Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.045258

ARTICLE

Multi-Stream Temporally Enhanced Network for Video Salient Object Detection

Dan Xu*, Jiale Ru and Jinlong Shi

School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, 212100, China

*Corresponding Author: Dan Xu. Email: xudan_zj@163.com

Received: 21 August 2023 Accepted: 08 November 2023 Published: 30 January 2024

ABSTRACT

Video salient object detection (VSOD) aims at locating the most attractive objects in a video by exploring the spatial and temporal features. VSOD poses a challenging task in computer vision, as it involves processing complex spatial data that is also influenced by temporal dynamics. Despite the progress made in existing VSOD models, they still struggle in scenes of great background diversity within and between frames. Additionally, they encounter difficulties related to accumulated noise and high time consumption during the extraction of temporal features over a long-term duration. We propose a multi-stream temporal enhanced network (MSTENet) to address these problems. It investigates saliency cues collaboration in the spatial domain with a multi-stream structure to deal with the great background diversity challenge. A straightforward, yet efficient approach for temporal feature extraction is developed to avoid the accumulative noises and reduce time consumption. The distinction between MSTENet and other VSOD methods stems from its incorporation of both foreground supervision and background supervision, facilitating enhanced extraction of collaborative saliency cues. Another notable differentiation is the innovative integration of spatial and temporal features, wherein the temporal module is integrated into the multi-stream structure, enabling comprehensive spatial-temporal interactions within an end-to-end framework. Extensive experimental results demonstrate that the proposed method achieves state-of-the-art performance on five benchmark datasets while maintaining a real-time speed of 27 fps (Titan XP). Our code and models are available at https://github.com/RuJiaLe/MSTENet.

KEYWORDS

Video salient object detection; deep learning; temporally enhanced; foreground-background collaboration

1 Introduction

The human vision system always pays more attention to objects that are more prominent, distinctive, or in motion. Video Salient Object Detection (VSOD) aims at identifying visually distinctive regions in a video and has been a hot topic in the field of computer vision research due to its applicability to downstream video tasks, such as video object segmentation [1], visual tracking [2,3], video compression [4], and video popularity prediction [5]. High-speed VSOD exhibits a broad range of applications, especially in the domains of autonomous vehicles [6] and traffic management [7].

There has been considerable development in still image-based salient object detection in the last few decades [8–11]. Salient object detection in videos is considered to be more challenging as compared to its image counterpart, as it is not only influenced by the complex spatial scenes but also affected by the temporal dynamics. To address this issue, several temporal modeling strategies have been proposed to guide the salient object detection for individual frames, including 3D convolution-based methods [12,13], ConvLSTM-based methods [14,15], and optical flow-based methods [16,17]. For a better understanding, Fig. 1 demonstrates currently the most representative temporal models and architectures used in VSOD approaches.

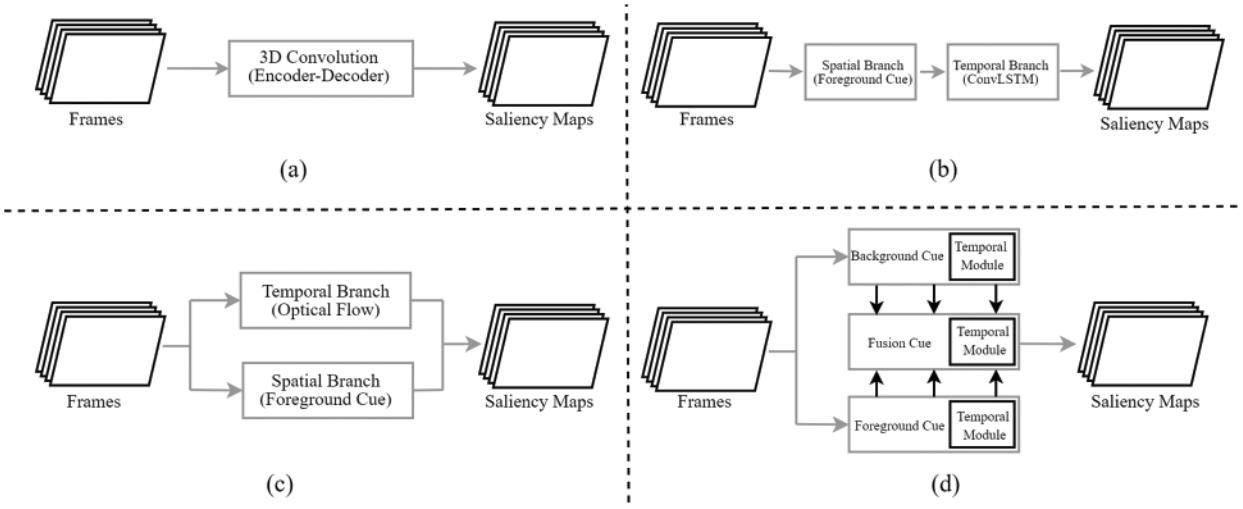

Figure 1: Temporal model and architecture comparison of the proposed model and other SOTA methods. (a) 3D convolution, (b) ConvLSTM, (c) optical flow, and (d) the proposed model

Although the existing video salient object detection methods achieve great improvement in spatiotemporal feature extraction and salient object detection, they suffer from two main problems. The first problem is the existing methods pay more attention to foreground feature extraction in the spatial domain, which cannot deal with the great diverse background scenes. The second problem is accumulative noise and expensive computation costs introduced by the temporal model of the methods. For instance, optical flow-based models suffer from expensive computation costs and a heavy dependence on the quality of optical flow maps. ConvLSTM-based methods take long-term temporal information into account, which may introduce accumulative noise from previous frames to the current frame and result in inaccurate salient regions.

In this paper, we introduce a novel approach, namely the Multi-Stream Temporally Enhanced Network (MSTENet), for video salient object detection. Our proposed multi-stream architecture is designed to facilitate the extraction of foreground, background, and foreground-background collaboration saliency cues in the spatial domain. This approach helps overcome the challenge of diverse background scenes and achieving precise salient regions. The temporally enhanced module aims at extracting motion information by enhancing distinct regions in the current frame in contrast to those in adjacent frames. The accumulative noise is avoided in our model as we consider the difference between adjacent frames instead of propagating the spatiotemporal information of previous frames into the current frame. In addition, as shown in Fig. 1, we develop an end-to-end framework that incorporates the temporally enhanced module into the multi-stream structure, enabling full spatial-temporal interactions.

Concretely, the main contributions of this work are listed as follows:

(1) We present a multi-stream fusion structure that encompasses foreground supervision, background supervision, and foreground-background fusion stream. The proposed structure effectively extracts collaborative saliency cues in the spatial domain and helps locate accurate salient regions.
(2) A simple and efficient difference enhancement module (DEM) is designed to acquire motion information between adjacent frames. The lightweight temporal enhancement module has the potential to avoid accumulative noise and high computation costs.
(3) Extensive experiments on five widely utilized datasets demonstrate that our method outperforms 18 state-of-the-art (SOTA) VSOD approaches in terms of three evaluation metrics. Furthermore, our method ensures a real-time speed of 27 fps on a single GPU.

The rest of the paper is arranged as follows: In Section 2, we briefly introduce the related work, including the progress of image salient object detection and video salient object detection. In Section 3, we describe our method in detail. In Section 4, we explain the experimental setting and comparison with existing SOTA VSOD methods. Conclusions are drawn in Section 5.

2 Related Work

2.1 Salient Object Detection

At the early stage of the saliency detection study, heuristic SOD approaches, which employ low-level saliency priors and handcrafted features, were developed to highlight the attentive object regions [18–20]. When entering the deep learning era, substantial improvement for SOD has been achieved, benefiting mostly from numerous labeled data and powerful learning methods. The convolutional neural network (CNN) [21,22] and fully convolutional networks (FCNs) [23,24] have been widely applied for SOD since its first introduction in 2015. Then, several works [25,26] demonstrate that shallow layers are more important in capturing detailed salient features; therefore, multi-level aggregation methods are proposed to improve the saliency detection performance. For example, Hou et al. [8] used a short connection to combine deeper layers and shallower layers to generate satisfactory saliency maps. CNN-based encoder-decoder models have profoundly inspired recent works in SOD [9]. Ren et al. [10] introduced a deep encoder-decoder network that captures discriminative saliency cues and generates confidence scores. In this paper, we adopt the encoder-decoder architecture and incorporate short connections due to their notable success in the task of SOD. Nevertheless, the SOTA works only focus on saliency cues in the foreground objects and pay little attention to the features included in the background regions. This leads to poor detection performance in complex background scenes. To this end, we propose a multi-stream framework, including foreground stream, background stream, and fusion stream. By leveraging foreground and background supervision, our proposed framework aims to extract salient regions more accurately and effectively.

2.2 Video Salient Object Detection

To extract the spatiotemporal features entangled in the video, the current VSOD models are devoted to building parallel networks, in which one network captures the static saliency over the spatial domain and the other network extracts the motion saliency over the temporal scale. Xu et al. [27] proposed a dual-stream spatiotemporal attention network that consists of two paths: the context feature path for modeling the long-range temporal features over frames, and a content feature path to effectively model the local spatial content information. Bak et al. [28] proposed a two-stream

network to investigate different fusion mechanisms for spatial and temporal information for saliency estimation. The static saliency detection methods that are conducted in the spatial domain have been fully discussed in Section 2.1. Now we pay more attention to motion modeling strategies in the temporal domain.

Early studies on VSOD relied on 3D convolution to capture the interframe temporal information. Tran et al. [29] adopted a 3D convolutional deep network with a convolution kernel to simultaneously model appearance and motion information. Le et al. [12] proposed a network that uses 3D filters in the spatiotemporal domain to directly learn both spatial information and temporal information to extract 3D deep features and then transfer the 3D deep features to pixel-level saliency prediction. To explore motion information from a longer period, ConvLSTM is used for temporal information capture. Fan et al. [14] proposed a baseline model equipped with a saliency-shift-aware ConvLSTM and a densely annotated VSOD dataset. Song et al. [15] proposed a fast VSOD model with a pyramid dilated convolution module and a deeper bidirectional ConvLSTM module for video salient object detection. In addition, to explicitly model motion cues, optical flow-based methods were proposed. Li et al. [16] developed a network for VSOD with an appearance branch for salient object detection in still images and a motion branch for motion saliency detection in optical flow images. Li et al. [30] designed an optical flow-guided current encoder for estimating the motion of each frame to enhance temporal coherence. To develop deeper insight into the dynamic nature of video data, Zhang et al. [31] proposed a dynamic context-sensitive filtering module and bidirectional dynamic fusion strategy for video salient object detection.

Despite the remarkable improvements, the aforementioned temporal modeling strategies are hampered by the issues of accumulating noise and high computational costs. In this study, we present an innovative lightweight temporal model along with a spatial-temporal interaction framework to effectively address these concerns.

3 The Proposed Method

3.1 Architecture Overview

The overall network structure of the proposed method is shown in Fig. 2. The MSTENet takes a video clip consisting of four consecutive frames I_t $(t = 1, 2, 3, 4)$ as input, generating saliency prediction S_t $(t = 1, 2, 3, 4)$. The network mainly contains three streams from top to bottom: background stream, fusion stream, and foreground stream. The background and foreground streams are supervised by the background mask and ground truth, respectively. A novel foreground-background fusion module (FBFM) is developed to gradually fuse the background and foreground saliency cues. The saliency prediction map generated by the fusion stream is considered the final saliency map of the MSTENet. To fully utilize the temporal information between consecutive frames, we design a DEM to capture temporal features by enhancing different regions between adjacent frames. The DEM module is embedded into three spatial streams in the decoder stage, enabling the losses from the decoder (temporal) to be propagated to the encoder (spatial) during backward propagation and facilitating the integration of temporal and spatial features. Instead of adding the temporal module after each block, we only put the DEM in the last two decoder blocks as they contain rich, detailed spatial structure information.

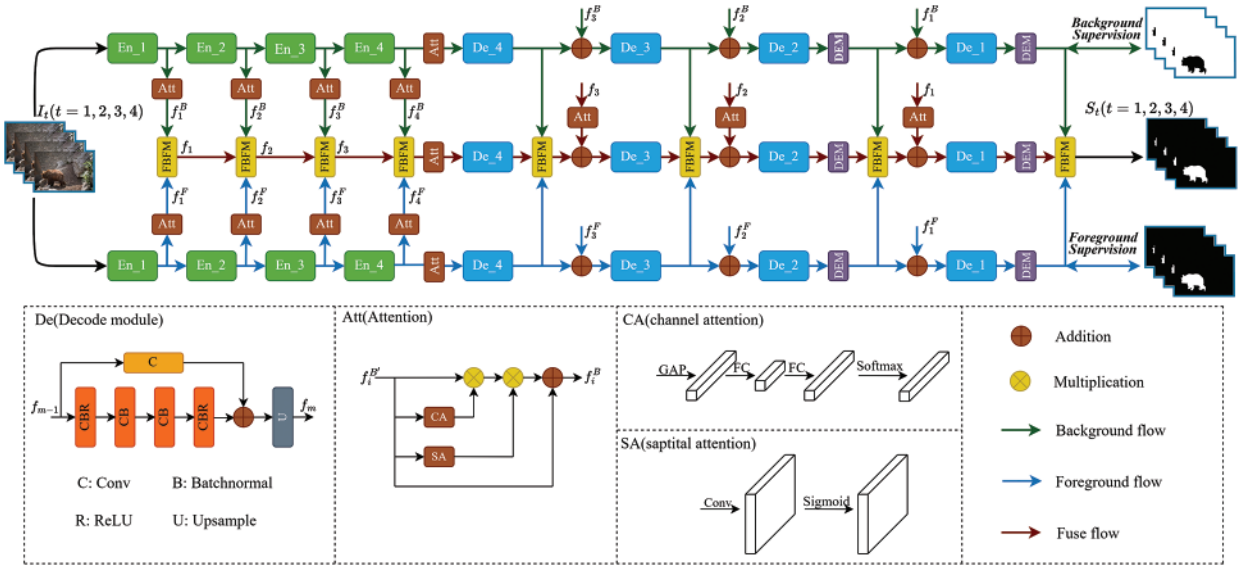

Figure 2: Overall architecture of our proposed method. The upper part of the figure shows the framework of the MSTENet, and the lower part of the figure further explains the components. The MSTENet contains three streams from top to bottom: Background stream, fusion stream and foreground stream. The foreground and background streams are supervised by the ground truth and background mask, respectively. Both foreground and background saliency cues are gradually integrated into the fusion stream which is supervised using the ground truth. The DEM module is embedded into three spatial streams in the decoder stage, facilitating the integration of temporal and spatial features. The MSTENet takes a video clip consisting of four consecutive frames I_t (t = 1, 2, 3, 4) as input. The saliency maps S_t (t = 1, 2, 3, 4), generated by the fusion stream, are regarded as the final output of the MSTENet

3.2 Multi-Stream Fusion Structure

3.2.1 Foreground and Background Streams

Both the background and foreground streams follow the encoder-decoder structure. The Shunted-Transformer-s [32] network is adopted as a backbone of the encoder to acquire a more powerful feature capture ability. The encoder consists of four stages, each generating a distinct feature map denoted as f_m $(m = 1, 2, 3, 4)$, characterized by different spatial resolutions and channel numbers. After feature extraction, the decoder is arranged into four stages. To make full use of the spatial saliency cues provided by the encoder, each decoder layer is connected to its corresponding encoder layer with a skip connection [11]. Furthermore, a channel-spatial attention module [33] is added to both the FBFM module and skip connection stage to pay more attention to discriminative channels (channel attention) and locations (spatial attention) of feature maps. For the foreground stream, the loss is calculated by comparing the ground truth with the predicted saliency map. Conversely, for the background stream, the loss is computed by comparing the background mask with the predicted background map.

3.2.2 Foreground-Background Fusion Module

Instead of employing a heuristic fusion approach, we propose a fusion stream to leverage the complementary saliency cues derived from both the background and foreground streams. The fusion

stream adopts an encoder-decoder structure, wherein the background and foreground saliency cues are progressively integrated with a novel FBFM module. The resulting saliency map obtained from the fusion stream represents the final output of the MSTENet model.

The detailed structure of the FBFM is shown in Fig. 3. The FBFM module takes f_m^B, f_m^F and f_{m-1} as input, which denotes the m-th level background feature, foreground feature, and (m−1)-th level fusion information, respectively. Then, we add them to obtain the immediate fusion information $f_m^{'}$ through a conv of 3 × 3 and a batchnormal layer:

$$f_m^{'} = Bconv_3\left(f_m^B + f_m^F + f_{m-1}\right) \tag{1}$$

where $Bconv_3(\cdot)$ is a sequential operation that consists of 3 × 3 convolution and batch normalization.

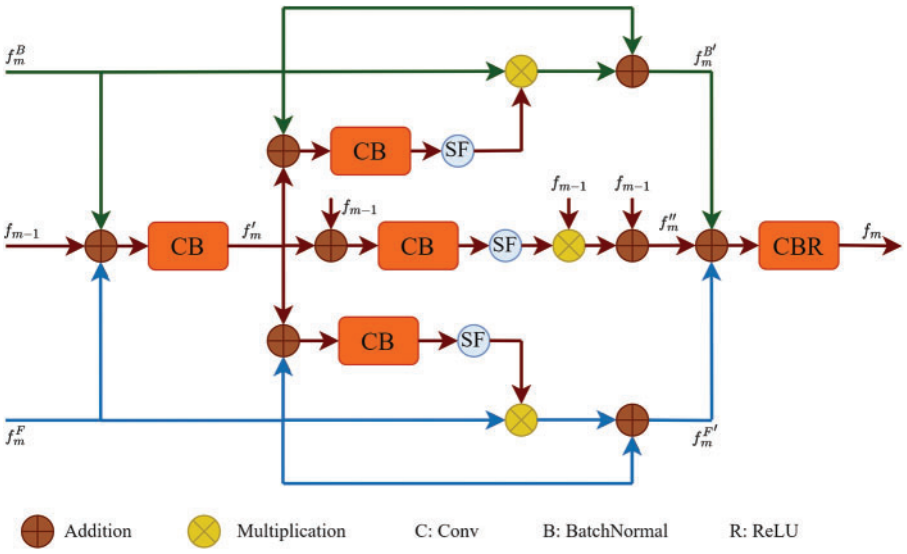

Figure 3: The detailed structure of the FBFM. 'SF' denotes the softmax function

The background feature f_m^B is added with $f_m^{'}$, followed by a CB (conv and batchnormal) layer and a softmax function to acquire a fusion weight of f_m^B. The element-wise multiplication operation is selected to enhance the specific feature in f_m^B. Inspired by [34], to preserve the original information, a short connection is adopted to combine the weighted feature with the original f_m^B. The foreground feature f_m^F and last layer fusion feature f_{m-1} follow the same fusion process, which is expressed as:

$$f_m^{B'} = f_m^B + f_m^B \otimes Softmax\left(Bconv_3\left(f_m^B + f_m^{'}\right)\right) \tag{2}$$

$$f_m^{''} = f_{m-1} + f_{m-1} \otimes Softmax\left(Bconv_3\left(f_{m-1} + f_m^{'}\right)\right) \tag{3}$$

$$f_m^{F'} = f_m^F + f_m^F \otimes Softmax\left(Bconv_3\left(f_m^F + f_m^{'}\right)\right) \tag{4}$$

where $\otimes$ denotes element-wise multiplication, $f_m^{B'}$, $f_m^{F'}$ and $f_m^{''}$ represent enhanced background, foreground, and fusion features in the m-th level, respectively. We add $f_m^{B'}$, $f_m^{F'}$ and $f_m^{''}$, then feed them to a CBR (conv, batchnormal, and ReLU) layer to obtain the final fusion feature f_m.

$$f_m = RBconv_3\left(f_m^{B'} + f_m^{F'} + f_m^{''}\right) \tag{5}$$

where $RBconv_3(\cdot)$ is a sequential operation that combines a 3 × 3 convolution followed by batch normalization and a ReLU function.

Notably, our FBFM module hierarchically fuses background and foreground saliency cues, which provides rich complementary information to promote saliency detection performance.

3.3 Difference Enhancement Module

In general, adjacent frames within a video exhibit typically subtle disparity. In this study, we introduce a Differential Enhancement Module (DEM) aiming to amplify the nuanced variances observed among consecutive frames. The DEM serves the purpose of extracting and highlighting temporal features from the enhanced differences. Note that we only add the DEM to the last two decoder blocks as abundant fine information is contained in these stages.

The details of the DEM are presented in Fig. 4. A bidirectional differential structure is adopted to further enhance the interframe differences. During forward and backward propagation, the current frame is compared with its last frame and next frame, respectively. In this way, we can more effectively capture temporal information than a single-directional structure. The input of the DEM is a feature map sequence $F = f_t, t = 1, 2, 3, 4$, which is generated by the decode module. A difference enhancement operation $DE(\cdot)$ is used to obtain the difference matrix between adjacent frames and to enhance the distinctive regions of the current frame compared with the last frame, followed by a short connection and a CBR layer. We denote the output of forward propagation as $F^{fw} = \left(f_t^{fw},\ t = 1, 2, 3, 4\right)$. Backward propagation is calculated with similar operations, and the corresponding output is $F^{bw} = \left(f_t^{bw},\ t = 1, 2, 3, 4\right)$. The bidirectional propagation of the DEM is formulated as:

$$f_t^{fw} = RBconv_3\left(DE\left(f_t,\ f_{t-1}\right) \oplus f_t\right) \tag{6}$$

$$f_t^{bw} = RBconv_3\left(DE\left(f_t^{fw}, f_{t+1}^{fw}\right) \oplus f_t^{fw}\right) \tag{7}$$

$$DE\left(f_i, f_j\right) = f_i \otimes \left(1 \oplus Softmax\left(Abs\left(f_i - f_j\right)\right)\right) \oplus f_i \tag{8}$$

where $RBconv_3(\cdot)$ is a sequential operation that combines a 3×3 convolution followed by batch normalization and a ReLU function. $\otimes$, $\oplus$ and Abs denote element-wise multiplication, addition, and absolute value, respectively.

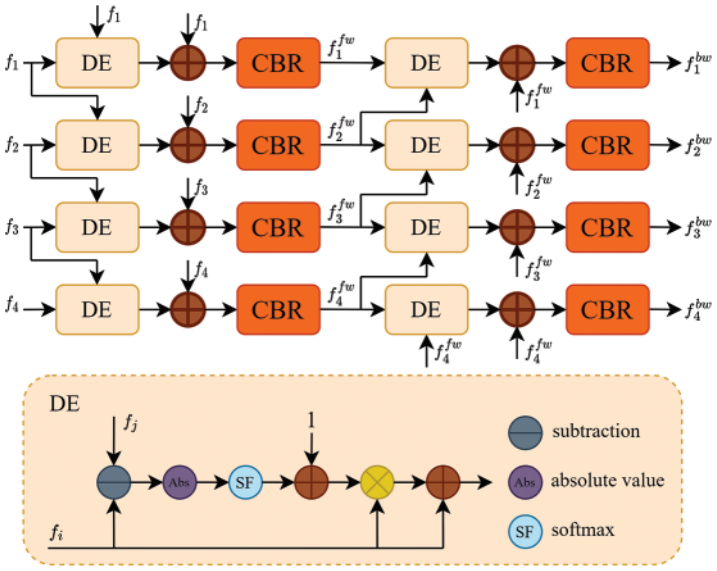

Figure 4: Detailed structure of the proposed difference enhancement module (DEM)

Note that for t = 1 in forward propagation and t = 4 in backward propagation, $DE(f_1, f_1)$ and $DE\left(f_4^{fw}, f_4^{fw}\right)$ are computed as there are no extra frames for the interframe differential.

The calculation flow $DE(\cdot)$ is shown at the bottom of Fig. 4. $Abs\left(f_i - f_j\right)$ is used to generate a difference matrix between frames f_i and f_j. The operation $1 \oplus Softmax(\cdot)$ maps the value of the difference matrix to the range [1,2], which is applied as the weight of enhancing distinctive regions in the frame f_i.

3.4 Loss Function

We use the hybrid loss proposed by BASNet [11] as our final loss function L:

$$L = l_{bce} + l_{ssim} + l_{iou} \tag{9}$$

where l_{bce}, l_{ssim}, l_{iou} denote BCE loss, SSIM loss and IoU loss, respectively.

BCE loss is a widely employed loss in binary classification and salient object detection. BCE loss is defined as follows:

$$l_{bce} = -\sum_{(r,c)} [G(r,c) \log(S(r,c)) + (1 - G(r,c)) \log(1 - S(r,c))] \tag{10}$$

where $G(r,c) \in 0, 1$ is the ground truth label of the pixel (r,c) and $S(r,c)$ is the predicted saliency value.

SSIM captures the structural information of an image; it was originally proposed for measuring the similarity of two images. Let x and y be the pixel values of two patches cropped from the saliency map S and the corresponding ground truth mask G. The SSIM of x and y is expressed as:

$$l_{ssim} = 1 - \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{11}$$

where μ_x, μ_y and σ_x, σ_y are the mean and standard deviation, respectively, of x and y. σ_{xy} is their covariance; and $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are used to avoid dividing by zero.

In object detection, IoU is applied to evaluate the coincidence degree between the predicted bounding box and the ground truth box. Now, IoU has been applied as a training loss with the following definition:

$$l_{iou} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} S(r,c)\, G(r,c)}{\sum_{r=1}^{H}\sum_{c=1}^{W} [S(r,c) + G(r,c) - S(r,c)\, G(r,c)]} \tag{12}$$

where H and W are the height and width of the image. $G(r,c)$ and $S(r,c)$ have the same meanings as those in Eq. (10).

4 Experiments

4.1 Datasets

There are five mainstream VSOD datasets, which are divided into two types: 1) key-frame annotation datasets, such as **ViSal** [35] and **VOS** [36], and 2) dense (per-frame) annotation datasets, including **SegV2** [37], **DAVIS** [38] and **DAVSOD** [14]. **ViSal** is the first dataset intended for video salient object detection. It contains 17 video sequences and 193 labeling frames; on average, every 5 images have a label. **VOS** consists of 200 video sequences; on average, every 15 frames have a label, and there are 7467 labeling frames in total. The limitation of this dataset is its low diversity and generality. Only simple indoor, stable camera scenarios are contained in the VOS dataset. **SegV2** is

a dataset originally provided for video segmentation that consists of 13 video sequences and 1065 densely annotated frames. **DAVIS** consists of 50 video sequences and 3455 densely labeled frames. The dataset is very challenging as occlusion, motion blur, and appearance change scenarios are included. **DAVSOD** is the up-to-date and largest dataset provided for video salient object detection; it consists of 226 video sequences and 23938 per-frame labeling frames. The main difference between DAVSOD and other datasets is that DAVSOD emphasizes the shift of salient objects in dynamic scenes. In addition, instance-level object annotation and real human eye fixation annotation are provided in DAVSOD for further study of video saliency detection. Table 1 provides a comprehensive description of the datasets.

Table 1: A comprehensive description of the datasets

Datasets	Video sequences	Labeling frames	Size	Training	Test	Validation	Annotation
ViSal	17	193	$320 * 240$	0	17	0	Key-frame
VOS	200	7467	$800 * 448$	160	40	0	Key-frame
DAVIS	50	3455	$854 * 480$	30	20	0	Per-frame
SegV2	13	1065	$640 * 360$	0	13	0	Per-frame
DAVSOD	226	23938	$640 * 360$	61	80	85	Per-frame

4.2 Implementation and Experimental Setup

First, we initialize our backbone with a Shunted-Transformer-s [32] pre-trained on ImageNet. Then, we pre-train the whole MSTENet on the image salient object detection dataset DUTS-TR [39] and fine-tune the pre-trained model on the training set of DAVIS and DAVSOD.

We resize all the frames to the same spatial size of 320×320 before feeding them to the network. The number of input frames each time is set to 4 due to the limitation of GPU memory. The training frames are augmented using various strategies, including random flipping, cropping, rotation, and color enhancement, to avoid overfitting. The optimization algorithm is AdamW, and the learning rate is set to 1e-4 and 1e-5 for pretraining and fine-tuning, respectively. Cosine decay is chosen as the learning rate schedule. 20 epochs are separately carried out for both the pretraining stage and fine-tuning stage. The batchsize is set to one. We implement our network based on the publicly available PyTorch 0.4.0 framework. The proposed method is evaluated on the test sets of DAVIS, SegV2, ViSal, VOS, and DAVSOD benchmarks. Both training and testing are implemented on a PC with an NVIDIA GeForce GTX Titan XP GPU.

4.3 Evaluation Metrics

Three indicators are employed to evaluate our method: mean absolute error (MAE), F-measure, and structural measurement (S-measure).

The MAE measures the difference between the resulting saliency map and the ground truth pixel by pixel with the following definition:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |f_i - y_i| \tag{13}$$

where f_i denotes the resulting saliency map value and y_i denotes the ground truth value. N represents the total number of pixels within the image.

The F-measure is a weighted harmonic mean of precision and recall and is defined as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{14}$$

We use the max F-measure over all thresholds from 0 to 255 and $\beta^2 = 0.3$.

The S-measure is used to evaluate the structural similarity between the predicted saliency map and the ground truth mask and is defined as follows:

$$S_\lambda = \alpha \times S_o + (1-\alpha) \times S_r \tag{15}$$

where $\alpha \in [0,\ 1]$ and is usually set to 0.5. S_o is the object-aware structural similarity measure, and S_r is the region-aware structural similarity measure.

4.4 Comparisons with State-of-the-Arts

We compare our method with 18 state-of-the-art deep learning-based video salient object detection methods: SCANet [40], DSNet [41], DCFNet [31], STVS [42], MQP [43], STANet [44], PCSA [45], TENet [17], LSDGCN [46], MGA [16], SSAV [14], PoolNet [47], CPD [48], DSSANet [27], RCRNet [49], PDB [15], MBNM [50], and SCOM [51]. Among these methods, DCFNet is based on dynamic filtering; SCANet and STVS are based on 3D convolution; STANet, SSAV, RCRNet, and PDB are ConvLSTM-based methods; DSNet, MQP, LSDGCN, MGA, SCNN, MBNM, and SCOM are optical flow-based methods; and TENet is ConvLSTM- and optical flow-based method.

4.4.1 Quantitative Evaluation

Table 2 shows the quantitative comparisons in terms of three metrics, including S_λ, F_β, MAE, on five widely utilized VSOD datasets. Our method achieves the performance of SOTA VSOD methods and ranks first with the DAVIS, SegV2, and ViSal datasets and second with the DAVSOD dataset. In particular, our method improves the S_λ, F_β by 5.5% and 10.3% with the SegV2 dataset and 3.2% and 5.7% with the DAVSOD dataset compared with [31]. The results on VOS are not as good as those on other datasets as our DEM only considers short temporal information. In the VOS dataset, every 15 frames have a label, which generates a long-term span between adjacent labeling frames. More quantitative comparisons are demonstrated in Figs. 5–7.

Table 2: Quantitative comparisons of three evaluation metrics on five widely utilized VSOD datasets. All the chosen methods are deep learning-based methods. "↑" & "↓" indicate that larger or smaller is better. The top three results are marked in **boldface**, red and green fonts

		DAVIS			SegV2			ViSal			VOS			DAVSOD		
Years	Method	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
18–19	SCOM	0.832	0.783	0.048	0.815	0.764	0.030	0.762	0.831	0.122	0.712	0.690	0.162	0.599	0.464	0.220
	MBNM	0.887	0.861	0.031	0.809	0.716	0.026	0.898	0.883	0.020	0.742	0.670	0.099	0.637	0.520	0.159
	PDB	0.882	0.855	0.028	0.864	0.800	0.024	0.907	0.888	0.032	0.818	0.742	0.078	0.698	0.572	0.116
	RCRNet	0.886	0.848	0.027	0.842	0.781	0.035	0.922	0.906	0.026	**0.873**	**0.833**	0.051	0.741	0.653	0.087
	CPD	0.863	0.826	0.030	0.851	0.839	0.018	0.944	0.944	0.013	0.800	0.731	0.065	0.697	0.595	0.086
	PoolNet	0.854	0.815	0.038	0.782	0.704	0.025	0.902	0.891	0.025	0.773	0.709	0.082	0.702	0.592	0.089
	SSAV	0.893	0.861	0.028	0.851	0.801	0.023	0.943	0.939	0.020	0.819	0.742	0.073	0.724	0.603	0.092
	MGA	0.912	0.892	0.022	0.864	0.821	0.030	0.941	0.940	0.016	0.792	0.767	0.063	0.751	0.656	0.081

(Continued)

Table 2 (continued)

Years	Method	DAVIS			SegV2			ViSal			VOS			DAVSOD		
		$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
	LSDGCN	0.897	0.891	0.021	0.880	0.866	0.018	0.950	0.952	0.012	0.850	0.792	0.050	0.768	0.689	0.075
	TENet	0.905	0.881	0.017	0.868	0.810	0.025	0.949	0.949	0.012	0.845	0.781	0.052	0.779	0.697	0.070
	PCSA	0.902	0.880	0.022	0.865	0.810	0.025	0.946	0.940	0.017	0.827	0.747	0.065	0.741	0.655	0.086
	STANet	0.892	0.883	0.025	0.872	0.769	0.031	0.910	0.893	0.021	0.804	0.701	0.062	0.706	0.549	0.107
20–	MQP	0.916	0.904	0.018	0.882	0.841	0.018	0.942	0.939	0.016	0.828	0.768	0.069	0.770	0.703	0.075
22	STVS	0.892	0.862	0.023	0.891	0.860	0.017	0.952	0.952	0.013	0.850	0.791	0.058	0.746	0.651	0.086
	DCFNet	0.914	0.900	0.016	0.883	0.839	0.015	0.952	0.953	0.010	0.846	0.791	0.060	0.741	0.660	0.074
	DSNet	0.914	0.891	0.018	0.875	0.832	0.028	0.949	0.950	0.013	0.855	0.801	0.060	0.729	0.627	0.077
	SCANet	0.902	0.881	0.021	0.906	0.890	0.026	**0.954**	0.955	0.011	0.872	0.828	**0.048**	**0.801**	**0.731**	0.064
	Ours	**0.924**	**0.914**	**0.012**	**0.932**	**0.926**	**0.010**	0.953	**0.961**	**0.009**	0.863	0.825	0.049	0.765	0.698	**0.063**

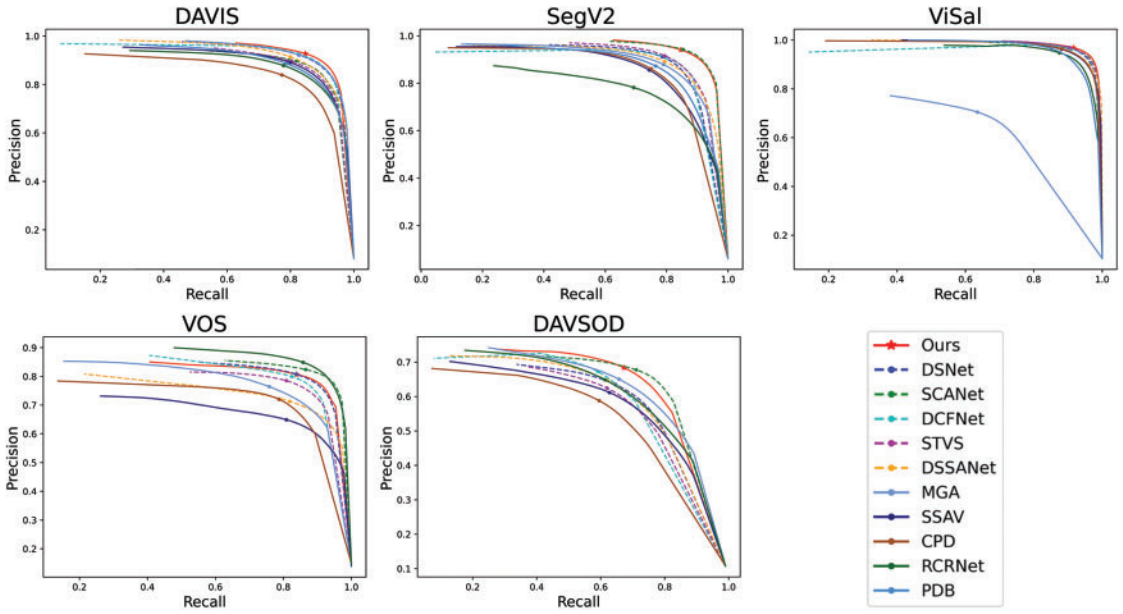

Figure 5: Precision-recall curve comparison of the MSTENet and other SOTA VSOD methods on five widely used datasets

Fig. 5 shows the precision-recall curves of MSTENet and other SOTA methods on five widely used datasets. The results indicate that MSTENet outperforms those methods in precision scores and recall scores in DAVIS, SegV2, and Visal datasets. Our method exhibits suboptimal performance on the VOS dataset compared to RCRNet and SCANet. This discrepancy can be attributed to the relatively inferior performance of our DEM module in handling long-term temporal information when contrasted with the ConvGRU module utilized in RCRNet and the 3D convolutional module employed in SCANet. Fig. 6 shows the F-measure curves of MSTENet and other SOTA methods on five widely used datasets. F-measure is a weighted harmonic mean of precision and recall, therefore, the performance conveyed through the F-measure is consistent with that depicted in Fig. 5. The S-measure-MAE scatter plots depicted in Fig. 7 measure the degree of similarity between the predicted saliency maps and the ground truth. In the S-measure-MAE scatter plot, the x-axis corresponds to the MAE while the y-axis corresponds to the S-measure. It is worth noting that superior algorithm

performance is indicated by the proximity of data points to the upper right corner of the plot. Fig. 7 demonstrates that the MSTENet achieves superior performance compared to other SOTA methods on four out of five datasets.

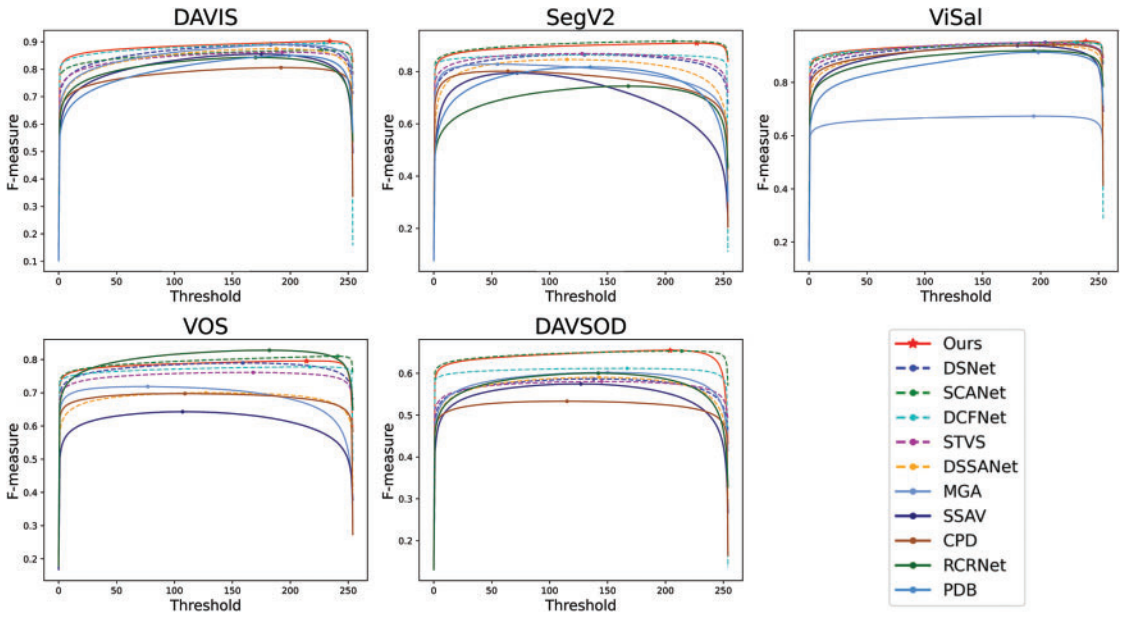

Figure 6: F-measure curve comparison of the MSTENet and other SOTA VSOD methods on five widely used datasets

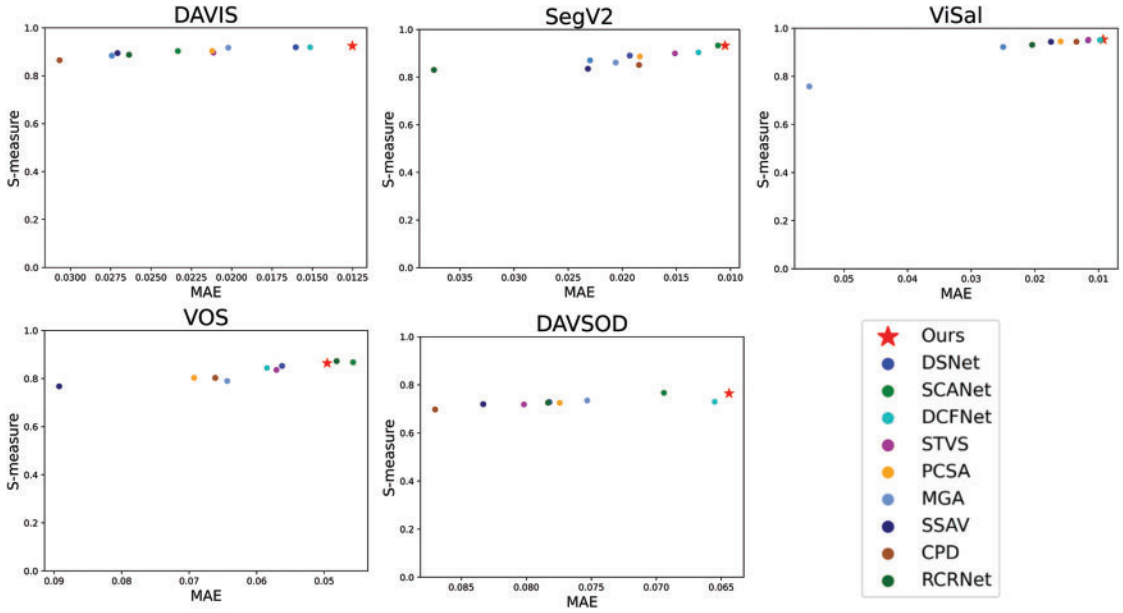

Figure 7: The S-measure-MAE scatter plots of the MSTENet and other SOTA VSOD methods on five widely used datasets. In each subfigure, the x-axis shows MAE and the y-axis represents the S-measure. The closer to the upper right corner, the better the algorithm performance

4.4.2 Qualitative Evaluation

To further illustrate the superior performance of our method, Fig. 8 shows several representative samples of our model and eleven other SOTA methods on the DAVIS dataset. Our method can accurately detect salient objects in various challenging scenarios, including images with some small but elaborate objects (1st and 2nd rows), large objects (3rd and 4th rows), a cluttered background (5th and 6th rows), low contrast and a complex foreground (7th and 8th rows). In addition, the visual results demonstrate that our method achieves clearer salient object boundaries and more accurate local salient regions with rich details.

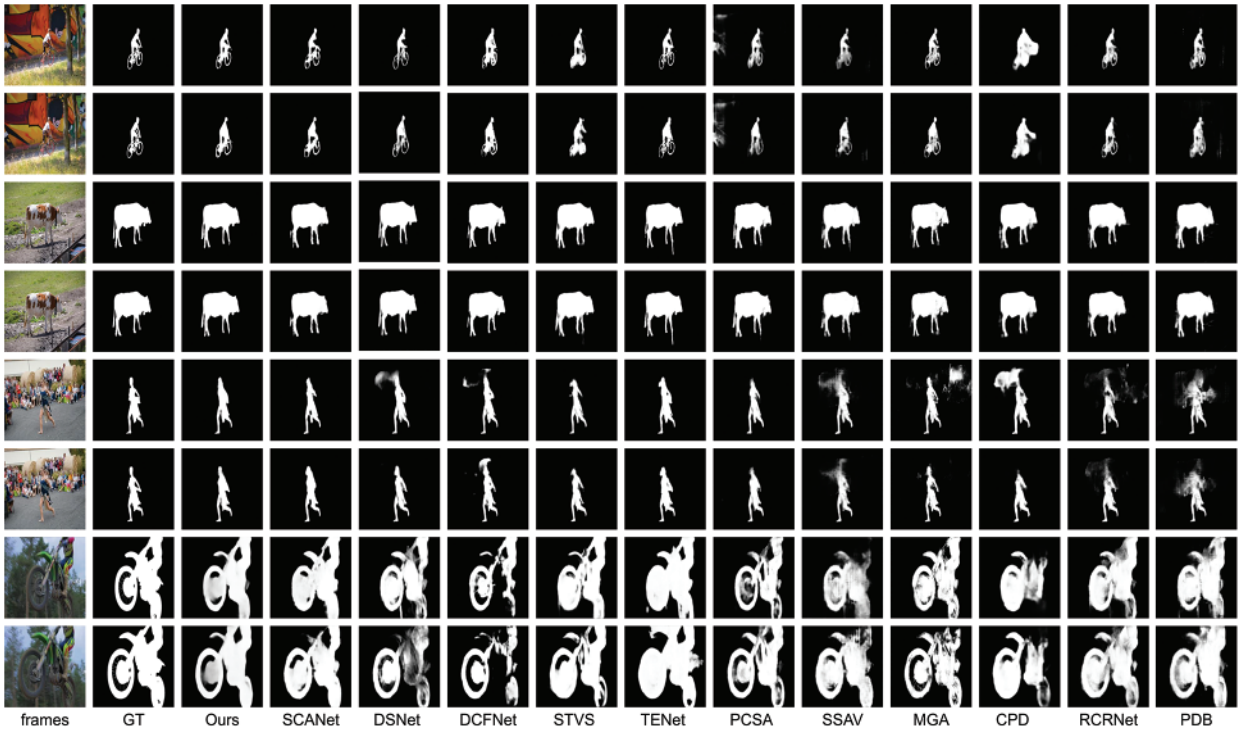

Figure 8: Qualitative comparisons of MSTENet with eleven existing state-of-the-art VSOD methods

4.4.3 Time Consumption Evaluation

We use the FLOPs (floating-point operations) to measure the computational complexity of the model. The FLOPs required for a convolution operation depend on multiple factors, such as the size of the convolution kernel, the dimensions of the input and output feature maps, and the number of channels involved in the feature maps. This can be expressed as follows:

$$FLOPs = H \times W \times (C_{in} \times K_w \times K_h + 1) \times C_{out} \tag{16}$$

where H, W, and C_{in} are the height, width, and number of channels of the input feature map, respectively. K_h and K_w are the kernel height and width, and C_{out} is the output channels. As presented in Table 3, the proposed method achieves minimal FLOPs and faster running speed compared with some optical flow-based methods and ConvLSTM-based methods. The complex structure of ConvLSTM and the need for the computation of optical flow maps result in high time consumption of ConvLSTM-based and optical flow-based methods.

Table 3: Time-cost comparison with other methods

Methods	Input size	Speed (FPS)	FLOPs (G)	Based	Platform
PDB	$473*473$	20	282.6	ConvLSTM	1080Ti
RCRNet	$448*448$	27	442.4	ConvGRU	1080Ti
SSAV	$473*473$	20	279.7	ConvLSTM	Titan Xp
MGA	$512*512$	14	494.3	Optical flow	1080Ti
LSDGCN	$256*256$	8	706.7	Optical flow	2080Ti
TENet	$256*256$	17	332.6	–	1080Ti
STANet	$473*473$	17	403.3	ConvLSTM	Tesla_V100
MQP	$352*352$	20	237.5	Optical flow	Titan Xp
DCFNet	$448*448$	16	373.2	Dynamic filtering	Titan Xp
Ours	$320*320$	27	209.4	–	Titan Xp

Fig. 9 depicts the relationship between running speed (frames per second, FPS) and accuracy evaluation metrics. We define Acc = s-measure + F-measure + (1-MAE), representing a comprehensive performance evaluation index that captures multiple aspects of algorithm performance. Closer to the upper right corner indicates both higher FPS and evaluation metrics. It visually demonstrates that our model achieves an optimal balance between time and accuracy.

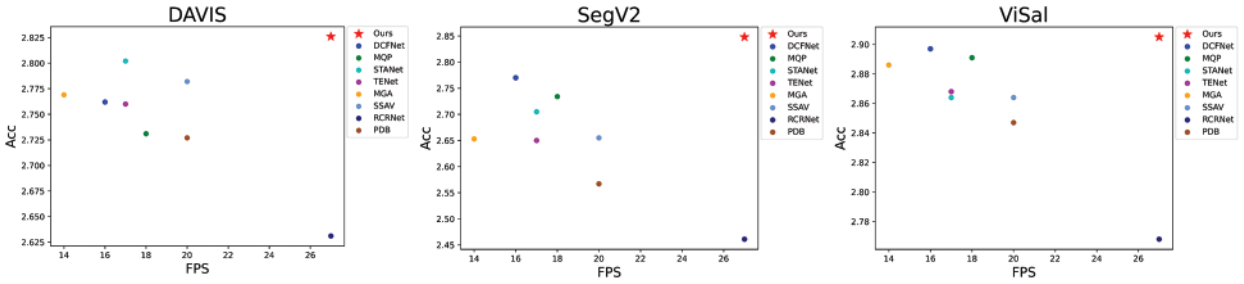

Figure 9: Relationship between FPS and evaluation metrics, where Acc = S-measure + F-measure + (1−MAE). Closer to the upper right corner indicates both higher FPS and evaluation metrics

4.5 Ablation Studies

In this section, we conduct extensive experiments to illustrate the effectiveness of each component of our method. The ablation study contains two parts: background stream ablation and DEM ablation.

4.5.1 Background Stream Ablation

To prove the effectiveness of the proposed foreground-background fusion strategy, we remove background supervision from the MSTENet. The visual samples in Fig. 10 and quantitative metrics in Table 4 show that the proposed foreground-background fusion strategy takes advantage of the complementary relationship between foreground saliency cues and background saliency cues, and outperforms the foreground supervised method by a large margin on all five experiment datasets. As depicted in Fig. 10, our method tends to generate saliency maps with both lower false positives

(2nd and 3rd columns) and false negatives (1st and 4th columns) compared to the foreground-based method.

Figure 10: Effectiveness of background supervision. From top to bottom: video frames, GT, and saliency maps generated by methods with background supervision and without background supervision

Table 4: The quantitative analysis of ablation studies. "Base" denotes the model without background stream (BS) and DEM, and "+" denotes adding the module to the base model

	DAVIS			SegV2			ViSal			VOS			DAVSOD		
Method	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
Base	0.898	0.883	0.018	0.928	0.922	0.011	0.950	0.957	0.010	0.822	0.805	0.063	0.751	0.671	0.068
+BS	0.909	0.892	0.016	0.929	0.924	**0.009**	0.951	0.958	0.010	0.845	0.806	0.051	0.755	0.696	0.065
+BS+DEM	**0.924**	**0.914**	**0.012**	**0.932**	**0.926**	0.010	**0.953**	**0.961**	**0.009**	**0.863**	**0.825**	**0.049**	**0.765**	**0.698**	**0.063**

4.5.2 DEM Ablation

To demonstrate the effectiveness of the proposed difference enhancement module, we conducted the ablation experiment by removing the module. It can be considered a degeneration from video to

still image saliency detection. Fig. 11 shows that noises and blurry boundaries will be introduced to changing regions in the resulting saliency map due to a lack of temporal information. Table 4 shows the quantitative comparison, which illustrates that the metrics with datasets DAVIS, VOS, and DAVSOD suffer from significant falls when removing the DEM from the proposed method.

Figure 11: Effectiveness of DEM. From top to bottom: video frames, GT, and saliency maps generated by methods with and without the DEM module

4.6 Limitations and Future Directions

The frames presented in Fig. 12 are extracted from the VOS dataset at an interval of 15 frames, specifically identified by the frame numbers 00614, 00629, and 00644. In contrast to the ground truth, which exactly depicts the dancing woman in motion, our method inadvertently identifies both the woman and the static chair as salient objects. Because our method takes 4 consecutive video frames as its input each time and only differences between adjacent frames are computed, its sensing scope over the temporal scale is quite limited and fails to eliminate the objects undergoing a long static period. In the future, we plan to improve the performance of VSOD methods from two aspects. First, we aim to explore lightweight modules for extracting long-term temporal information, to alleviate the trade-off between accuracy and computational complexity. Second, we intend to design novel spatial-temporal interaction strategies, enabling the collective improvement of both temporal and spatial features.

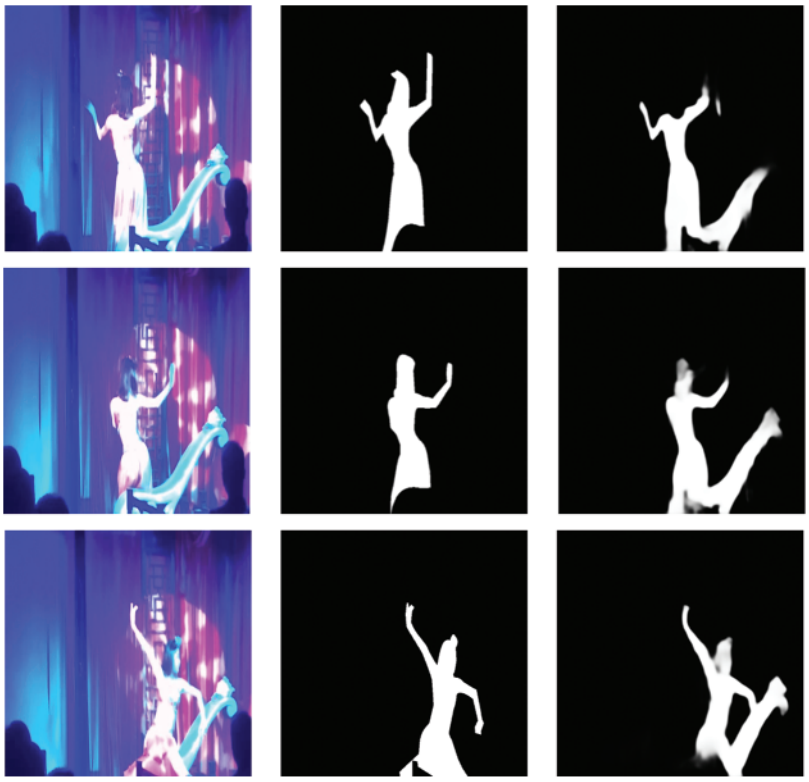

Figure 12: Some failure samples of the proposed method. From left to right: Video frames, GT, and saliency maps generated by the proposed method

5 Conclusion

In this paper, we work on the challenge of accurate video saliency detection in real-time. The major highlights of our approach can be summarized as follows:

First, we propose a multi-stream structure consisting of background, foreground, and fused saliency cues, wherein the background and foreground saliency cues are progressively integrated with a novel FBFM module. The structure ensures precise salient detection in the spatial domain. Second, a simple and effective difference enhancement module is proposed to capture motion information in the temporal domain. Our lightweight temporal module can be treated as a plug-in to be inserted into the decoder stage, enabling the original spatial decoder to sense temporal information. Last, we conduct extensive quantitative and qualitative evaluations to show the advantages of the proposed model and verify the effectiveness of each of its main components with ablation analysis. The evaluation metrics indicate that our model outperforms SOTA methods, ranking first in terms of performance on the DAVIS, SegV2, and ViSal datasets, and second on the DAVSOD dataset. The comparison of time consumption reveals that the proposed model achieves optimal computational performance, exhibiting lower FLOPs.

The MSTENet can be transferred into other computer vision tasks. The multi-stream structure, along with the incorporated fusion module FBFM, exhibits a high degree of generalizability for diverse computer vision tasks that concern multi-stream fusion, including RGB-D saliency detection and

video classification. The lightweight DEM module can seamlessly serve as a plug-in, enabling its integration into various video processing tasks, such as object tracking and video surveillance. In addition to its extensive utilization in computer vision tasks, MSTENet finds significant application within the domain of autonomous vehicles and traffic management, wherein the multi-stream structure of MSTENet enhances the accuracy of detection results in complex scenarios, while the DEM module guarantees the real-time requirements of the associated applications.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Natural Science Foundation China (NSFC) under Grant No. 62203192.

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: Dan Xu; data collection: Jiale Ru; analysis and interpretation of results: Jinlong Shi, Jiale Ru; draft manuscript preparation: Dan Xu, Jinlong Shi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. P. Ji, K. Gu, Z. Wu, D. P. Fan, J. Shen *et al.,* "Full-duplex strategy for video object segmentation," in *Int. Conf. on Computer Vision*, Montreal, QC, Canada, pp. 4922–4933, 2021.

[2] M. F. Alotaibi, M. Omri, S. A. Khalek, E. Khalil and R. F. Mansour, "Computational intelligence-based harmony search algorithm for real-time object detection and tracking in video surveillance systems," *Mathematics*, vol. 10, no. 5, pp. 733, 2022.

[3] H. Lee and D. Kim, "Salient region-based online object tracking," in *IEEE Winter Conf. on Applications of Computer Vision*, Lake Tahoe, NV, USA, pp. 1170–1177, 2018.

[4] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2014.

[5] M. U. Nisa, D. Mahmood, G. Ahmed, S. Khan, M. A. Mohammed *et al.,* "Optimizing prediction of YouTube video popularity using XGBoost," *Electronics*, vol. 10, no. 23, pp. 2962, 2021.

[6] H. Singh, M. Verma and R. Cheruku, "DSNet: Efficient lightweight model for video salient object detection for IoT and WoT applications," in *WWW'23 Companion: Companion Proceedings of the ACM Web Conference 2023*, Austin, TX, USA, pp. 1286–1295, 2023.

[7] E. Bas, A. M. Tekalp and F. S. Salman, "Automatic vehicle counting from video for traffic flow analysis," in *IEEE Intelligent Vehicles Symp.*, Istanbul, Turkey, pp. 392–397, 2007.

[8] Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu *et al.,* "Deeply supervised salient object detection with short connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 815–828, 2019.

[9] Y. Ji, H. Zhang, Z. Zhang and M. Liu, "CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances," *Information Sciences*, vol. 546, pp. 835–857, 2021.

[10] Q. Ren and R. Hu, "Multi-scale deep encoder-decoder network for salient object detection," *Neurocomputing*, vol. 316, pp. 95–104, 2018.

[11] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan *et al.,* "BASNet: Boundary-aware salient object detection," in *Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 7479–7489, 2019.

[12] T. N. Le and A. Sugimoto, "Deeply supervised 3D recurrent FCN for salient object detection in videos," in *British Machine Vision Conf.*, London, UK, pp. 31–38, 2017.
[13] T. N. Le and A. Sugimoto, "Video salient object detection using spatiotemporal deep features," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5002–5015, 2018.
[14] D. P. Fan, W. Wang, M. M. Cheng and J. Shen, "Shifting more attention to video salient object detection," in *Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 8554–8564, 2019.
[15] H. Song, W. Wang, S. Zhao, J., Shen and K. M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," *European Conference on Computer Vision*, vol. 11215, pp. 715–731, 2018.
[16] H. Li, G. Chen, G. Li and Y. Yu, "Motion guided attention for video salient object detection," in *Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 7274–7283, 2019.
[17] S. Ren, C. Han, X. Yang, G. Han and S. He, "TENet: Triple excitation network for video salient object detection," *European Conference on Computer Vision*, vol. 12350, pp. 212–228, 2020.
[18] A. Borji and L. Ltti, "Exploiting local and global patch rarities for saliency detection," in *Computer Vision and Pattern Recognition*, Providence, RI, USA, pp. 478–485, 2012.
[19] Q. Yan, L. Xu, J. Shi and J. Jia, "Hierarchical saliency detection," in *Computer Vision and Pattern Recognition*, Portland, OR, USA, pp. 1155–1162, 2013.
[20] R. Fu, C. Chen, S. Yan, A. A. Heidari, X. Wang *et al.,* "Gaussian similarity-based adaptive dynamic label assignment for tiny object detection," *Neurocomputing*, vol. 543, pp. 126285, 2023.
[21] R. Zhao, W. Ouyang, H. Li and X. Wang, "Saliency detection by multi-context deep learning," in *Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 1265–1274, 2015.
[22] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li *et al.,* "Non-local deep features for salient object detection," in *Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 6609–6617, 2017.
[23] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 478–487, 2016.
[24] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya and R. V. Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 5781–5790, 2016.
[25] N. Liu, J. Han and M. H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3089–3098, 2018.
[26] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 3085–3094, 2019.
[27] C. Xu, Z. Gao, H. Zhang, S. Li and V. H. C. Albuquerque, "Video salient object detection using dual-stream spatiotemporal attention," *Applied Soft Computing*, vol. 108, pp. 107433, 2021.
[28] C. Bak, A. Kocak, E. Erdem and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1688–1698, 2018.
[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Int. Conf. on Computer Vision*, Santiago, Chile, pp. 4489–4497, 2015.
[30] G. Li, Y. Xie, T. Wei, K. Wang and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3243–3252, 2018.
[31] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao *et al.,* "Dynamic context-sensitive filtering network for video salient object detection," in *Int. Conf. on Computer Vision*, Montreal, QC, Canada, pp. 1553–1563, 2021.
[32] S. Ren, D. Zhou, S. He, J. Feng and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 10843–10852, 2022.
[33] W. Zhao, J. Zhang, L. Li, N. Barnes, N. Liu *et al.,* "Weakly supervised video salient object detection," in *Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 16821–16830, 2021.
[34] T. Zhou, H. Fu, G. Chen, Y. Zhou, D. P. Fan *et al.,* "Specificity-preserving RGB-D saliency detection," in *Int. Conf. on Computer Vision*, Montreal, QC, Canada, pp. 4661–4671, 2021.
[35] W. Wang, J. Shen and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015.

[36] J. Li, C. Xia and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 349–364, 2018.
[37] F. Li, T. Kim, A. Humayun, D. Tsai and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Int. Conf. on Computer Vision*, Sydney, NSW, Australia, pp. 2192–2199, 2013.
[38] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross *et al.,* "A benchmark dataset and evaluation methodology for video object segmentation," in *Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 724–732, 2016.
[39] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang *et al.,* "Learning to detect salient objects with image-level supervision," in *Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 3796–3805, 2017.
[40] T. Chen, J. Xiao, X. Hu, G. Zhang and S. Wang, "Spatiotemporal context-aware network for video salient object detection," *Neural Computing and Applications*, vol. 34, pp. 16861–16877, 2022.
[41] J. Liu, J. Wang, W. Wang and Y. Su, "DS-Net: Dynamic spatiotemporal network for video salient object detection," *Digital Signal Processing*, vol. 130, pp. 103700, 2022.
[42] C. Chen, G. Wang, C. Peng, Y. Fang, D. Zhang *et al.,* "Exploring rich and efficient spatial temporal interactions for real-time video salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3995–4007, 2021.
[43] C. Chen, J. Song, C. Peng, G. Wang and Y. Fang, "A novel video salient object detection method via semisupervised motion quality perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2732–2745, 2022.
[44] H. B. Bi, D. Lu, H. H. Zhu, L. N. Yang and H. P. Guan, "STA-Net: Spatial-temporal attention network for video salient object detection," *Applied Intelligence*, vol. 51, pp. 3450–3459, 2021.
[45] Y. Gu, L. Wang, Z. Wang, Y. Liu, M. M. Cheng *et al.,* "Pyramid constrained self-attention network for fast video salient object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 10869–10876, 2020.
[46] B. Wang, W. Liu, G. Han and S. He, "Learning long-term structural dependencies for video salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 9017–9031, 2020.
[47] J. J. Liu, Q. Hou, M. M. Cheng, J. Feng and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3912–3921, 2019.
[48] Z. Wu, L. Su and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3902–3911, 2019.
[49] P. Yan, G. Li, Y. Xie, Z. Li, C. Wang *et al.,* "Semi-supervised video salient object detection using pseudo-labels," in *Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 7284–7293, 2019.
[50] S. Li, B. Seybold, A. Vorobyov, X. Lei and C. C. Jay Kuo, "Unsupervised video object segmentation with motion-based bilateral networks," *European Conference on Computer Vision*, vol. 11207, pp. 207–223, 2018.
[51] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu *et al.,* "SCOM: Spatiotemporal constrained optimization for salient object detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3345–3357, 2018.