IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 17, 2020, accepted November 4, 2020, date of publication November 6, 2020, date of current version November 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036533

Cross Complementary Fusion Network for Video Salient Object Detection

ZIYANG WANG, JUNXIA LI, (Member, IEEE), AND ZEFENG PAN

Jiangsu Key Lab of Big Data Analysis Technology (B-DAT), Nanjing University of Information Science and Technology, Nanjing 210044, China
Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China

Corresponding author: Junxia Li (junxiali99@163.com)

This work was supported in part by the National Science Fund of China under Grant 61702272, in part by the Startup Foundation for Introducing Talent of NUIST under Grant 2243141701034, and in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant SJCX20_0300.

ABSTRACT Recently, optical flow guided video saliency detection methods have achieved high performance. However, the computation cost of optical flow is usually expensive, which limits the applications of these methods in time-critical scenarios. In this article, we propose an end-to-end cross complementary network (CCNet) based on fully convolutional network for video saliency detection. The CCNet consists of two effective components: single-image representation enhancement (SRE) module and spatiotemporal information learning (STIL) module. The SRE module provides robust saliency feature learning for a single image through a pyramid pooling module followed by a lightweight channel attention module. As an effective alternative operation of optical flow to extract spatiotemporal information, the STIL introduces a spatiotemporal information fusion module and a video correlation filter to learn the spatiotemporal information, the inner collaborative and interactive information between consecutive input groups. In addition to enhancing the feature representation of a single image, the combination of SRE and STIL can learn the spatiotemporal information and the correlation between consecutive images well. Extensive experimental results demonstrate the effectiveness of our method in comparison with 14 state-of-the-art approaches.

INDEX TERMS Video saliency detection, pyramid pooling, self-attention mechanism, multi-channel concatenation, structural information.

I. INTRODUCTION

Video salient object detection (VSOD) aims at finding the most obvious object in each video group. It can be applied as a basic component in many visual tasks, such as video object segmentation [1], [2], video compression [3], object tracking [4] and so on. VSOD can be roughly divided into two categories: human eye fixation prediction and mask prediction (salient object detection). The purpose of human eye fixation prediction is to find the focus of human eyes when watching and mask prediction focuses on the most attention-grabbing objects. In this article, we focus on the latter task.

Similar to the saliency detection methods for static images, many effective VSOD methods [5], [6], [11] combine deep convolutional neural networks (CNNs) along with traditional techniques to acquire higher accuracy. For example, Chen *et al.* [7] exploit the global motion clues to guide the fusion of color saliency and motion saliency. In order to guarantee the temporal smoothness of the detection results, a low-rank coherency guided saliency diffusion strategy is designed by constructing the temporal saliency correspondence among the cross-frame superpixels. Liu *et al.* [8] propose a superpixel-based spatiotemporal saliency model, which extracts the motion histograms and the color histograms as local features at the superpixel level, and the global features at frame level. Jiang *et al.* [9] design a two-layer convolutional long short-term memory (2C-LSTM) network to learn spatiotemporal features for predicting the inter-frame saliency.

In addition to learning the detailed information and semantic information in a single image, VSOD also needs to learn the spatiotemporal information in time dimension via

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

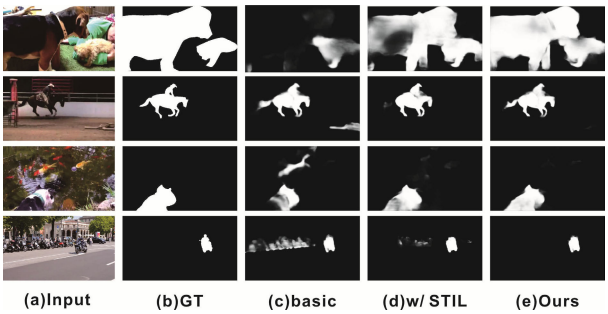


FIGURE 1. Saliency detection results by the proposed method with different settings. From left to right: (a) input image; (b) ground truth; (c) saliency predication result by the backbone; (d) saliency result by adding SITL module; and (e) our final saliency map.

considering the motion contrast and correlation between frames. So far, most methods solve the above-mentioned problems through two ways: the optical flow guided network and the ConvLSTM/ConvGRU related network. Optical flow can be used to describe the instantaneous motion state of moving objects and offer explicit motion information. Therefore, many traditional and deep learning based methods [10], [13], [14] use changes in the time domain of pixels in the image sequence and the correlation between adjacent frames to find the correspondence between the previous frame and the current frame. ConvLSTM/ ConvGRU (*e.g.*, Pyramid Dilated Bidirectional ConvLSTM (PDB-ConvLSTM) [11], bidirectional ConvGRU (DBConvGRU) [12]) are designed to capture the long and short-term memory of video sequences, fuse spatiotemporal information and learning temporal motion cues. However, the generation of optical flow map will incur significant computational cost, and the blocks based on ConvLSTM have a few cost for GPU memory in the training process.

To address the above issues, in this article, we propose an end-to-end cross complementary network to learn the spatiotemporal information jointly. Specifically, to improve the learning ability of single image representation, we use a static salient object detection network as the pre-trained model and introduce a single-image representation enhancement model. In order to enhance the spatial information learning and capture the temporal motion cues, we compress the video sequence into a five dimensional tensor and introduce it into the spatiotemporal information learning (STIL) module, which is composed of a spatiotemporal information fusion block and a video correlation filter. In addition, we design a mixed training strategy to maintain the strong feature representation ability of static image. Figure 1 shows some visual examples of the proposed method and its variants. Compared to the baseline, ''w/ STIL'' is able to roughly detect salient objects in continuous frames and our final method (''Ours'') can pop out the whole salient object regions and sharpen boundaries. The above design contributes to a powerful and very fast (with speed at 23-26 fps on GPU) deep video object detection model, which achieves the state-of-the-art performance on four popular video datasets. We consider that the proposed method is a robust network, which does not need optical flow guiding network and long short-term memory blocks such as ConvLSTM and ConvGRU. In total, the contributions of this article can be summarized as follows.

- We innovatively establish a novel end-to-end cross complementary network (CCNet) for VSOD. It is composed of a SRE module for spatial feature learning in static images, and a STIL module for the spatial feature representation enhancement and the inter-frame cues capturing.
- Considering the importance of different frames to a video group, we design a novel video correlation filter to adaptively distribute a set of weights according to the importance of input video clips and thus to optimize the feature maps respectively.
- The proposed CCNet achieves the state-of-the-art performance in terms of both accuracy and speed.

The rest of this article is organized as follows. Section II discusses related work, including the relevant models for salient object detection and video salient object detection. Section III introduces the details of the proposed video saliency model, and meanwhile describes the SRE module and the SIL module in detail. Section Section IV reports the experimental results and comparisons with the state-of-the-art methods. Section V shows the drawback of our proposed network and conclusions are finally given in Section VI.

II. RELATED WORK

A. IMAGE SALIENT OBJECT DETECTION

With the development of convolutional neural networks (CNNs), many deep learning methods [15]–[19], [25] for salient object detection have been proposed. For example, Li and Yu [20] propose a refinement method that introduces a neural network architecture and fully connected layers on top of CNNs to aggregate multiple saliency maps. Wang *et al.* [21] design a deep neural network (DNN-L) to detect local saliency and integrate it with global search, and take deep neural network (DNN-G) to predict the saliency score of each object region based on the global features. Recently, many works [23]–[25] based on fully convolutional networks (FCN) [22] have made great progress in generating pixel-wise saliency prediction. Liu and Han [26] propose a novel end-to-end deep hierarchical saliency network (DHSNet) to make a coarse global prediction by automatically learning various global structured saliency cues, and then take hierarchical recurrent convolutional neural network (HRCNN) to better refine the details of saliency map. Wang *et al.* [27] use a recurrent full convolutional network to refine previous predictions and incorporate saliency prior into the network to facilitate training and reasoning. Zhang *et al.* [28] capture context information with convolution layers in multi-scale and incorporate multi-level convolutional features with a gated bidirectional passing model. Compared with image SOD, VSOD not only focuses on the feature representation learning of a single image, but

also needs to learn the temporal relationship and coherence between consecutive video frames. In this article, we train the backbone based on an effective SOD model.

B. VIDEO SALIENT OBJECT DETECTION

Traditional VSOD models [7], [8], [29], [30], [44] mainly rely on handcrafted features. For example, Huang *et al.* [29] present a fast trajectory-based approach to detect salient regions in videos by motion removal. They exploit long-term object motions to filter out short-term noises and employ one-class SVM to remove consistent trajectories in motion. Chen *et al.* [7] advocates a novel video saliency detection method based on the spatial-temporal saliency fusion and low-rank coherency guided saliency diffusion. In detail, they fuse color saliency with global motion clues in a batch-wise fashion to avoid incorrect low-level saliency map and propose a low-rank coherency guided spatial-temporal saliency diffusion to guarantee the temporal smoothness of saliency maps. Liu *et al.* [8] extract features at the superpixel level and thus propose a superpixel-based spatiotemporal saliency model for saliency detection in videos. Considering background priors are effective clues to find salient objects in images, Xi *et al.* [30] propose a saliency based method to detect the visual objects by using background priors.

With the success of deep learning in static image salient object detection, more and more deep CNNs based methods have made great progress in VSOD. Wang *et al.* [31] firstly introduce deep learning into VSOD and propose a novel static and dynamic saliency information coding scheme. Recently, optical flow guided neural network and ConvLSTM(or ConvGRU) have made great progress in VSOD. More specifically, Li *et al.* [14] introduce a flow guided recurrent neural encoder framework to extend FCN based static-image saliency detector to VSOD, in which an optical flow network is used to estimate the motion of each frame. Song *et al.* [11] propose a pyramid dilated bidirectional ConvLSTM (PDB-ConvLSTM). It mainly includes a pyramid dilated convolution module for simultaneously extracting spatial features at multiple scales, and forward and backward ConvLSTM units to extract multi-scale spatiotemporal information. Li *et al.* [32] introduce a flow guided recurrent neural encoder framework to enhance the temporal coherence modeling of the per-frame feature representation, and exploit a ConvLSTM to capture the evolution of appearance contrast in temporal domain. Although these methods are efficient for the VSOD task, they are time-consuming. To overcome this deficiency, we propose an end-to-end cross complementary network, which costs fewer time, to eliminate the influence of optical flow or ConvLSTM.

III. THE PROPOSED ARCHITECTURE

A. MOTIVATION AND OVERALL ARCHITECTURE

Our goal is to design a network with less computation, which can effectively extract rich spatiotemporal information and low-level and high-level features to generate a group of pixel-wise salient object maps with high quality. Therefore, extracting the spatiotemporal features and ensuring the learning ability of single image have become two key points.

Thus, we elaborate on the details of the proposed video salient object detection model, which consists of two key components. First, single-image representation enhancement (SRE) module (the pink block in Fig. 2) is established for maintaining the single image representation when learning the spatiotemporal information. SRE is composed of Pyramid Pooling Module (PPM) followed by channel attention. Second, we design a spatiotemporal information learning (STIL) module (the blue block in Fig. 2) for spatiotemporal information learning and frame-to-frame relationship (time-dimension) learning. STIL is composed of Spatiotemporal Information Fusion (SIF) and Video Correlation Filter (VCF).

B. SINGLE-IMAGE REPRESENTATION ENHANCEMENT MODULE

The proposed SRE module consists of two blocks: a pyramid pooling module (PPM) and a lightweight channel attention block (CA). PPM is used to learn the multi-scale information and CA is exploited to assign weights more optimally according to the importance of channels.

1) PYRAMID POOLING MODULE

Different from the atrous spatial pyramid pooling (ASPP), it uses different sizes of atrous convolution to learn rich global information. PPM generates feature maps in different levels by pyramid pooling and fuses them to make the network adaptively learn better features. Intuitively, this multi-scale pooling indeed retains global information at different scales. In detail, we first take the adaptive averaging pooling to generate four small-size feature maps, which correspond to 1×1, 2×2, 3×3, 6×6 respectively and use bilinear upsampling to enlarge these blocks to the same size with $layer_5$ (29×29). Then, we concatenate four same feature blocks in the channel dimension and exploit 1×1 convolution to compress these channels to the original number.

2) CHANNEL ATTENTION

The purpose of the designed CA module is to learn the weight of each channel through the attention module and attribute attention weight in the channel domain. The values of different channels are multiplied by different weights, which can enhance the attention in the key channel domain. Specifically, we first use the channel-wise global averaging pooling to squeeze the global information in each feature map:

$$z_c = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} X_c(i,j), \tag{1}$$

where $X_c \in R^{W \times H \times C}$ is the input of CA, W, H and C correspond to width, height and channel dimension (i.e., number of filters), respectively. Then, we take two fully connected layers followed by a Rectified Linear Unit (ReLU) and the sigmoid

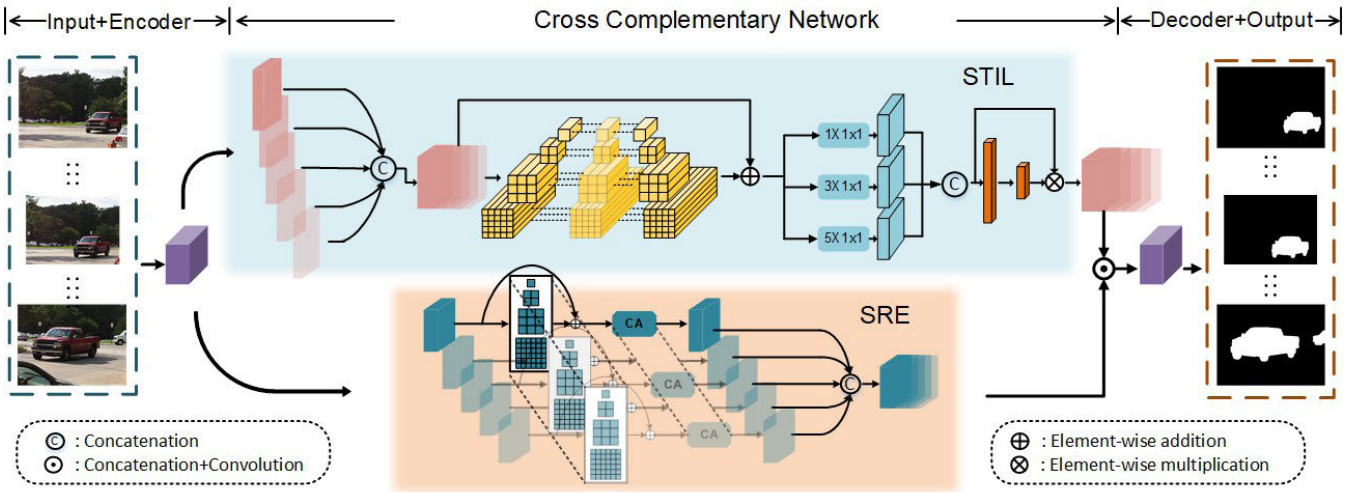


FIGURE 2. The overall architecture of our proposed model. The network is composed by two components: Spatiotemporal Information Learning (STIL) Module and Single-image representation enhancement (SRE) module. A video clip with T frames (here we set T=5) is fed into Encoder to extract features. The STIL extract spatiotemporal information and motion related information based on these features and the SRE is used to enhance the learning ability of single image features.

activation function to enhance the feature representation and increase the nonlinearity of the network.

$$S = \sigma(w_2\,\delta(w_1\,z_c)), \tag{2}$$

where σ is sigmoid activation function and δ is ReLU activation function. w_1, w_2 are parameters of two fully connected layers. The size of the two fully connected layers are set to $C/r \times C$. In order to reduce the model complexity, here we set r=16.

C. SPATIOTEMPORAL INFORMATION LEARNING MODULE

Our proposed spatiotemporal information learning (STIL) module consists of two blocks: spatiotemporal information fusion (SIF) and video correlation filter (VCF). SIF is elaborately designed to learn the spatiotemporal information based on the fusion of T single input blocks, and VCF is used to learn the influence of each individual block corresponding to the input group in the network and redistribute the weights according to their importance.

1) SPATIOTEMPORAL INFORMATION FUSION

Similar to saliency detection, the feature learning of single image in the video saliency task is only applied for representation extraction. Therefore, in order to learn the correlation and spatiotemporal information between consecutive frames more effectively, we first stack the t feature maps corresponding to consecutive t input images in the time dimension, and design a time-dimensional pyramid structure and multi-scale 3D convolution combination module.

Time-dimensional Pyramid Structure. The average pooling block in the pyramid pooling module (PPM) only acts in the w and h dimensions, which enhances the representation for SOD task, but cannot extract spatiotemporal information. Therefore, we propose a novel time-dimensional pyramid structure based on PPM to extract the spatiotemporal information and motion related information. In the time-dimensional pyramid structure, we use 3D adaptive-average pooling to extract features and compress T to n ($n < T$) as time dimension. In this way, the feature information at different times can be merged. Then, we adopt the similar settings to PPM in the two dimensions of w and h to enhance the feature extraction ability of the fused information, and learn more global spatiotemporal correlation information in the time dimension. Suppose n is set to 3, we use the adaptive average pooling to generate four sets of maps with different sizes ($3 \times 512 \times 1 \times 1$, $3 \times 512 \times 2 \times 2$, $3 \times 512 \times 4 \times 4$, $3 \times 512 \times 6 \times 6$) respectively, and use bilinear interpolation to upsample these small-scale feature maps to its original size ($3 \times 512 \times 29 \times 29$). Then, we stack the four branches with the original input on the channel dimension, and take a 3×3 convolution to combine the mainstream and the supplementary features. Different from PPM, our time-dimensional pyramid structure pays more attention on the integrated feature extraction and the spatiotemporal feature enhancement.

Multi-scale 3D Convolution Combination. As shown in the middle block "STIL" in Fig. 2, the STIL takes the output of time-dimensional pyramid structure (five feature maps with 512 channels) as input, and extracts spatiotemporal information in another way. Considering the limited spatiotemporal correlation information obtained from the time-dimensional pyramid structure, and the methods based on LSTM (DB-ConvGRU [12] and PDB-ConvLSTM [11]) are very time-consuming, we propose a multi-scale 3D convolution combination module. However, the amount of calculation for $n \times n \times n$ 3D convolution is very large and the effect of the last two dimensions (h, w) in convolution is not applied to the learning of spatiotemporal information,

since our representation learning is enhanced by the backbone and SRE. Therefore, in order to overcome the above shortcomings, we use $n \times 1 \times 1$ 3D convolution to extract the time dimension information. Specifically, we introduced three 3D convolutions of different receptive fields ($1 \times 1 \times 1$, $3 \times 1 \times 1, 5 \times 1 \times 1$) to extract multi-scale temporal and spatial information, and then stack these three blocks in channel dimension and use a 3D convolution to fuse these features. It is noted that the maximum value of T can only be 8 due to the GPU restriction, so the 3D convolution of $5 \times 1 \times 1$ is sufficient to extract features in the time-dimension. $1 \times 1 \times 1$ 3D convolution does not destroy the original features, so we do not join the residual network.

2) VIDEO CORRELATION FILTER

The channel attention (CA) mechanism mentioned above mainly focuses on redistributing weights according to the importance of channels in each feature map blocks. Inspired by CA, in order to optimize our network in time dimension, we propose a video correlation filter in STIL to learn the importance of T feature map blocks corresponding to T input frames in each group. That is to say, we want to assign weights for each blocks rather than each feature maps in their block.

Considering that the number of feature map blocks (pink cuboid in STIL) in time dimensions is relatively small, we compress the input $X \in R^{(1\times512\times T\times29\times29)}$ only in w and h dimensions and generate $\widetilde{X} \in R^{(1\times512\times T\times1\times1)}$. Specifically, we turn each feature map in two-dimensional ($w \times h$) into a real number, which to some extent has a global receptive field. Then, we compress the five-dimensional vector into a three-dimensional vector $z \in R^{(1\times1\times512*T)}$ to fuse features in the dimensions of time and channel. Similar to softmax, we use two fully connected layers to map the learned distributed feature representation to the sample label space and generate a vector of $Y \in R^{(1\times1\times T\times1\times1)}$. Finally, we multiply the input of video correlation filter with the newly learned filter weight Y and thus generate optimized feature map blocks.

3) BASE NETWORK

Our network is built upon a widely used backbone ResNet-50. For the Encoder, we take five convolution blocks $layer_{1-5}$ to reduce the resolution of feature maps and learn visual representation from low-level to high-level. $layer_1$ takes a sequence of images in video group with resolution 448×448 as input and takes 7×7 kernel size, stride of 2, followed by a batch normalization and a ReLU function to generate five 64-channel feature maps. Different from $layer_1$, the rest four layers just use 3×3 kernel size and add a residual bottleneck architecture (bottleneck). In detail, these four residual layers contain 3, 4, 6, 3 bottlenecks and generate 256, 512, 1024, 2048 channel feature maps, respectively. The strides of these four residual layers are set to 2, 2, 2, 1, and thus the size of the output feature maps is $1/16$ of the original size. Finally, we take a 3×3 convolution to process $layer_5$ and generate 512-channel feature maps, then we feed it into spatiotemporal information learning module, single-image representation enhancement module respectively for the representation enhancement and spatiotemporal information learning.

For the Decoder, through the fusion of low-level and high-level features, $F(\mathrm{F}_{1-5})$ denotes five different previous spatial size of feature maps layer by layer, and the video salient object groups with high-resolution can be predicted with accurate semantic information and object boundary. In order to reduce the impact of detail information loss caused by down-sampling, three refinement blocks are fused for each layer: its corresponding feature map connected from the top-down stream, side-output of feature maps F_i and its previous output feature maps in high-level layer. Bilinear interpolation is applied to the up-sampling small-scale feature maps in high-level layer ($layer_{i+1}$) and F_i, ensuring that they are in the same size with $layer_i$. Note that, channel numbers are reduced to 128, 256, 256, 512, 512 corresponding to $layer_{1-5}$ in the refinement process. Finally, we use a convolution (kernel = 3) to get the final prediction.

IV. EXPERIMENTS AND RESULTS

In this section, we introduce the experimental setup including implementation details, utilized datasets and evaluation metrics, and report the performance of the proposed method. Besides, a number of ablation experiments are performed to analyze the role and importance of each component of the proposed approach.

A. EXPERIMENTAL SETUP

1) IMPLEMENTATION DETAILS

We implement the proposed method based on the publicly available framework: PyTorch-1.0. A PC with a NVIDIA 2080Ti GPU is used for both training and testing, and the operating system is Ubuntu 16.04. We remove STIL and SRE (the blue and pink block in Fig. 2) module from our network and pre-train it as backbone with the training set of DUTS-TR [33] dataset, which has 10553 images. We utilize the Adam optimizer [37] with learning rate $5e-5$ and weight decay 0.0005 to train our pre-train network until it is converged. Pre-training process takes about 10 hours with 24 epoches. After pre-training process, we restore the network to its original appearance and choose the dataset of DUTS [33], DAVIS [34], DAVSOD [36], and FBMS [35] as our training set. Note that we resize input video groups to 448×448 and utilize the Adam optimizer with learning rate $1e-5$ in this process, which takes about 15 hours with 12 epoches.

2) DATASETS

We evaluate the performance of our proposed method on four extensively used video object detection public benchmark datasets: DAVIS [34], FBMS [35], ViSal [5] and DAVSOD [36], all of which are available online. Freiburg-Berkeley motion segmentation (FBMS) is an early adopted dataset which comprises a large, heterogeneous

TABLE 1. Quantitative evaluation in terms of maximum F-measure (maxF), S-measure (S-m) and MAE scores in four popular datasets. ↑ indicates higher scores on the metric are better and ↓ presents lower scores on the metric are better. "-" indicates no reported. "*" indicates traditional methods and others are deep learning based methods. The best results are shown in green and the worst results are shown in red.

	DAVIS2016			FBMS			ViSal			DAVSOD		
	max-F↑	S-m↑	MAE↓	max-F↑	S-m↑	MAE↓	max-F↑	S-m↑	MAE↓	max-F↑	S-m↑	MAE↓
MDB_{15}*	0.470	0.597	0.177	0.487	0.609	0.206	0.692	0.726	0.129	0.342	0.538	0.228
MST_{16}*	0.429	0.583	0.165	0.500	0.613	0.177	0.673	0.749	0.095	0.344	0.532	0.211
$STBP_{17}$*	0.544	0.677	0.096	0.595	0.627	0.152	0.622	0.629	0.163	0.410	0.568	0.160
$SFLR_{17}$*	0.727	0.790	0.056	0.660	0.699	0.117	0.779	0.814	0.062	0.478	0.624	0.132
$SCOM_{18}$	0.783	0.832	0.048	0.797	0.794	0.079	0.831	0.762	0.122	0.464	0.599	0.220
$SCNN_{18}$	0.714	0.783	0.064	0.762	0.794	0.095	0.831	0.847	0.071	0.532	0.674	0.128
$DLVS_{18}$	0.708	0.794	0.061	0.759	0.794	0.091	0.852	0.881	0.048	0.521	0.657	0.129
$FGRN_{18}$	0.783	0.838	0.043	0.767	0.809	0.088	0.848	0.861	0.045	0.573	0.693	0.098
$MBNM_{18}$	0.861	0.887	0.031	0.816	0.857	0.047	0.883	0.898	0.020	0.520	0.637	0.159
$PDBM_{18}$	0.855	0.882	0.028	0.821	0.851	0.064	0.888	0.907	0.032	0.572	0.698	0.116
RCR_{19}	0.848	0.886	0.027	0.859	0.872	0.053	0.906	0.922	0.026	0.653	0.741	0.087
$SSAV_{19}$	0.860	0.892	0.028	0.865	0.879	0.040	0.938	0.943	0.021	0.603	0.724	0.092
$LSTI_{19}$	0.880	0.827	0.031	0.795	0.816	0.084	0.909	0.922	0.027	-	-	-
$PSCA_{20}$	0.880	0.902	0.022	0.837	0.868	0.040	0.940	0.946	0.017	0.655	0.741	0.086
Ours	0.881	0.907	0.023	0.883	0.900	0.035	0.954	0.954	0.013	0.671	0.761	0.084

benchmark with 59 sequences and pixel-accurate ground truth annotation of moving objects. ViSal is the first dataset designed for video object detection task and contains 17 video sequences with obvious objects. DAVIS is a famous dataset deigned for video object segmentation and consists of 50 sequences, 3455 annotated frames. DAVSOD is a newly proposed challenging dataset with 226 video sequences and totally 23938 frames. Note that DAVSOD have not only pixel-accurate ground truth annotation but also eye fixation labels.

3) EVALUATION METRICS

We use five widely used and standard metrics, F-measure, S-measure (S-m), mean absolute error (MAE), precision-recall (PR) curve and F-measure curve to evaluate the performance of the proposed network, and compare it with other state-of-the-art networks. In detail, the F-measure [38] is a harmonic mean weight of Precision and Recall:

$$F_\beta = \frac{\left(1+\beta^2\right) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{3}$$

where *Precision* means the proportion of pixels that are correctly predicted to be positive account for all predictions that are positive in the ground truth and *Recall* means the proportion of pixels that are correctly predicted to be positive account for all actually positive in the ground truth. We set $\beta^2 = 0.3$ to weigh the precision value more important than recall. The higher F-measure, the prediction map is closer to ground truth. The PR curve and F-measure curve are created by varying the saliency threshold from 0 to 255.

The MAE score indicates the similarity in pixel level between the generated saliency map S and the binary ground truth G:

$$MAE = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |S(x,y) - G(x,y)|, \tag{4}$$

where (x, y) is the coordinate position, and W and H represent the width and the height of the predicted saliency map. The smaller the value of MAE, the more similar the saliency map is to the ground truth.

The S-measure is a new way to evaluate non-binary foreground maps proposed by Fan [38], which can simultaneously evaluate based on region-aware structural similarity measure and object-aware structural similarity measure.

B. MODEL ABLATION ANALYSIS

Our proposed method consists of two modules: spatiotemporal information learning (STIL) module and single-image representation enhancement (SRE) module. In this part, we conduct experiments to prove the necessity of the STIL and SRE. Table 2 shows the effect of STIL and SRE in terms of MAE, maxF and S-m. Table 3 shows the effect of the pre-training in training process. Table 4 gives the importance of different blocks in STIL module. From Table 2, we can find that STIL and SRE module are both important, and our

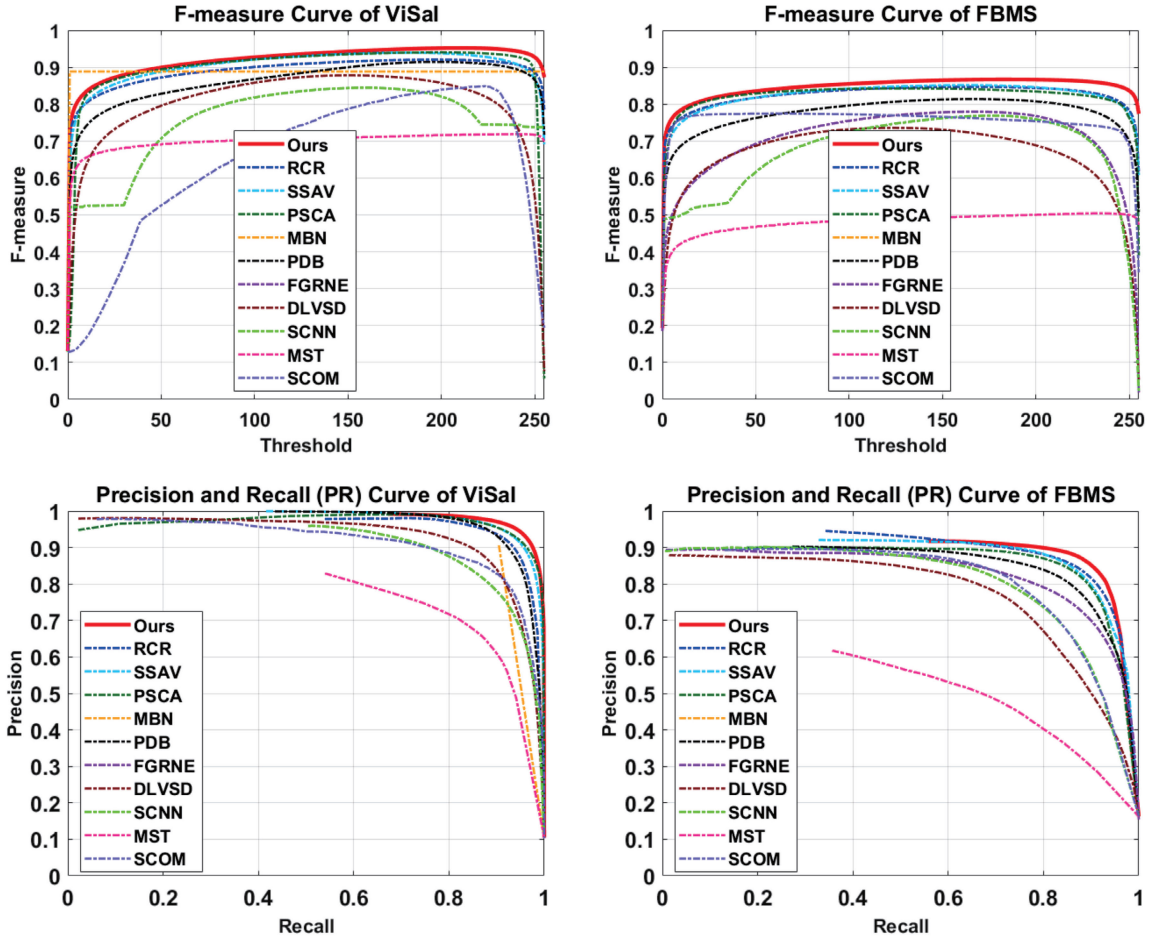


FIGURE 3. Comparisons of the proposed method and other state-of-the-art methods on FBMS and ViSal datasets in terms of PR curves and F-measure curves. Clearly, our method performs well on these two datasets.

TABLE 2. Effect of each module in terms of maximum F-measure, S-measure and MAE on the FBMS datasets. "Backbone" means basic network. "STIL" denotes spatiotemporal information learning module and "SRE" denotes single-image representation enhancement module. The best results are shown in **bold**.

Backbone	STIL	SRE	maxF	S-m	MAE
✓			0.856	0.865	0.052
✓	✓		0.877	0.891	0.037
✓		✓	0.869	0.889	0.040
✓	✓	✓	**0.883**	**0.900**	**0.035**

proposed network with STIL and SRE modules achieves the best performance.

1) EFFECTIVENESS OF PRE-TRAINING

In this experiment, we aims at testing the importance of the pre-training process in training methods. Table 3 shows the result of "w/o Pre-training" and "w/ Pre-training" respectively in terms of maxF, S-measure and MAE in two popular datasets: ViSal and FBMS, where "w/o Pre-training" means without pre-training process in our training

TABLE 3. Maximum F-measure (maxF), S-measure (S-m) and MAE scores on the FBMS (left) and ViSal (right) datasets with respect to "w/o Pre-training" and "w/ Pre-training" in our training process. The best results are shown in **bold**.

	maxF	S-m	MAE	maxF	S-m	MAE
w/o Pre-training	0.687	0.751	0.108	0.821	0.863	0.049
w/ Pre-training	**0.856**	**0.865**	**0.052**	**0.896**	**0.914**	**0.026**

and "w/ Pre-training" denotes adding pre-training process in the training. Obviously, the results without pre-training are much worse than those with pre-training. This proves the importance of learning the representation information of a single image (like the SOD task) through pre-training.

2) EFFECTIVENESS OF DETAILS OF STIL

In Table 2, we learn that STIL is truly effective. To provide more comprehensive analysis of STIL, we further evaluate its performance with respective to spatiotemporal information fusion (SIF) and video correlation filter (VCF) in STIL module. In this experiment, we test the necessity of SIF and VCF in STIL module. Table 4 illustrates the result of "w/o

TABLE 4. Maximum F-measure (maxF), S-measure (S-m) and MAE scores on the FBMS (left) and ViSal (right) datasets with respect to "w/o SIF", "w/o VCF" and "Ours" in spatiotemporal information learning (STIL) module. The best results are shown in **bold**.

	maxF	S-m	MAE	maxF	S-m	MAE
w/o SIF	0.865	0.889	0.039	0.948	0.948	0.016
w/o VCF	0.874	0.894	0.035	0.949	0.948	**0.013**
Ours	**0.883**	**0.900**	**0.035**	**0.954**	**0.954**	**0.013**

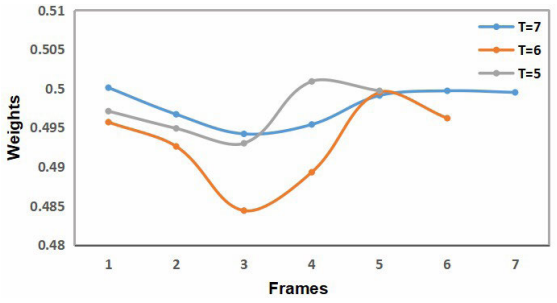


FIGURE 4. Visualization of the weights distribution corresponding to different input video group numbers T.

TABLE 5. F-measure, S-measure and MAE scores on the FBMS dataset with respect to different image numbers T in input block. The best results are shown in **bold**.

	3	4	5	6	7	8
maxF	0.945	0.950	**0.954**	0.948	0.948	0.944
S-m	0.946	0.949	**0.954**	0.947	0.948	0.940
MAE	0.018	**0.013**	**0.013**	0.015	0.015	0.017

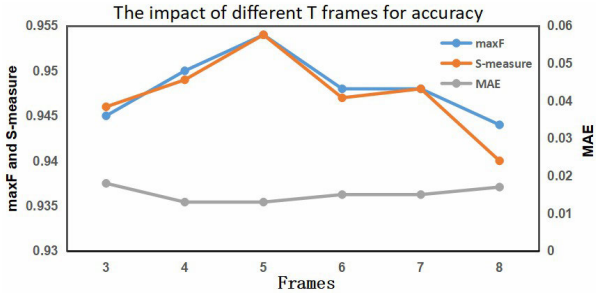


FIGURE 5. Visualization in terms of maxF, S-measure and MAE scores on the ViSal dataset with respect to different image numbers T in input block.

SIF" and "w/o VCF" respectively in terms of Maximum F-measure, S-measure and MAE in ViSal and FBMS datasets in detail, where "w/o SIF" means without spatiotemporal information fusion block in our network and "w/o VCF" means without video correlation filter in our network. "Ours" performs better than "w/o SIF" and "w/o VCF", which demonstrates the effect of SIF and VCF in our network.

Next, through Fig. 4, we visualize the weights generated by the two fully connected layers in VCF. In detail, we list three different situations $T = 5/6/7$ with three evaluation metrics: maxF, S-m and MAE for comparison. Clearly, we find that for T continuous feature maps, the weights trained by the middle feature maps will be lower than feature maps in the front and back which means that our network pay more attention to both sides of video groups.

3) EFFECTIVENESS OF DIFFERENT AMOUNT OF INPUT VIDEO FRAMES

As described in Table 5, we discuss the sensitivity of different input video group numbers to the network's ability of learning spatiotemporal information. We take six number settings: $m_1 = 3$, $m_2 = 4$, $m_3 = 5$, $m_4 = 6$, $m_5 = 7$, $m_6 = 8$, where "m" means the number of input video frames. By repeating the experiment above, we show the performance of CCNet trained with different number of input video frames in Table 5. Note that our maximum T can only be set to 8 due to the GPU limitations. According to Fig. 5 and Table 5, it can be clearly seen that $m = 5$ obtains the best performance. Thus, we keep $m = 5$ in all experiments.

C. COMPARISON WITH THE STATE-OF-THE-ART METHODS

1) QUANTITATIVE COMPARISON

We compare our video saliency detection network with other 14 state-of-the-art models, including MDB [39], MST [44], STBP [30], SFLR [7], SCOM [43], SCNN [45], DLVS [31], FGRN [14], MBNM [46], PDBM [11], RCR-Net [12], SSAV [36], PSCA [47], and LSTI [48]. For fair comparison, we take the code provided by Fan *et al.* [36] to compute these metrics on our video saliency maps. As seen from Table 1, our method gets the best result on four test datasets in terms of maximum F-measure (maxF), S-measure and MAE, which demonstrates the effectiveness of our proposed method. As shown in Fig. 3, the PR curves and F-measure curves of our method are higher than other methods, which demonstrates our method is more robust than other approaches, even on the challenging datasets.

From Table 1 above, we learn that the index of maxF, S-measure and MAE get the best result on the four datasets (DAVIS, FBMS, ViSal and DAVSOD). To be specific, our method gets much improvement, comparing with the best existing approach on ViSal dataset and FBMS dataset, 0.954, 0.883 in the index of maxF, 0.954, 0.900 in the index of S-measure and 0.013, 0.035 in the index of MAE, respectively. Since DAVSOD is a newly proposed dataset, our method demonstrates 2.4% maxF and 2.2% S-measure higher than the second best model PSCA. Moreover, quantitative evaluation results of our proposed method in terms of index PR curve, F-measure curve are listed in Fig. 3. Note that due to the lack of video groups on the DAVIS2016 and DAVSOD datasets, we only provide comparisons on the FBMS and ViSal databases.

2) VISUAL COMPARISON

Fig. 6 shows visual comparisons of our proposed model with 11 previous state-of-the-art methods. From the results given above, we find that the video saliency maps generated by our method are more accurate and more similar with the ground truth. Specifically, our method is more accurate in the recognition of continuous salient objects in video. For the

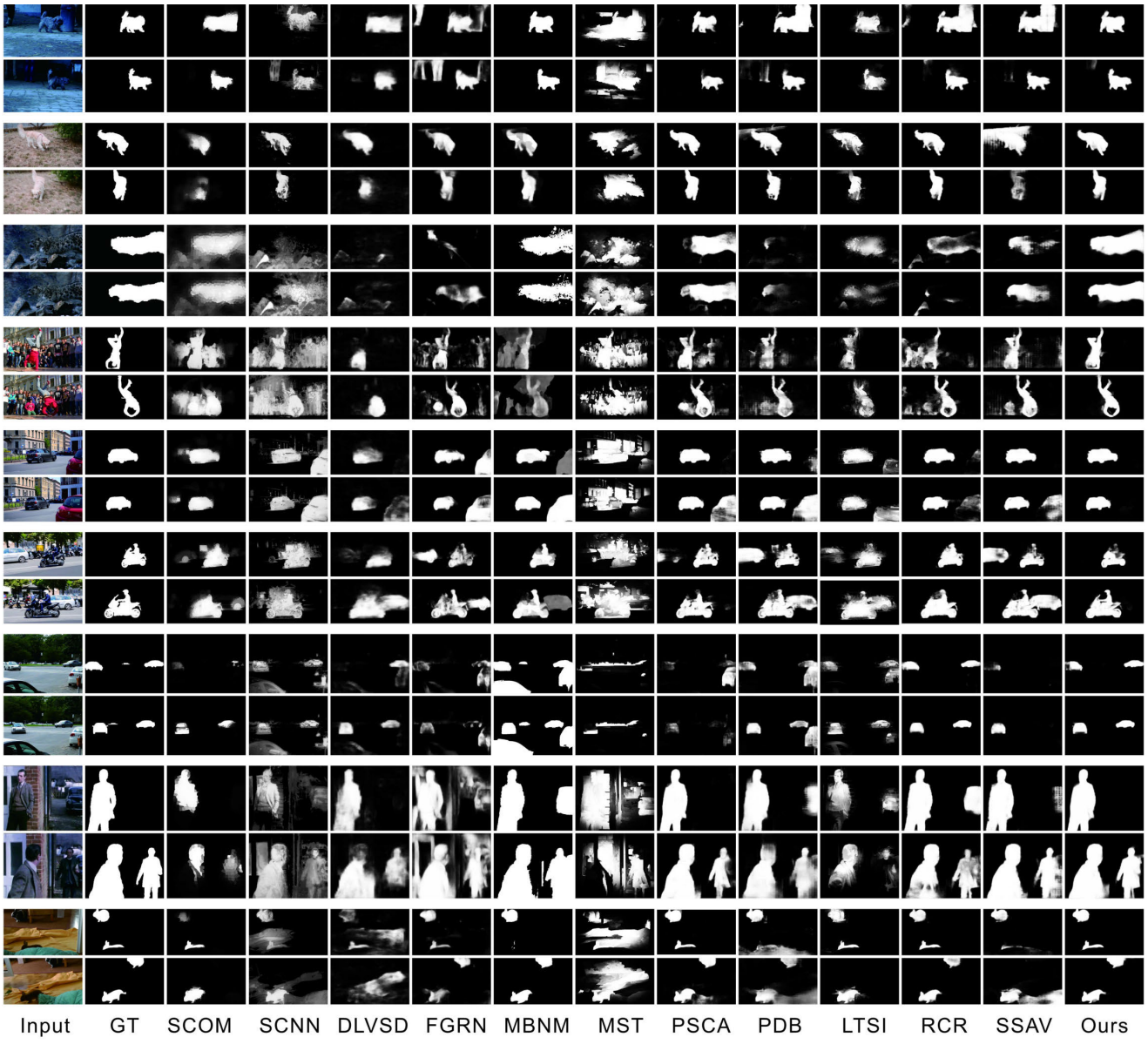


FIGURE 6. Visual comparisons of the proposed method and the state-of-the-art algorithms. From left to right: the input image, ground truth, the saliency maps produced by our proposed method, GT, SCOM [43], SCNN [45], DLVSD [31], FGRN [14], MBNM [46], MST [44], PSCA [47], PDB [11],LSTI [48], RCR [12], SSAV [36]. Our method consistently produces saliency maps closest to the ground truth. "GT" indicates Ground Truth. "Input" indicates input image.

most of the challenging video groups, they mainly are with low contrast between objects and background (*e.g.*, row 1, 2, 3 and 9 in Fig. 6), multi-objects or multi-salient objects overlap (*e.g.*, row 7, 8 and 9 in Fig. 6). Our proposed method can effectively divide the foreground and the background (*e.g.*, row 1, 3 and 4 in Fig. 6) and distinguish continuous salient objects (*e.g.*, row 5, 6 and 8 in Fig. 6).

Besides, our model can highlight both small-scale or large-scale salient objects in the video groups, no matter having a large difference (*e.g.*, row 2, 4, 6 and 8 in Fig. 6) or small difference (*e.g.*, row 3 and 7 in Fig. 6) in the movement process. These results demonstrate the robustness of our method, and confirm the effectiveness of the proposed cross complementary network in obtaining the discriminant feature representations and spatiotemporal information for video salient object detection.

TABLE 6. Average running time cost comparison (in a single video frame) for optical flow based methods."-" indicates no reported.

Methods	EC	SP	SA	SCOM	bMRF	Ours
Total time cost	8.1	11.8	2.56	38.8	2.6	**0.04**
Optical flow cost	-	10.1	-	36.7	1	-

3) RUNTIME COMPARISON

We compare our video saliency detection network with optical flow based methods and all methods in runtime

TABLE 7. Average running time cost (in second) for a single video frame. The first row corresponds to various state-of-the-art image/video saliency methods. The second row corresponds to average time cost for one video frame in different methods.

Method	MDB	MST	EC	SA	SP	STBP	SCOM	SCNN	bMRF	MBNM	FGRN	PDBM	RCR	SSAV	Ours
Time cost	0.02	0.02	8.1	2.56	11.8	49.49	38.8	38.5	2.6	2.63	0.09	0.05	0.04	0.05	0.04

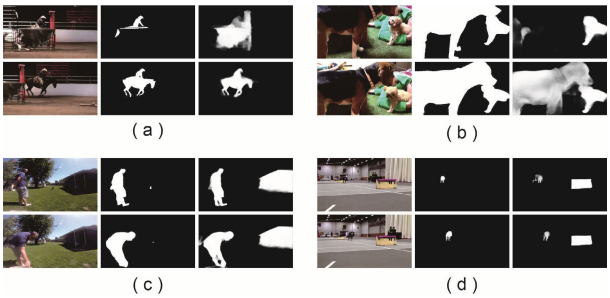


FIGURE 7. Four failure cases of the proposed method. The three lines in each case from left to right correspond to: input image, ground truth and our final predication result.

cost respectively in Table 6 and Table 7. In Table 6, we show total time cost and optical time cost in five optical flow based methods: EC [40], SP [8], SA [42], SCOM [43] and bMRF [41]. For most optical flow based methods, they account for a large proportion in total time cost.

Table 7 shows runtime comparisons of our proposed model with 14 previous state-of-the-art methods (some are not mentioned), including MDB [39], MST [44], EC [40], SA [42], SP [8], STBP [30], SCOM [43], SCNN [45], FGRN [14], bMRF [41], MBNM [46], PDBM [11], RCRNet [12] and SSAV [36]. Though some traditional methods (*e.g.*, MDB, MST in Table 7) are truly efficient in time cost, they perform worse than deep learning based methods in F-measure, S-measure and MAE. It can be seen from Table 7 that the average time of a single image of our method takes 0.04s, which is close to the real-time level. It is far better than many deep learning based methods (*e.g.*, FGRN, PDB, SSAV in Table 7) and optical flow based methods (*e.g.*, SA, SP, SCOM, SCNN and MBNM in Table 7). From the results given above, single image average time of our method takes 0.04s which is close to real-time level, which is much better than many deep learning based methods (*e.g.*, FGRN, PDB, SSAV in Table 7) and optical flow based methods (*e.g.*, SA, SP, SCOM, SCNN, MBNM and FGRN in Table 7).

V. LIMITATIONS

Although satisfactory results have been achieved, there are still some limitations in our cross complementary network. Fig. 7 shows some failure examples. First, our proposed method is difficult to deal with some challenging video groups, in which the salient objects are occluded or incomplete (*e.g.*, (*a*) and (*b*) in Fig. 7). Second, some static salient objects that may be salient in a single image are also considered as salient objects in video salient object detection (*e.g.*, (*c*) and (*d*) in Fig. 7).

TABLE 8. Quality comparison with 8 unsupervised video object segmentation (UVOS) methods in DAVIS16 using the region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$. The best scores are marked in bold.

Dataset	Metrics	FST	SFL	LMP	FSEG	LVO	PDB	RCR	AGS	Ours
DAVIS	$\mathcal{J}\uparrow$	55.8	67.4	70.0	70.7	75.9	77.2	74.7	79.7	**80.8**
	$\mathcal{F}\uparrow$	51.1	66.7	65.9	65.3	72.1	74.5	73.3	77.4	**79.5**
FBMS	$\mathcal{J}\uparrow$	47.7	35.7	35.7	68.4	65.1	72.3	75.9	-	**80.1**

VI. PERFORMANCE ON UNSUPERVISED VIDEO OBJECT SEGMENTATION

We compare our method on the DAVIS and FBMS dataset with 8 state-of-the-art methods: FST [49], SFL [50], LMP [51], FSEG [52], LVO [53], PDB [11], RCR [12] and AGS [54]. Following the evaluation setting of unsupervised video object segmentation (UVOS), we adopt the mean Jaccard index $\mathcal{J}$ (intersection-over-union) and mean boundary accuracy $\mathcal{F}$ as metrics for fair comparison. In Table 8, our method achieve the best accuracy in UVOS task.

VII. CONCLUSION AND FUTURE WORK

In this article, we concern video salient object detection both from emphasizing spatiotemporal information and better fusion in spatiotemporal information and saliency information in a single image. We present an end-to-end cross complementary network, which consists of a spatiotemporal information learning (STIL) module for spatiotemporal information extraction and a single-image representation enhancement (SRE) module for feature supplement for single image. Besides, in order to maintain the ability of extracting image features, we combine video dataset and image dataset as our training set. Experimental evaluation on four datasets demonstrates that our proposed approach provides more accurate video saliency maps as compared to the state-of-the-art video saliency detection methods. In further work, we will focus on replacing our backbone with a lightweight network to increase the detection speed and while maintaining the accuracy.

REFERENCES

[1] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[2] Y. Liang, Y. Zhang, Y. Wu, S. Tu, and C. Liu, "Robust video object segmentation via propagating seams and matching superpixels," *IEEE Access*, vol. 8, pp. 53766–53776, 2020.

[3] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[4] H. Wu, G. Li, and X. Luo, "Weighted attentional blocks for probabilistic object tracking," *Vis. Comput.*, vol. 30, no. 2, pp. 229–243, Feb. 2014.

[5] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[6] W. Wang, J. Shen, F. Guo, M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4894–4903.

[7] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.

[8] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[9] L. Jiang, M. Xu, and Z. Wang, "Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM," 2017, *arXiv:1709.06316*. [Online]. Available: http://arxiv.org/abs/1709.06316

[10] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[11] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 715–731.

[12] P. Yan, G. Li, Y. Xie, Z. Li, C. Wang, T. Chen, and L. Lin, "Semi-supervised video salient object detection using pseudo-labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7284–7293.

[13] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[14] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3243–3252.

[15] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 478–487.

[16] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.

[17] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[18] L. Han, X. Li, and Y. Dong, "Convolutional edge constraint-based U-Net for salient object detection," *IEEE Access*, vol. 7, pp. 48890–48900, 2019.

[19] J. Li, Z. Wang, and Z. Pan, "Double structured nuclear norm-based matrix decomposition for saliency detection," *IEEE Access*, vol. 8, pp. 159816–159827, 2020.

[20] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[21] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

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

[23] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P. Heng, "R3Net: Recurrent residual refinement network for saliency detection," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[24] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[25] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.

[26] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[27] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[28] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1741–1750.

[29] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1336–1349, Aug. 2014.

[30] T. Xi, W. Zhao, H. Wang, and W. Lin, "Salient object detection with spatiotemporal background priors for video," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3425–3436, Jul. 2017.

[31] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[32] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7274–7283.

[33] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

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

[35] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 282–295.

[36] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8554–8564.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[38] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[39] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 FPS," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1404–1412.

[40] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.

[41] C. Chen, S. Li, H. Qin, Z. Pan, and G. Yang, "Bilevel feature learning for video saliency detection," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3324–3336, Dec. 2018.

[42] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3395–3402.

[43] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis, "SCOM: Spatiotemporal constrained optimization for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3345–3357, Jul. 2018.

[44] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2334–2342.

[45] Y. Tang, W. Zou, Z. Jin, Y. Chen, Y. Hua, and X. Li, "Weakly supervised salient object detection with spatiotemporal cascade neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1973–1984, Jul. 2019.

[46] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C. Kuo, "Unsupervised video object segmentation with motion-based bilateral networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 207–223.

[47] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S.-P. Lu, "Pyramid constrained self-attention network for fast video salient object detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 10869–10876, Apr. 2020.

[48] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, "Improved robust video saliency detection based on long-term spatial-temporal information," *IEEE Trans. Image Process.*, vol. 29, pp. 1090–1100, 2020.

[49] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1777–1784.

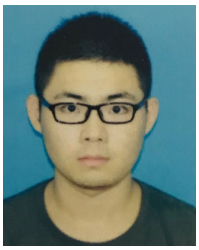
[50] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3271–3279.

[51] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3386–3394.

[52] S. D. Jain, B. Xiong, and K. Grauman, "FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2126.
[53] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4481–4490.
[54] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. H. Hoi, and H. Ling, "Learning unsupervised video object segmentation through visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3064–3074.

ZIYANG WANG is currently pursuing the M.S. degree with the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China. His current research interests include image processing and computer vision.

JUNXIA LI (Member, IEEE) received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2017. From 2014 to 2015, she was a Visiting Student with Nanyang Technological University, Singapore. She is currently a Lecturer with the School of Automation, Nanjing University of Information Science and Technology, Nanjing. Her research interests include visual saliency detection, image segmentation, and computer vision.



ZEFENG PAN is currently pursuing the M.S. degree with the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China. His current research interests include static image saliency detection and video saliency detection.