN can achieve competitive performance than the other methods in evaluation criteria of the PR curve, MAE and F-measure.

Table 3
Comparison average run time (seconds per frame) on DAVIS dataset.

Method	ST	CS	SS	CG	SA
Time(s)	28.193	1.175	37.176	38.075	38.751
Method	SFCN	MSST	SCNN	PDB	DCL
Time(s)	0.473	2.021	2.53	0.05	0.670
Method	RFCN	DSMT	MDF	DSS	DHSN
Time(s)	4.580	0.14	11.33	0.453	0.465
Method	Amulet	UCF	WSS	STAN	
Time(s)	5.299	0.151	0.024	2.144	

Table 4
Average run time (seconds per frame) of each component in the proposed approach on DAVIS Dataset.

Model	Component	Time (s)	Ratio (%)
STAN	Optical flow computation	0.739	34.5
	Motion prior generation	0.823	38.3
	Neural network processing	0.102	4.8
	Saliency refinement	0.480	22.4
	Total	2.144	100

6. Acknowledge

In this paper, we propose a novel spatiotemporal attention neural network for video salient objects detection. Our network is composed of the complementary components, two network streams, ConvLSTM units, 3D convolutional operation and spatiotemporal attentive module. The network can effectively extract spatial features and robust motion information. Besides, through the spatiotemporal attentive module, the proposed STAN is able to further integrate the spatial and temporal cues to generate high-quality saliency maps. Meanwhile, to ensure the network training, a number of coarse pixel-wise labels are generated to improve the robustness of the proposed network. In the end, the experiments on the FBMS and DAVIS indicate that our STAN can achieve competitive performance than the other methods in evaluation criteria of the PR curve, MAE and F-measure.

References

[1] Brox, T., Malik, J., 2010. Object segmentation by long term analysis of point trajectories, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 282–295.

[2] Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L., 2017. One-shot video object segmentation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 221–230.

[3] Chen, C., Li, S., Wang, Y., Qin, H., Hao, A., 2017a. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Transactions on Image Processing 26, 3156–3170.

[4] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S., 2017b. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 6298–6306.

[5] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017c. Deeplab: Semantic image segmentation with deep convolu-

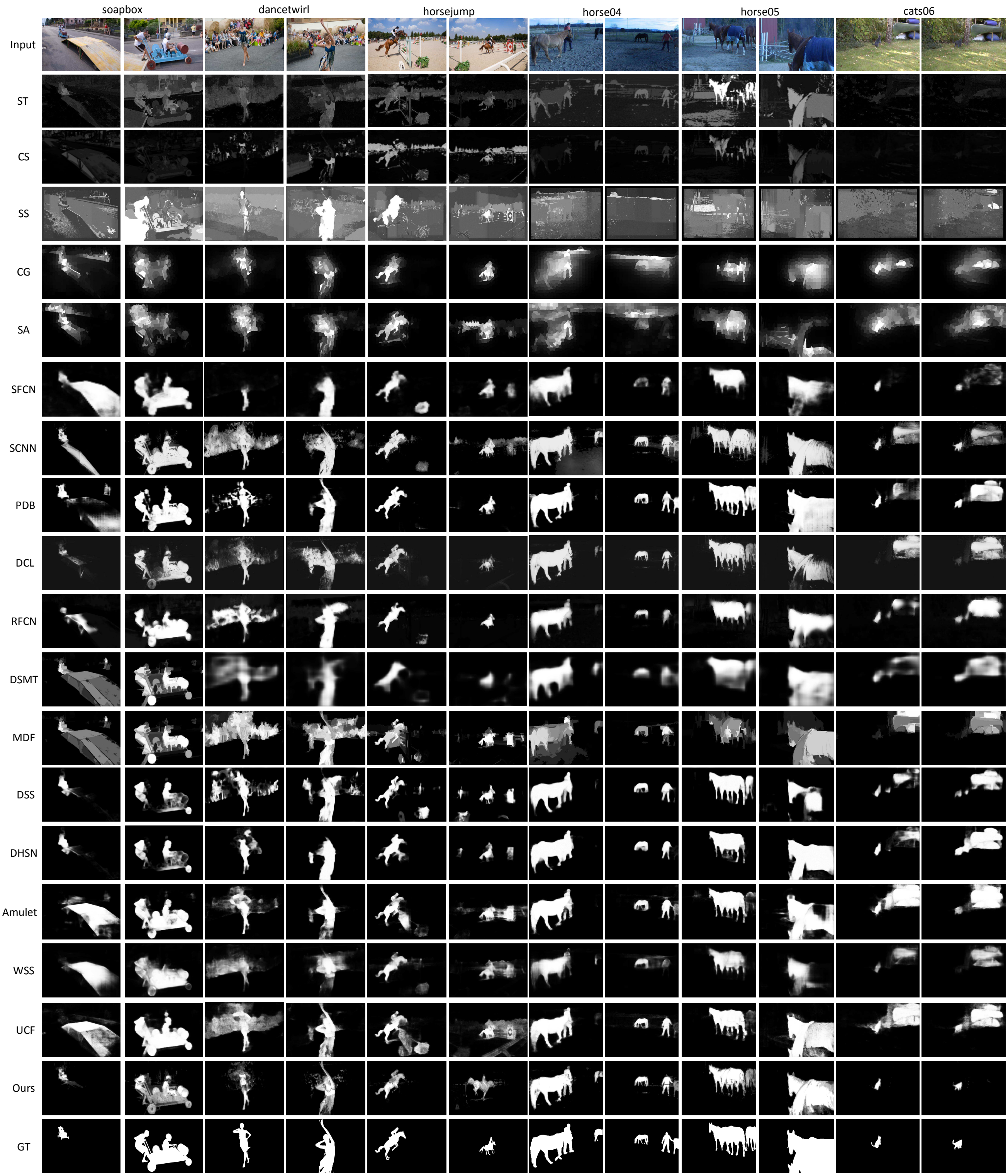


Figure 8: The qualitative comparison between other state-of-the-art methods and the proposed STAN. The left and right three columns are the video frames from DAVIS and FBMS datasets, respectively.

tional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848.

[6] Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H., 2017. Segflow: Joint learning for video object segmentation and optical flow, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 686–695.

[7] Cornia, M., Baraldi, L., Serra, G., Cucchiara, R., 2018. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. IEEE Transactions on Image Processing .

[8] Dong, X., Shen, J., Wang, W., Liu, Y., Shao, L., Porikli, F., 2018.

Hyperparameter optimization for tracking with continuous deep q-learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 518–527.

[9] Dong, X., Shen, J., Wu, D., Guo, K., Jin, X., Porikli, F., 2019. Quadruplet network with one-shot learning for fast visual object tracking. IEEE Transactions on Image Processing 28, 3516–3527.

[10] Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59, 167–181.

[11] Fu, H., Cao, X., Tu, Z., 2013. Cluster-based co-saliency detection. IEEE Transactions on Image Processing 22, 3766–3778.

[12] Fu, J., Zheng, H., Mei, T., 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 3.

[13] Fu, Y., Wang, X., Wei, Y., Huang, T., 2019. Sta: Spatial-temporal attention for large-scale video-based person re-identification, in: Proceedings of the Association for the Advancement of Artificial Intelligence.

[14] Goferman, S., Zelnik-Manor, L., Tal, A., 2012. Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 1915–1926.

[15] Guo, F., Wang, W., Shen, J., Shao, L., Yang, J., Tao, D., Tang, Y.Y., 2017. Video saliency detection using object proposals. IEEE Transactions on Cybernetics 48, 3159–3170.

[16] Hou, Q., Cheng, M.M., Hu, X., Tu, Z., Borji, A., 2017. Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE.

[17] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1647–1655.

[18] Jain, S.D., Xiong, B., Grauman, K., 2017. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, IEEE. pp. 2117–2126.

[19] Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S., 2011. Automatic salient object segmentation based on context and shape prior., in: Proceedings of the British Machine Vision on Conference, p. 9.

[20] Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S., 2013. Salient object detection: A discriminative regional feature integration approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2083–2090.

[21] Kalboussi, R., Abdellaoui, M., Douik, A., 2017. A spatiotemporal model for video saliency detection, in: Proceedings of the Image Processing, Applications and Systems, IEEE. pp. 1–6.

[22] Keuper, M., Andres, B., Brox, T., 2015. Motion trajectory segmentation via minimum cost multicuts, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3271–3279.

[23] Le, T.N., Sugimoto, A., 2018. Video salient object detection using spatiotemporal deep features. IEEE Transactions on Image Processing 27, 5002–5015. doi:10.1109/TIP.2018.2849860.

[24] Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M., 2013. Video segmentation by tracking many figure-ground segments, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE. pp. 2192–2199.

[25] Li, G., Xie, Y., Lin, L., Yu, Y., 2017. Instance-level salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 247–256.

[26] Li, G., Xie, Y., Wei, T., Wang, K., Lin, L., 2018a. Flow guided recurrent neural encoder for video salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3243–3252.

[27] Li, G., Yu, Y., 2016a. Deep contrast learning for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 478–487.

[28] Li, G., Yu, Y., 2016b. Visual saliency detection based on multiscale deep CNN features. IEEE Transactions on Image Processing 25, 5012–5024.

[29] Li, J., Xia, C., Chen, X., 2018b. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. IEEE Transactions on Image Processing 27, 349–364. doi:10.1109/TIP.2017.2762594.

[30] Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J., 2016. DeepSaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing 25, 3919–3930.

[31] Li, Z., Liu, J., Tang, J., Lu, H., 2015. Robust structured subspace learning for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 2085–2098.

[32] Liu, N., Han, J., 2016. Dhsnet: Deep hierarchical saliency network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 678–686.

[33] Liu, Z., Zhang, X., Luo, S., Le Meur, O., 2014. Superpixel-based spatiotemporal saliency detection. IEEE Transactions on Circuits and Systems for Video Technology 24, 1522–1540.

[34] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3431–3440.

[35] Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering, in: Proceedings of the Advances In Neural Information Processing Systems, pp. 289–297.

[36] Papazoglou, A., Ferrari, V., 2013. Fast object segmentation in unconstrained video, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE. pp. 1777–1784.

[37] Pei, W., Baltrušaitis, T., Tax, D.M., Morency, L.P., 2017. Temporal attention-gated model for robust sequence classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 820–829.

[38] Peng, Y., Zhao, Y., Zhang, J., 2018. Two-stream collaborative learning with spatial-temporal attention for video classification. IEEE Transactions on Circuits and Systems for Video Technology 29, 773–786.

[39] Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A., 2012. Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 733–740.

[40] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A., 2016. A benchmark dataset and evaluation methodology for video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 724–732.

[41] Qin, Y., Lu, H., Xu, Y., Wang, H., 2015. Saliency detection via cellular automata, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 110–119.

[42] Rahtu, E., Kannala, J., Salo, M., Heikkilä, J., 2010. Segmenting salient objects from images and videos, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 366–379.

[43] Shen, J., Peng, J., Dong, X., Shao, L., Porikli, F., 2017. Higher order energies for image segmentation. IEEE Transactions on Image Processing 26, 4911–4922.

[44] Shen, J., Peng, J., Shao, L., 2018. Submodular trajectories for better motion segmentation in videos. IEEE Transactions on Image Processing 27, 2688–2700.

[45] Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M., 2018. Pyramid dilated deeper convlstm for video salient object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–731.

[46] Song, S., Lan, C., Xing, J., Zeng, W., Liu, J., 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Proceedings of the Association for the Advancement of Artificial Intelligence.

[47] Sun, D., Roth, S., Black, M.J., 2014. A quantitative analysis of current

practices in optical flow estimation and the principles behind them. International Journal of Computer Vision 106, 115–137.

[48] Tang, Y., Zou, W., Jin, Z., Chen, Y., Hua, Y., Li, X., 2018a. Weakly supervised salient object detection with spatiotemporal cascade neural networks. IEEE Transactions on Circuits and Systems for Video Technology , 1–1doi:10.1109/TCSVT.2018.2859773.

[49] Tang, Y., Zou, W., Jin, Z., Li, X., 2018b. Multi-scale spatiotemporal conv-lstm network for video saliency detection, in: Proceedings of the ACM International Conference on Multimedia Retrieval, ACM. pp. 362–369.

[50] Tokmakov, P., Alahari, K., Schmid, C., 2017. Learning motion patterns in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3386–3394.

[51] Tong, N., Lu, H., Ruan, X., Yang, M.H., 2015. Salient object detection via bootstrap learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1884–1892.

[52] Voigtlaender, P., Leibe, B., 2017. Online adaptation of convolutional neural networks for video object segmentation, in: Proceedings of the British Machine Vision Conference.

[53] Wang, L., Lu, H., Ruan, X., Yang, M.H., 2015a. Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3183–3192.

[54] Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Xiang, R., 2017a. Learning to detect salient objects with image-level supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3796–3805.

[55] Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X., 2016a. Saliency detection with recurrent fully convolutional networks, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 825–841.

[56] Wang, W., Shen, J., Dong, X., Borji, A., 2018a. Salient object detection driven by fixation prediction, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[57] Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A., 2018b. Revisiting video saliency: A large-scale benchmark and a new model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4894–4903.

[58] Wang, W., Shen, J., Ling, H., 2018c. A deep network solution for attention and aesthetics aware photo cropping. IEEE transactions on pattern analysis and machine intelligence 41, 1531–1544.

[59] Wang, W., Shen, J., Porikli, F., 2015b. Saliency-aware geodesic video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3395–3402.

[60] Wang, W., Shen, J., Porikli, F., Yang, R., 2018d. Semi-supervised video object segmentation with super-trajectories. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 985–998.

[61] Wang, W., Shen, J., Shao, L., 2015c. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing 24, 4185–4196.

[62] Wang, W., Shen, J., Shao, L., 2018e. Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing 27, 38–49.

[63] Wang, W., Shen, J., Sun, H., Shao, L., 2017b. Video co-saliency guided co-segmentation. IEEE Transactions on Circuits and Systems for Video Technology 28, 1727–1736.

[64] Wang, W., Shen, J., Yu, Y., Ma, K.L., 2016b. Stereoscopic thumbnail creation via efficient stereo saliency detection. IEEE transactions on visualization and computer graphics 23, 2014–2027.

[65] Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S., 2017. Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2314–2320.

[66] Wu, D., Zou, W., Li, X., Zhao, Y., 2017. Kernalised multi-resolution convnet for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 66–73.

[67] Xi, T., Zhao, W., Wang, H., Lin, W., 2017. Salient object detection with spatiotemporal background priors for video. IEEE Transactions on Image Processing 26, 3425–3436. doi:10.1109/TIP.2016.2631900.

[68] Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P., 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in: Proceedings of The IEEE International Conference on Computer Vision.

[69] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A., 2015. Describing videos by exploiting temporal structure, in: Proceedings of the IEEE International Conference on Computer Vision.

[70] Yu, D., Fu, J., Mei, T., Rui, Y., 2017. Multi-level attention networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 4187–4195.

[71] Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X., 2017a. Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE.

[72] Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B., 2017b. Learning uncertain convolutional features for accurate saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision.

[73] Zhao, R., Ouyang, W., Li, H., Wang, X., 2015. Saliency detection by multi-context deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1265–1274.

[74] Zheng, W., Ren, J., Dong, Z., Sun, M., Jiang, J., 2018. A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing 287, S0925231218301097.

[75] Zhou, F., Bing Kang, S., Cohen, M.F., 2014. Time-mapping using space-time saliency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3358–3365.

[76] Zhu, Z., Wu, W., Zou, W., Yan, J., 2018. End-to-end flow correlation tracking with spatial-temporal attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 548–557.

[77] Zou, W., Kpalma, K., Liu, Z., Ronsin, J., 2013. Segmentation driven low-rank matrix recovery for saliency detection, in: Proceedings of the British Machine Vision on Conference, pp. 1–13.