Information Sciences 628 (2023) 134–147

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

STI-Net: Spatiotemporal integration network for video saliency detection

Xiaofei Zhou [a], Weipeng Cao [b,*], Hanxiao Gao [a], Zhong Ming [b], Jiyong Zhang [a]

[a] School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China
[b] Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen), Shenzhen 518107, China

ARTICLE INFO

Keywords:
Spatiotemporal saliency
Feature aggregation
Saliency prediction
Saliency fusion

ABSTRACT

Image saliency detection, to which much effort has been devoted in recent years, has advanced significantly. In contrast, the community has paid little attention to video saliency detection. Especially, existing video saliency models are very likely to fail in videos with difficult scenarios such as fast motion, dynamic background, and nonrigid deformation. Furthermore, performing video saliency detection directly using image saliency models that ignore video temporal information is inappropriate. To alleviate this issue, this study proposes a novel end-to-end spatiotemporal integration network (STI-Net) for detecting salient objects in videos. Specifically, our method is made up of three key steps: feature aggregation, saliency prediction, and saliency fusion, which are used sequentially to generate spatiotemporal deep feature maps, coarse saliency predictions, and the final saliency map. The key advantage of our model lies in the comprehensive exploration of spatial and temporal information across the entire network, where the two features interact with each other in the feature aggregation step, are used to construct boundary cue in the saliency prediction step, and also serve as the original information in the saliency fusion step. As a result, the generated spatiotemporal deep feature maps can precisely and completely characterize the salient objects, and the coarse saliency predictions have well-defined boundaries, effectively improving the final saliency map's quality. Furthermore, "shortcut connections" are introduced into our model to make the proposed network easy to train and obtain accurate results when the network is deep. Extensive experimental results on two publicly available challenging video datasets demonstrate the effectiveness of the proposed model, which achieves comparable performance to state-of-the-art saliency models.

1. Introduction

Saliency detection is a popular research topic that aims to identify the most visually appealing objects in an image or video. It is especially useful in a variety of applications such as video saliency detection [18,34], image retrieval [22,48], and video-based traffic control [43,8]. In general, there are two settings for saliency detection: one estimates where humans look in an image, while the other attempts to pop-out all salient objects in an image or video. This paper is concerned with the latter. Furthermore, saliency models are typically classified as static image saliency models or dynamic video saliency models based on the input. So far, a large

* Corresponding author.
E-mail addresses: zxforchid@outlook.com (X. Zhou), caoweipeng@gml.ac.cn (W. Cao), gaohx@hdu.edu.cn (H. Gao), mingz@szu.edu.cn (Z. Ming), jzhang@hdu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.ins.2023.01.106
Received 16 May 2021; Received in revised form 18 January 2023; Accepted 24 January 2023

Available online 28 January 2023

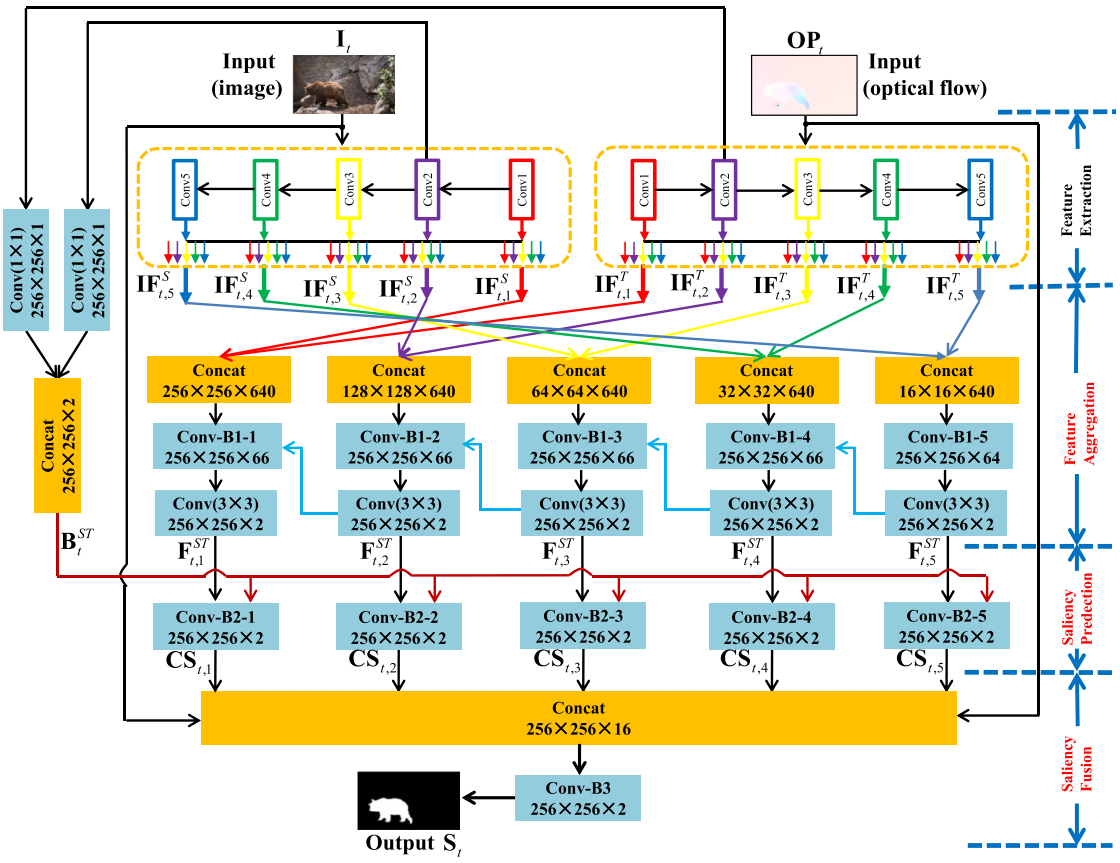

Fig. 1. Illustration of the proposed spatiotemporal integration network (STI-Net). Given the current frame $\mathbf{I}_t$ (256×256×3) and its optical flow image $\mathbf{OP}_t$ (256×256×3), feature extraction modules are first employed to generate the initial spatial deep features $\{\mathbf{IF}_{t,i}^S\}_{i=1}^5$ and the initial temporal deep features $\{\mathbf{IF}_{t,i}^T\}_{i=1}^5$. Then, the feature aggregation step is performed on these initial deep features, yielding the spatiotemporal deep feature map $\mathbf{F}_{t,i}^{ST}$ $(i = 1, \cdots, 5)$. After that, by considering the spatiotemporal boundary cue $\mathbf{B}_t^{ST}$ originated from spatial and temporal domains, we perform saliency prediction on these spatiotemporal deep feature maps and acquire the coarse saliency prediction $\mathbf{CS}_{t,i}$ $(i = 1, \cdots, 5)$. Finally, by incorporating the original information including $\mathbf{I}_t$ and $\mathbf{OP}_t$, these prediction results are further fused to obtain the final saliency map $\mathbf{S}_t$.

number of studies have focused on detecting saliency in static images. In stark contrast, efforts on video salient object detection have received remarkably little attention due to the high complexity of video saliency computation and the scarcity of massive pixel-wise annotated video data.

The primary distinction between video and image lies in the temporal information, which is quite appealing to people's attention. Obviously, for video saliency detection, we should pay attention to one critical point: how to adequately integrate the spatial information and temporal information using some efficient utilization methods. According to the crucial point, many video saliency models have been developed based on various aspects such as the center-surround mechanism [29], machine learning [25], information fusion [5], and regional saliency assessment [49]. The aforementioned video saliency models can achieve comparable performance, but these will deteriorate in handling some challenging scenarios such as dynamic background, non-rigid deformation, and fast motion.

Fortunately, the performance of image saliency models has improved significantly since the successful use of convolutional neural networks. As a result, an intuitive thought occurs to us: can we directly apply these image saliency models to perform video saliency detection? The path is obviously unavailable due to the lack of temporal information in image saliency models, which are unable to maintain visual coherence and temporal correlation in the generated saliency maps. Meanwhile, convolutional neural networks have recently been extended to video saliency models [39,9].

Motivated by the aforementioned discussion, we present a novel end-to-end spatiotemporal integration network (STI-Net) for video saliency detection, as shown in Fig. 1, which contains three key steps: feature aggregation, saliency prediction, and saliency fusion. The key advantage of our model lies in that spatial and temporal information can be adequately integrated throughout the overall network. Firstly, in the feature aggregation step, the network is designed to enable an interaction between spatial information and temporal information via concatenation, convolution and recurrent connections. In this way, the generated spatiotemporal deep features can effectively characterize salient objects in videos. Secondly, to ensure that the saliency maps produced have well-defined boundaries, the spatiotemporal boundary cue originated from both spatial and temporal domains is incorporated into the saliency prediction step. Thirdly, in order to fuse the coarse prediction results effectively, the original or initial information, *i.e.* current frame

and optical flow image, is used to provide the complementary information for the correction of prediction results in the fusion step. Furthermore, to aid network optimization and obtain more accurate results as the depth of our network increases, the shortcut connections adopted in [11] are imposed on the convolutional blocks located in the feature aggregation and saliency fusion steps, *i.e.* Conv-B1 and Conv-B3, as shown in Figs. 2 and 5, respectively. Consequently, the proposed spatiotemporal saliency network (STI-Net) can achieve very promising results in difficult scenarios.

The most related work (denoted as Amulet) to this paper is proposed in [45], which investigates the aggregation of multilevel convolutional feature maps for saliency detection in still images. The differences between our model and Amulet are evident in both input and model architecture. Firstly, with the introduction of temporal information, *i.e.* optical flow image, we design a more comprehensive architecture for feature aggregation compared to Amulet. In the feature aggregation step of our model, feature interaction occurs not only between features originating from different layers, but also between features originating from spatial and temporal information. Differently, Amulet only achieves the interaction between features originating from different layers. Furthermore, to make our model easy to train and obtain promising results, we design the "shortcut connections" based convolutional block Conv-B1-i shown in Fig. 2, which is also another difference between our model and Amulet. Certainly, inspired by Amulet, this block also introduces the recurrent connections to further accelerate the interaction of deep features. Secondly, the generation and utilization of boundary information are not the same. In the saliency prediction step of our model, we generate the spatiotemporal boundary cue from both spatial and temporal information, where the obtained boundary cue is concatenated with the deep features for saliency prediction. In contrast, Amulet only adopts the spatial information as the boundary cue, which is employed in a summation way. Thirdly, unlike Amulet, our model uses the initial information, namely current frame and optical flow image, to correct the predicted results in saliency fusion step. Further, in the fusion step, we also deploy the "shortcut connections" based convolutional block Conv-B3 shown in Fig. 5 to accelerate model optimization.

Other typical works related to this paper are [5,27,36], which are all built on hand-crafted low-level features (*e.g.*, color, edge, *etc*), and are dependent on motion information originated from optical flow. Furthermore, the predicted saliency map suffers from low accuracy due to the limited representability of low-level features. In contrast, our model makes adequate use of spatial information and temporal information throughout the network, resulting in the effective exploitation of deep features. As a result, the obtained saliency maps are of high quality (see Figs. 6 and 7). Thus, the difference between our model and the aforementioned models lies in the utilization of features.

Overall, the main contributions of our paper are summarized as follows:

1. We propose a novel end-to-end spatiotemporal saliency network (STI-Net) for salient object detection in videos that consists of three key steps containing feature aggregation, saliency prediction, and saliency fusion.
2. Spatial and temporal information are fully exploited throughout the network, namely interacting with one another, constructing boundary cue, and serving as the original information in feature aggregation step, saliency prediction step and saliency fusion step, respectively.
3. To make our model easy to train and obtain promising results, we first equip the "shortcut connections" to two types of convolutional blocks, *i.e.* Conv-B1 and Conv-B3, and then reuse the parameters of Amulet for initialization.

The rest of this paper is organized as follows. The related works are reviewed in Section 2. The proposed model is described in Section 3. The experimental results and the related analyses are presented in Section 4. Finally, we conclude this paper in Section 5.

2. Related works

This section reviews some works on detecting salient objects in static images and dynamic video sequences.

2.1. Image saliency detection

Saliency detection for static images has been sufficiently studied for many years. The pioneering model was proposed by Itti et al. in [17], in which orientation, color and luminance are used to compute center-surround difference. Referring to this famous framework, in [7], center-surround difference is measured using global contrast, which incorporates spatial distance of each pair of regions. Besides, in [28], Liu et al. employ the binary tree to model the saliency of each superpixel. Recently, some saliency models are also constructed on some machine learning techniques. For instance, in [26], conditional random field is utilized to aggregate multiple features and generate saliency maps. In [19], features, which are used to represent regions, are mapped to saliency scores by using a random forest regressor. Furthermore, some saliency models are built on convolutional neural networks. In [46], the multi-context deep model is proposed to handle the scene whose attributes are low contrast background and confusing visual appearance. In [45], multi-level deep features are effectively integrated by using the proposed bidirectional learning network. In [47], a simple gated network is proposed to control the information interactions between encoder and decoder. In [31], a two-level nested U-structure based network is designed to acquire contextual information and increase network's depth. In [40], a decomposition and completion network is employed to integrate edge and skeleton cues, which localizes boundaries and interiors of salient objects, respectively. In [21], a recursive CNN architecture equipped with a contour-saliency blending module is proposed to promote the information exchange between saliency and contour.

Overall, the above works focus on image saliency detection. However, it is inappropriate to directly employ them for video saliency detection, because the spatial and temporal cues should be taken into account simultaneously for videos. In this paper, to

perform salient object detection in videos, we adopt an off-the-shelf deep saliency model [45] as feature extraction module, and fully exploit the spatial information and temporal information of a video sequence by devising specific operations in the following three contents including feature aggregation, saliency prediction, and saliency fusion as shown in Fig. 1.

2.2. Video saliency detection

In contrast to image saliency detection, it is full of challenging for video saliency detection, due to the high computational complexity, the complex scene, and the lack of large-scale pixel-wise annotated video data. Generally speaking, the existing models can be categorized into two classes, one is the conventional models constructed using center-surround framework, information theory, control theory, traditional machine learning techniques, or information fusion method, *etc*, and the other one is deep learning based models.

Referring to the classical center-surround framework [17], in [16], multiple features consisting of motion energy, luminance, color and so on are employed to estimate feature difference. In [29], saliency scores are obtained by performing Kullback-Leibler divergence on dynamic features. Traditional machine learning techniques have also been employed by some video saliency models. For example, in [14], saliency map is constructed based on the diffusion of trajectories, which can be found by using one-class support vector machines. In [50], a random forest based video saliency model is learned on each frame. Besides, some saliency models are performed on segmented or superpixel regions, which are often used to constitute a graph. For instance, in [27], saliency is computed by using propagation method based on a superpixel-level graph. Furthermore, some other theoretical methods such as background priors and quantum cuts are also used to construct saliency models [5], which also achieves encouraging performances.

Recent few years, convolutional neural networks are also adopted by some video saliency models. For instance, in [39], an attention mechanism based two-stream spatiotemporal network is proposed to perform video salient object detection. In [37], static/dynamic saliency models are constructed using fully convolutional neural networks. Besides, in [33,9], the ConvLSTM [41] is employed to detect salient objects in videos. In [24], the video saliency model can be obtained by weakly retraining a pretrained image saliency deep model. In [20], the guidance and teaching network is proposed to conduct implicit guidance and explicit teaching strategies for video salient object detection. In [42], graph convolution networks based video saliency model is proposed to highlight complete salient objects with well-defined boundaries. In [6], a confidence-guided adaptive gate module and a dual differential enhancement module are used to purify and merge spatial and temporal cues. In [44], the proposed dynamic context sensitive filtering network employs the bidirectional dynamic fusion strategy to promote the interaction of spatial and temporal information. In addition, some other efforts are also devoted to the video research. For example, in [1], an evaluation method is proposed to simulate most of the details in a wireless automated video surveillance framework. In [12], the key frame is extracted by a bio-inspired method which is applied to optimize the quality of the frames after embedding watermark. In [8], the spatial and temporal relations are defined by mapping between vocabulary and the object movements.

Compared with all the aforementioned saliency models, our model can sufficiently exploit spatial information and temporal information, where the two modal cues interact with one another, are used to construct boundary information, and are served as the original information. Compared with the contemporary video saliency models, the proposed model achieves a comparable performance.

3. The proposed model

In this section, we give a detailed presentation for constructing and training the proposed spatiotemporal saliency network.

3.1. Overall architecture

Our model, as illustrated in Fig. 1, consists of three key steps including feature aggregation, saliency prediction, and saliency fusion. These three steps are trained in an end-to-end manner jointly. Concretely, the input of our model is the current frame $\mathbf{I}_t$ and optical flow image $\mathbf{OP}_t$. Here, $\mathbf{OP}_t$ is obtained by using large displacement optical flow method (LDOF) [4], which is further converted to a optical flow image [3] coded by 3-channel (R/G/B) color. The current frame and its optical flow image are first fed into feature extraction modules, which produce the initial spatial and temporal deep features. Secondly, the feature aggregation is performed on initial deep features, where we can obtain the spatiotemporal deep features. Thirdly, by incorporating the boundary cues, which can be acquired by using spatial and temporal cues, we perform saliency prediction on these spatiotemporal deep feature maps and obtain the coarse saliency predictions. Lastly, by incorporating the original information, *i.e.* current frame and optical flow image, we deploy saliency fusion on the prediction results, generating the final saliency map $\mathbf{S}_t$.

For the feature extraction module, we adopt an off-the-shelf deep saliency model, *i.e.* Amulet [45], which is built on VGG-16 model [32]. In our model, the feature extraction module that provides the initial deep features is only a part of Amulet, where we only adopt the output of resolution-based feature combination structure (*i.e.* RFC) and the second layer of VGG-16 (*i.e.* conv1-2) in Amulet to perform feature aggregation and saliency prediction, respectively. Concretely, the current frame $\mathbf{I}_t$ and optical flow image $\mathbf{OP}_t$ are sent to the feature extraction module to obtain the corresponding initial spatial deep features $\{\mathbf{IF}_{t,i}^{S}\}_{i=1}^{5}$ and initial temporal deep features $\{\mathbf{IF}_{t,i}^{T}\}_{i=1}^{5}$, respectively. Notably, the number of channels of these initial deep features are all set to 320, and these initial deep features are with different spatial resolutions, which can be obtained by using the five convolutional blocks in VGG-16.

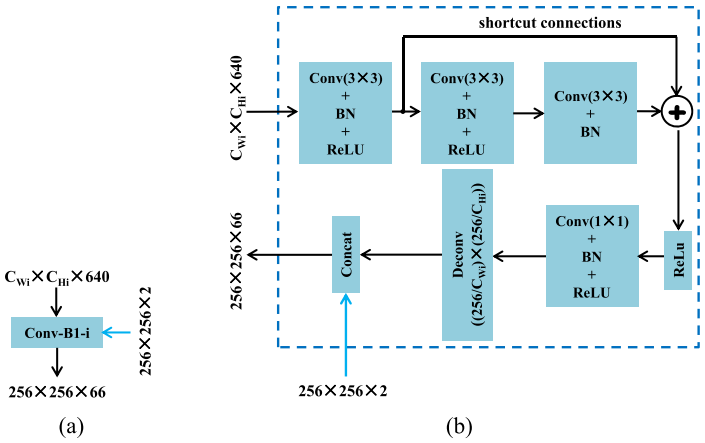

Fig. 2. Illustration of the convolutional block Conv-B1. (a): the thumbnail of Conv-B1, (b): the detailed configuration of Conv-B1, which consists of the convolutional layers with 3×3 kernel size and 1×1 kernel size, the deconvolutional layer with $(256/C_{Wi})\times(256/C_{Hi})$ kernel size, batch normalization layers (BN), and ReLU layers. Besides, in Conv-B1, the "shortcut connections" are also adopted. Here, C_{Wi} and C_{Hi} represent the width and height of input feature maps.

3.2. Feature aggregation

To characterize the salient objects in the video effectively, we perform feature aggregation on the initial spatial and temporal deep features $\{\mathbf{IF}_{t,i}^{S}\}_{i=1}^{L}$ and $\{\mathbf{IF}_{t,i}^{T}\}_{i=1}^{L}$ via concatenation, convolution and recurrent connections, in which spatial information and temporal information can interact with each other. Following this way, we acquire the spatiotemporal deep features $\{\mathbf{F}_{t,i}^{ST}\}_{i=1}^{L}$.

Firstly, we mix the initial spatial and temporal deep features via a simple concatenation operation. Concretely, since the initial deep features are obtained by using five convolutional blocks, which leads to five different resolution features. Therefore, we concatenate the initial spatial deep features and temporal deep features belonging to the same resolution along the channel direction, which is formulated as:

$$\mathbf{IF}_{t,i}^{ST} = \mathbf{Cat}\left(\mathbf{IF}_{t,i}^{S}, \mathbf{IF}_{t,i}^{T}\right), i = 1, \cdots, L, \tag{1}$$

where **Cat** denotes the cross channel concatenation, and $\mathbf{IF}_{t,i}^{ST}$ refers to the initial spatiotemporal deep features with five different spatial resolutions including $256\times256\times640$, $128\times128\times640$, $64\times64\times640$, $32\times32\times640$, and $16\times16\times640$.

Secondly, inspired by the recurrent connections [45] which facilitates the information exchange, we mix $\mathbf{IF}_{t,i}^{ST}$ by using the convolutional blocks Conv-B1-i ($i = 1, \cdots, 5$), which are connected via recurrent connections. Here, each convolutional block is followed by a 3×3 convolutional layer, as shown in Fig. 1. Moreover, to make our model easy to train, we also equip the convolutional blocks with "shortcut connections" [11]. For simplicity, we denote each of these convolutional blocks as Conv-B1-i ($i = 1, \cdots, 5$) shown in Fig. 2(a).

Specifically, in each level i, the convolution block Conv-B1-i takes $\mathbf{IF}_{t,i}^{ST}$ and the high-level aggregation result $\mathbf{F}_{t,i+1}^{ST}$ as input, and produces the new aggregation result. According to the two blue boxes located in the aggregation step shown in Fig. 1 and the convolutional block Conv-B1-i shown in Fig. 2, this process can be defined as:

$$\mathbf{F}_{t,i}^{ST} = \begin{cases} \mathbf{W}_i^a * \mathbf{Cat}\left(f_i^{B1}\left(\mathbf{IF}_{t,i}^{ST}\right), \mathbf{F}_{t,i+1}^{ST}\right) + \boldsymbol{b}_i^a & i < L \\ \mathbf{W}_i^a * f_i^{B1}\left(\mathbf{IF}_{t,i}^{ST}\right) + \boldsymbol{b}_i^a & i = L \end{cases}, \tag{2}$$

where $\mathbf{F}_{t,i}^{ST}$ denotes the spatiotemporal deep feature map. The kernel weights $\mathbf{W}_i^a$ and bias $\boldsymbol{b}_i^a$ refer to the parameters of the 3×3 convolutional layers, *i.e.* the bottom blue boxes in the feature aggregation step shown in Fig. 1. f_i^{B1} denotes the convolutional block Conv-B1-i shown in Fig. 2, which includes convolutional layers, batch normalization layers (BN) [2], an ReLU activation function, concatenation layer, and deconvolutional layers. Specifically, when $i = L$, the convolutional block Conv-B1-i can be denoted as $f_i^{B1}\left(\mathbf{IF}_{t,i}^{ST}\right)$, namely the recurrent connections are removed from the high-level aggregation result. When $i < L$, Conv-B1-i is represented by $\mathbf{Cat}\left(f_i^{B1}\left(\mathbf{IF}_{t,i}^{ST}\right), \mathbf{F}_{t,i+1}^{ST}\right)$.

According to Eq. (2) and Fig. 1, we can see that the initial low-resolution spatiotemporal deep features are recurrently utilized to generate the deep features with high resolution. This facilitates the interaction between spatial information and temporal information effectively, and the generated spatiotemporal deep features can give an effective representation for salient objects in videos.

3.3. Saliency prediction

Based on the obtained spatiotemporal deep feature maps, we perform saliency prediction. Furthermore, inspired by the boundary preserved efforts [45,21], we first generate the spatiotemporal boundary cue by using the initial spatial and temporal deep features,

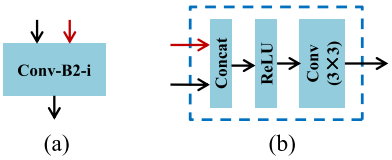

Fig. 3. Illustration of the convolutional block Conv-B2. (a): the thumbnail of Conv-B2, (b): the detailed configuration of Conv-B2, which consists of the 3×3 convolutional layer, concatenation layer and ReLU layer.

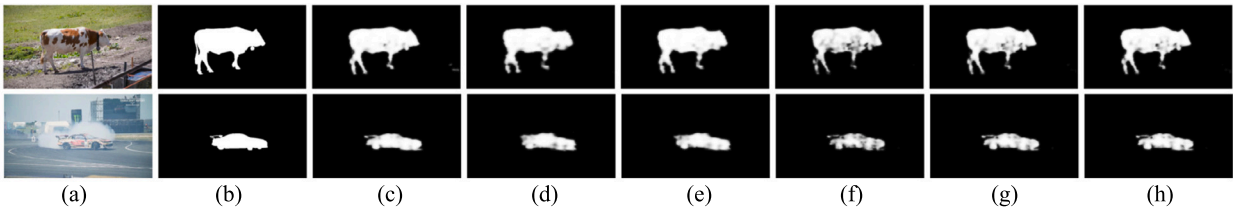

Fig. 4. Intermediate results of our model. (a): Input video frames, (b): binary ground truths, (c): the final saliency maps $\mathbf{S}_t$, coarse saliency maps generated by the saliency prediction step (d): $\mathbf{CS}_{t,1}$, (e): $\mathbf{CS}_{t,2}$, (f): $\mathbf{CS}_{t,3}$, (g): $\mathbf{CS}_{t,4}$, (h): $\mathbf{CS}_{t,5}$.

which are the output of the feature extraction module (*i.e.* conv1-2). Then, we introduce the boundary cue to the prediction step shown in Fig. 1.

Firstly, we apply two 1×1 convolutional layers to conv1-2 layers that correspond to spatial and temporal information, yielding the spatial boundary cue $\mathbf{B}_t^S$ and temporal boundary cue $\mathbf{B}_t^T$, respectively.

Secondly, we concatenate the two boundary cues, namely:

$$\mathbf{B}_t^{ST} = \mathbf{Cat}\left(\mathbf{B}_t^S, \mathbf{B}_t^T\right), \tag{3}$$

where $\mathbf{B}_t^{ST}$ refers to the spatiotemporal boundary cue.

Finally, by incorporating the spatiotemporal boundary cue $\mathbf{B}_t^{ST}$, the saliency prediction deploys the convolutional block Conv-B2 on each spatiotemporal deep feature map $\mathbf{F}_{t,i}^{ST}$, and we can obtain the coarse saliency map $\mathbf{CS}_{t,i}$ shown in Fig. 4(d-h), namely:

$$\mathbf{CS}_{t,i} = \mathbf{W}_i^p * \sigma\left(\mathbf{Cat}\left(\mathbf{B}_t^{ST}, \mathbf{F}_{t,i}^{ST}\right)\right) + \boldsymbol{b}_i^p, \tag{4}$$

where the convolutional block Conv-B2 consists of Conv-B2-1, Conv-B2-2, Conv-B2-3, Conv-B2-4 and Conv-B2-5, as shown in Fig. 1. For simplicity, each one is denoted as Conv-B2-i ($i = 1, \cdots, 5$) shown in Fig. 3(a). According to Fig. 3(b), we can see that Conv-B2-i consists of a concatenation layer, an ReLU activation function, and a 3×3 convolutional layer. $\mathbf{W}_i^p$ and $\boldsymbol{b}_i^p$ denote the kernel weights and the convolutional layer's bias in Conv-B2-i, respectively. Besides, σ denotes the ReLU activation function.

3.4. Saliency fusion

The obtained coarse saliency maps $\{\mathbf{CS}_{t,i}\}_{i=1}^L$, that are computed by using five different resolution spatiotemporal deep feature maps, complement each other [13]. Furthermore, to obtain high quality saliency maps and inspired by [35], we incorporate the original (or initial) information, *i.e.* current frame $\mathbf{I}_t$ and optical flow image $\mathbf{OP}_t$, into the saliency fusion step. In the saliency fusion step, we first concatenate the original information and the obtained coarse saliency maps into a 16-channel image $\mathbf{FOC}_t$, which can be written as:

$$\mathbf{FOC}_t = \mathbf{Cat}\left(\mathbf{F}_t, \mathbf{OP}_t, \{\mathbf{CS}_{t,i}\}_{i=1}^L\right). \tag{5}$$

After that, the 16-channel image $\mathbf{FOC}_t$ is fed into the convolutional block Conv-B3 shown in Fig. 5, namely:

$$\left(\mathbf{S}_t^f, \mathbf{S}_t^b\right) = f^{B3}\left(\mathbf{FOC}_t\right), \tag{6}$$

where f^{B3} denotes the convolutional block Conv-B3, which consists of the 3×3 convolutional layers, batch normalization layers (BN), and ReLU activation function. Here, we should note that the channel number of the 3×3 convolutional layers in Conv-B3 is set to 2. Besides, similar to convolutional block Conv-B1, we also employ the "shortcut connections" to make our model easy to train. In this way, we can sufficiently exploit the complementary information of these coarse saliency maps via a nonlinear manner to obtain the final high-quality saliency map. $\mathbf{S}_t^f$ and $\mathbf{S}_t^b$ mean the probability of each pixel belonging to salient objects or background.

Finally, the final saliency map $\mathbf{S}_t$, as illustrated in Fig. 4(c), is computed in a contrast way, namely:

$$\mathbf{S}_t = \sigma\left(\mathbf{S}_t^f - \mathbf{S}_t^b\right), \tag{7}$$

where σ denotes the ReLU activation function. Besides, according to Fig. 4, it can be found that the coarse saliency maps highlight the mainbody of the salient objects shown in Fig. 4(d-h). Meanwhile, we can also find that the final saliency map $\mathbf{S}_t$ shown in Fig. 4(c)

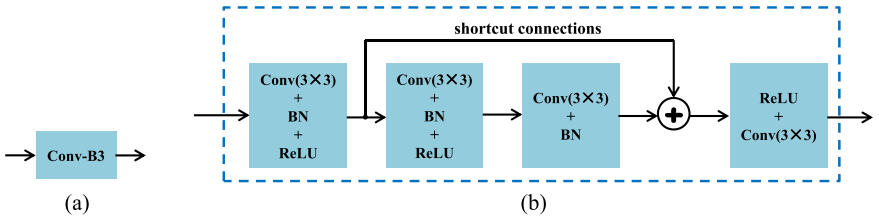

Fig. 5. Illustration of the convolutional block Conv-B3. (a): the thumbnail of Conv-B3, (b): the detailed configuration of Conv-B3, which consists of the 3×3 convolutional layers, batch normalization layers (BN) and ReLU layers. Besides, in Conv-B3, the "shortcut connections" are also adopted.

both highlights the salient object regions more uniformly and shows more precise detail, such as the bell around the neck of cow (top example) and the spoiler on the car (bottom example).

3.5. Model learning

In our model, the three key steps including feature aggregation, saliency prediction and saliency fusion are trained in an end-to-end manner. Given the training set $\mathbb{D}_{train} = \{(\mathbf{I}_n, \mathbf{OP}_n, \mathbf{Y}_n)\}_{n=1}^{N}$ with N training samples, in which $\mathbf{I}_n = \{\mathbf{I}_n^j, j = 1, \cdots, N_p\}$, $\mathbf{OP}_n = \{\mathbf{OP}_n^j, j = 1, \cdots, N_p\}$, and $\mathbf{G}_n = \{\mathbf{G}_n^j, j = 1, \cdots, N_p\}$ represent current frame, optical flow image and binary ground truth with N_p pixels, respectively. $\mathbf{G}_n^j = 1$ denotes that the j-th pixel belongs to salient object, and $\mathbf{G}_n^j = 0$ otherwise. Inspired by the effort in [45], the weighted cross-entropy loss fusion is employed to approach the imbalanced data (*i.e.* salient and non-salient pixels). Here, we drop the subscript n and use $\{\mathbf{I}, \mathbf{OP}\}$ to represent each frame. Thus, the loss function is presented as:

$$\begin{aligned}\mathcal{L}(\mathbf{W}, \boldsymbol{b}) = &-\alpha \sum_{j \in \mathbf{G}_+} \log P\left(\mathbf{G}^j = 1 | \mathbf{I}, \mathbf{OP}; \mathbf{W}, \boldsymbol{b}\right) \\ &-(1-\alpha) \sum_{j \in \mathbf{G}_-} \log P\left(\mathbf{G}^j = 0 | \mathbf{I}, \mathbf{OP}; \mathbf{W}, \boldsymbol{b}\right),\end{aligned} \tag{8}$$

where $\boldsymbol{b}$ and $\mathbf{W}$ denotes convolutional layers' bias and kernel weights, respectively. $\mathbf{G}_+$ and $\mathbf{G}_-$ are pixel sets of salient object regions and background regions, respectively. The notation α denotes the proportion of salient object pixels, namely $\alpha = |\mathbf{G}_+|/|\mathbf{G}_-|$. The expression $P\left(\mathbf{G}^j = 1 | \mathbf{I}, \mathbf{OP}; \mathbf{W}, \boldsymbol{b}\right)$ is the probability of a pixel belonging to salient regions, while the probability of a pixel being background can be denoted as $P\left(\mathbf{G}^j = 0 | \mathbf{I}, \mathbf{OP}; \mathbf{W}, \boldsymbol{b}\right)$.

The entire model is trained in an end-to-end way, and the pre-trained model Amulet [45] is used to initialize the parameters of feature extraction module. The parameters of remaining layers are initialized using "msra" [10]. To minimize the loss function in Eq. (8), we adopt the stochastic gradient descent algorithm (SGD) to train our model, namely:

$$\left(\mathbf{W}^*, \boldsymbol{b}^*\right) = \arg\min \mathcal{L}(\mathbf{W}, \boldsymbol{b}), \tag{9}$$

where $\boldsymbol{b}^*$ and $\mathbf{W}^*$ denotes the optimal solution of convolutional layers' bias and kernel weights, respectively. Here, some parameters of SGD such as momentum, iterations, base learning rate, mini-batch size and weight decay are set to 0.9, 22000, 10^{-8}, 32, and 0.0001, respectively. Besides, when the training loss is convergence, the learning rate is declined by 0.1. During the training stage of our model, we don't adopt the validation datasets to prevent over-fitting. Meanwhile, after the training loss is convergence, the performance of the generated models is very close to each other. Therefore, we randomly select a model (12000 training iterations) as the reported model.

4. Experimental results

In this part, the video datasets and implementation details are first introduced. Then, the comprehensive qualitative and quantitative comparison results are provided in Section 4.1. Some validation tests are performed in Section 4.2. After that, we present some failure cases in Section 4.3. Finally, the computation cost of our model is presented in Section 4.4.

For training and test our model, three video datasets including DAVIS [30], UVSD [27] and SegTrackV2 [23] are employed in this paper. For each frame in each video of these video datasets, they provide pixel-wise annotation. Similar to [37], the whole SegTrackV2 dataset and the training set of DAVIS dataset are collected as training data. For test, the performance of the proposed model is evaluated on the UVSD dataset and the test set of DAVIS dataset. Besides, we also adopt some augmentation operations to obtain good performance without over-fitting, where we perform rotation ($90°, 180°$, and $270°$) and mirror reflection on each image of the training set successively. In this way, compared to the original training set, our training data can be augmented by 8 times.

4.1. Performance comparison

To evaluate the performance of the proposed STI-Net, we perform qualitative and quantitative comparisons against the state-of-the-art saliency models including SGSP [27], GD [38], SFCD [5], VSFCN [37], PDB [33], SSAV [9], MC [46] and Amulet [45] on the whole UVSD dataset and the test set of DAVIS dataset. The former four models are state-of-the-art video saliency models

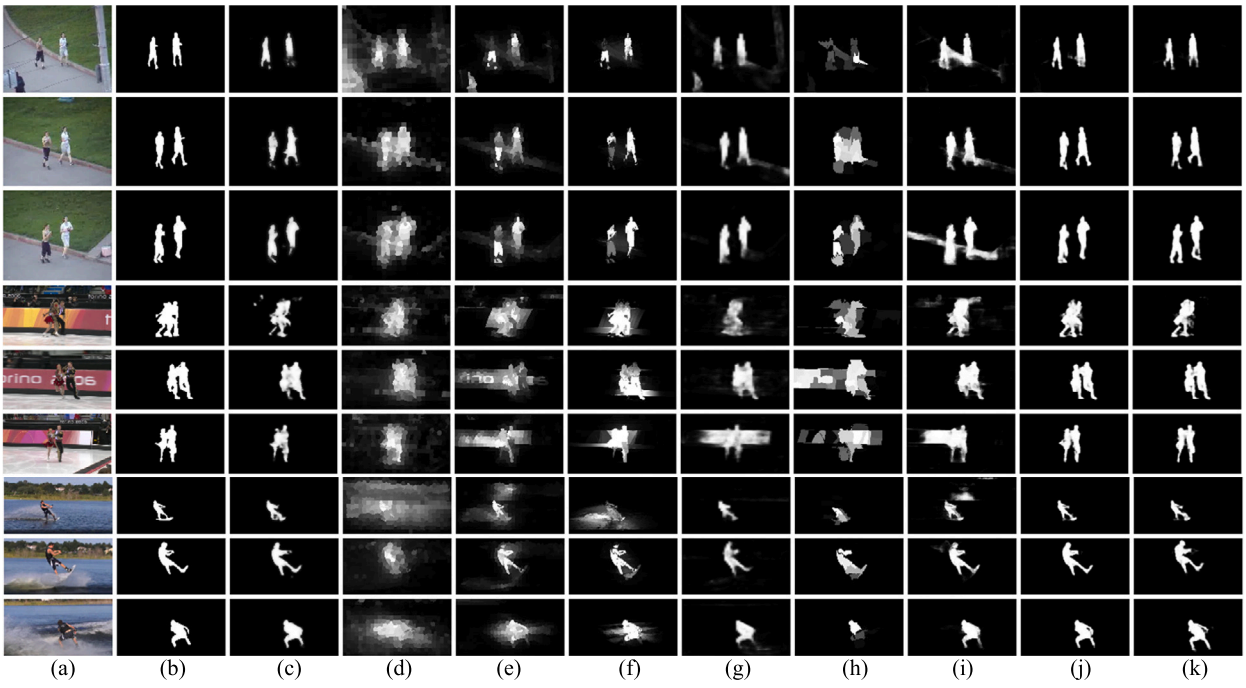

Fig. 6. Visualization comparison of different saliency models on some videos in the UVSD dataset. (a): Input video frames, (b): binary ground truths, (c): Ours, (d): SGSP [27], (e): GD [38], (f): SFCD [5], (g): VSFCN [37], (h): MC [46], (i): Amulet [45], (j): PDB [33], (k): SSAV [9].

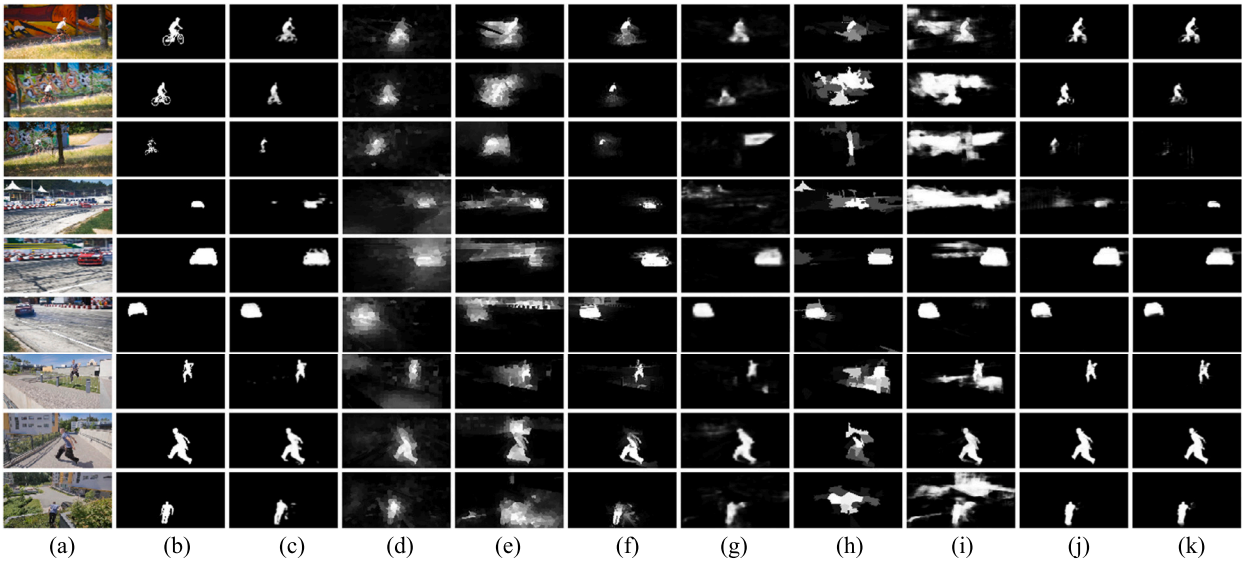

Fig. 7. Visualization comparison of different saliency models on some videos in the DAVIS dataset. (a): Input video frames, (b): binary ground truths, (c): Ours, (d): SGSP [27], (e): GD [38], (f): SFCD [5], (g): VSFCN [37], (h): MC [46], (i): Amulet [45], (j): PDB [33], (k): SSAV [9].

while the latter two are deep learning based image saliency models. For a fair comparison, the source codes of SGSP, GD, SFCD, VSFCN, MC and Amulet are directly provided by their authors, and the saliency maps generated by different models are normalized into the same resolution as original videos with pixel values ranging from 0 to 255. In the following, quantitative and qualitative comparisons are performed successively.

4.1.1. Qualitative evaluation

Figs. 6 and 7 show qualitative comparisons for our model and the state-of-the-art saliency models on UVSD and DAVIS, respectively. According to Figs. 6(a) and 7(a), it can be seen that these videos exhibit various complex scenarios such as shape

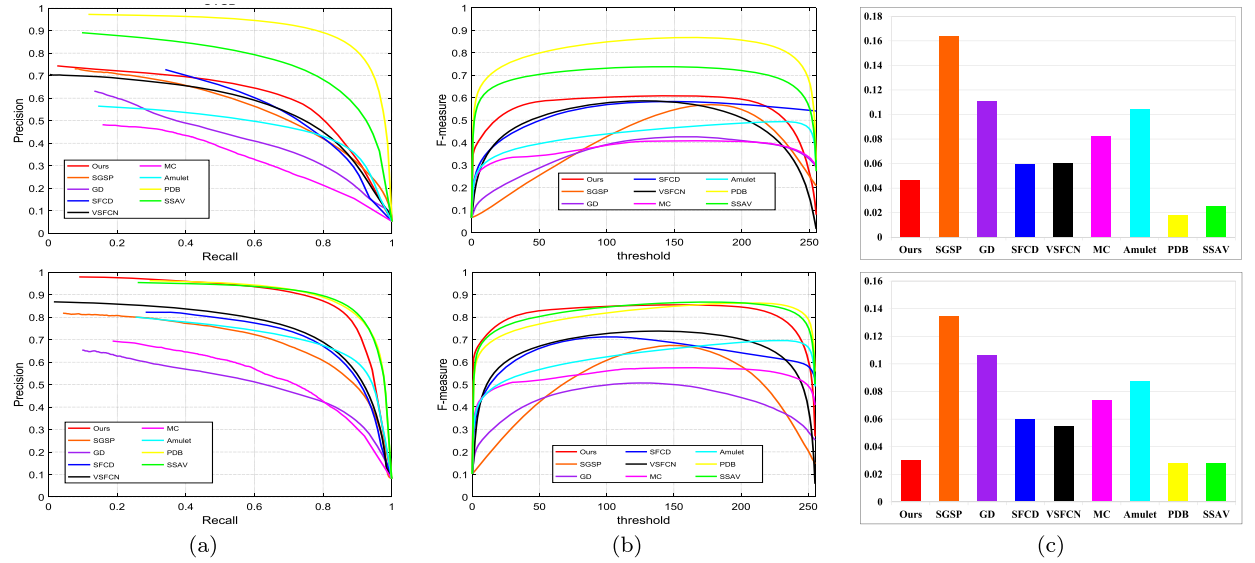

Fig. 8. Quantitative results of different saliency models: (a) PR curves, (b) F-measure curves, and (c) MAE values. From top to down, each row presents the results of compared models on UVSD and DAVIS datasets, respectively.

complexity, occlusion and non-rigid deformation, motion blur and so on. Therefore, it is a challenging task for popping-out salient objects in videos. The results of video saliency models including SGSP, GD and SFCD are shown in Figs. 6(d, e, f) and 7(d, e, f). We can find that they highlight the main parts of salient objects. However, they mistakenly highlight some background regions around the salient objects. For the deep learning based image saliency models such as MC and Amulet shown in Figs. 6(h, i) and 7(h, i), we can see that more promising performances exhibit when compared to the aforementioned video saliency models. Unfortunately, due to the lack of temporal information, the two deep learning based image saliency models often highlight some background regions incorrectly. Further, for deep learning based video saliency model VSFCN shown in Figs. 6(g) and 7(g), we can see that some background regions are also popped-out in the videos which are with fast motion and non-rigid deformation such as the bottom example in Fig. 6 and the middle example in Fig. 7. The two top-level deep learning-based models PDB and SSAV achieve satisfactory results shown in Figs. 6(j, k) and 7(j, k). Compared to PDB and SSAV, the results of our model shown in Figs. 6(c) and 7(c) achieve a comparable performance on all videos even for the cluttered background (the top example in Fig. 7) and fast motion with non-rigid deformation (the middle example in Fig. 6). We can find that our saliency maps can not only give a complete prediction for salient objects, but also are with precise boundary details. Overall, all the results shown in Fig. 6 and Fig. 7 clearly demonstrate the effectiveness and superiority of our model (STI-Net), which generates high-quality saliency maps.

4.1.2. Quantitative evaluation

We employ three standard metrics, precision-recall (PR) curve, F-measure curve, and mean absolute error (MAE) value, to quantitatively evaluate the performances of different saliency models. Here, the default settings of these three metrics in prior works [7,37] are adopted.

The PR curves of all saliency models are presented in Fig. 8(a). Except the top-level models including PDB and SSAV, our model achieves the best performance on both datasets. Especially, on DAVIS dataset, our model can pop-out the salient objects in videos precisely. The F-measure curves are shown in Fig. 8(b), in which our model performs best when compared with other saliency models including SGSP, GD, SFCD, VSFCN, MC, and Amulet. Similarly, the MAE values presented in Fig. 8(c) also show the same conclusion that our model gains the lowest MAE value when compared with the six saliency models including SGSP, GD, SFCD, VSFCN, MC, and Amulet. Here, we should note that the performance of our model is lower than the two top-level models including PDB and SSAV in terms of PR curves, F-measure curves, and MAE values, as shown in the top row of Fig. 8. Interestingly, as shown in the bottom row of Fig. 8, our model presents a comparable performance when compared to PDB and SSAV on the DAVIS dataset. This also indicates that our model is suitable for the DAVIS dataset. Overall, the PR curves, F-measure curves and MAE values shown in Fig. 8 convincingly demonstrate the effectiveness of the proposed spatiotemporal integration network (STI-Net).

Thus, according to the qualitative and quantitative comparison results show in Figs. 6, 7 and 8, we can firmly demonstrate the effectiveness of our model, and the reason behind this can be attributed to the sufficient exploitation of spatial and temporal information via feature aggregation, saliency prediction and saliency fusion, which gives a powerful characterization for the salient objects in the complex video scenarios. Besides, we also insert the boundary cues into the saliency prediction step, which guarantees the accuracy of the generated saliency maps. Based on this, our model characterizes salient objects in videos effectively, and thus can perform very well on some complex scenes.

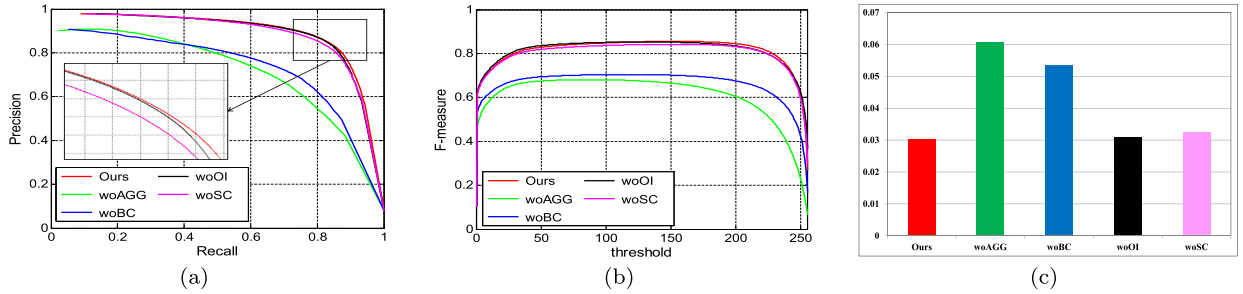

Fig. 9. Quantitative comparison for the model analysis on DAVIS dataset: (a) PR curves, (b) F-measure curves, and (c) MAE values.

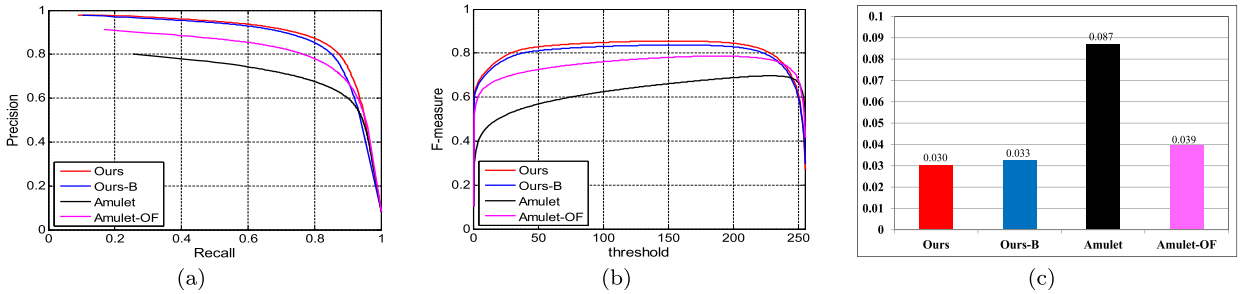

Fig. 10. Quantitative comparisons for the differences between our model and Amulet on DAVIS dataset: (a) PR curves, (b) F-measure curves, and (c) MAE values.

4.2. Validation of the proposed model

In this subsection, we first study the contribution of each component (namely feature aggregation, boundary cues, original information and "shortcut connections") in our model. Then, we analyze the differences between our model and Amulet [45]. Thirdly, we deeply explore the effect of boundary information and temporal information, respectively. Lastly, we validate the effects of five-level feature blocks in our network. Firstly, to validate the contributions of feature aggregation, boundary cues, original information (current frame and optical flow image), and "shortcut connections" to the final performance, we performed the ablation studies as shown in Fig. 9. Our model without "shortcut connections", original information, boundary cues and feature aggregation are denoted as "woSC", "woOI", "woBC" and "woAGG", respectively. Here, "woOI" means that we don't incorporate the original information, namely current frame and its optical flow image, in the saliency fusion step. "woBC" denotes that we don't adopt the boundary cues obtained from conv1-2 in the saliency prediction step. "woAGG" refers to that our model treats current frame and optical flow image separately, and there is no interaction between them. Specifically, firstly, our model acquires the best performance when compared with woAGG, woBC, and woSC in terms of PR curves, F-measure curves, and average MAE values. In particular, our model outperforms woAGG and woBC with a large margin, which clearly demonstrates the effectiveness of interaction in the feature aggregation step and the usefulness of boundary cues in the saliency prediction step. Besides, our model also performs better than woSC, which indicates the necessity of "shortcut connections". Secondly, woOI achieves competitive performance with our model in terms of F-measure curves and averages MAE values as shown in Figs. 9(b) and (c). Nonetheless, we can also find that in terms of the PR curves shown in Fig. 9(a), our model outperforms woOI with a small margin. Therefore, in terms of all three metrics, our model performs better than woOI, which indicates the effectiveness of the original information adopted in the saliency fusion step. According to the above studies, we see that each component in our model is beneficial to obtain satisfactory performance, and thus the rationality of our model has been verified.

Secondly, our model is constructed based on Amulet [45], therefore it is essential to analyze the differences between our model and Amulet. Some comparisons are performed in this part, as shown in Fig. 10. We make a concatenation for the temporal data (*i.e.* optical flow image) and the current frame as the input of Amulet. The Amulet is retrained with the same training data as ours, and the retrained Amulet is denoted as "Amulet-OF". Besides, for the boundary information, the differences between our model and Amulet lie in two aspects, namely the generation and usage of boundary information. The boundary information is generated by using motion information and appearance information in our model, as shown in Fig. 1. Differently, Amulet only utilizes the appearance information. As for the usage of boundary information, Amulet is performed in a summation way, where the boundary information is added to the saliency prediction. In contrast, our model adopts the concatenation way. Thus, we make a modification for our model, namely adopting the same way as Amulet for the usage of boundary information. This new version of model is denoted as "Ours-B". We can see from Fig. 10 that our model (*i.e.* "Ours") performs best when compared with Amulet, Ours-B and Amulet-OF in terms of all three metrics. This clear validates the differences between Amulet and our model, where our model provides the sufficient exploitation of temporal information and the effective design using boundary information.

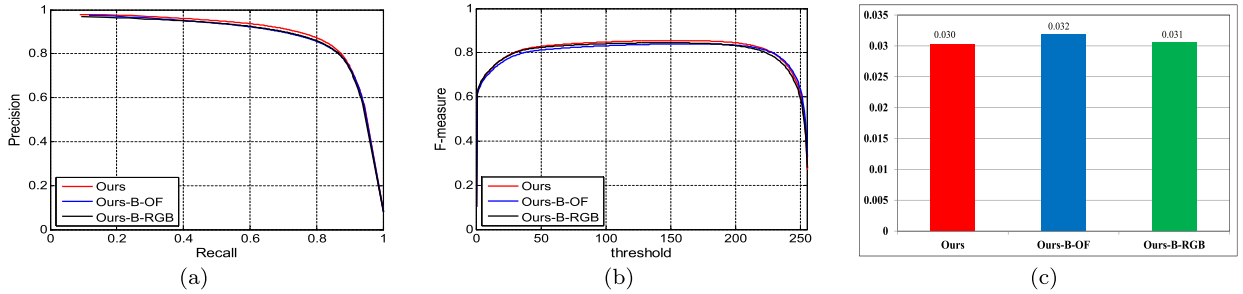

Fig. 11. Quantitative results of boundary information's effect on DAVIS dataset: (a) PR curves, (b) F-measure curves, and (c) MAE values.

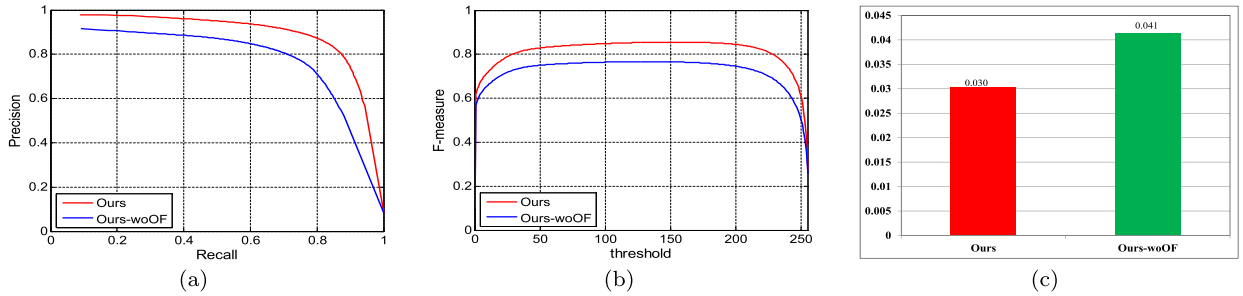

Fig. 12. Quantitative results of temporal information's effect on DAVIS dataset: (a) PR curves, (b) F-measure curves, and (c) MAE values.

Thirdly, to further analyze the effects of boundary information, we perform some comparisons among different versions of our model. In our model, if the boundary information is originated from the current frame only, we denote it as "Ours-B-RGB". If the boundary information is originated from the optical flow image only, we denote it as "Ours-B-OF". Thus, according to the experimental results shown in Fig. 11, it can be seen that our model achieves the best performance when compared with Ours-B-OF and Ours-B-RGB in terms of PR curves, F-measure curves, and average MAE values. Furthermore, by comparing with our model and woBC shown in Fig. 9, we can demonstrate our model's effectiveness and the rationality of our boundary design.

Besides, as for the temporal information is the key point for video saliency detection, we conduct the performance evaluation for our model and our model without the optical flow input branch, which is denoted as "Ours-woOF". The corresponding results are shown in Fig. 12. It can be seen that our model performs better than Ours-woOF with a significant margin in terms of PR curves, F-measure curves, and average MAE values. This clearly indicates the effect of temporal information, which is very important for video saliency detection. Furthermore, we also conduct a qualitative comparison, as shown in Fig. 14, on two situations: 1) the salient object is moving between the two frames so that it will be highlighted by the optical flow visualization used in our implementation; 2) the salient object is not the one that is moving between the two frames so that it will not be highlighted by the optical flow visualization. The top two examples belong to the first situation, *i.e.* the salient object is highlighted in the optical flow visualization. It can be seen that the obtained saliency map shown in Fig. 14(d) is very similar to the ground truth shown in Fig. 14(b). As for the bottom two examples, which correspond to the second situation, where the salient object moves slowly or is stationary. It can be found that the optical flow images of the bottom two examples do not highlight the goat and the mallard. But fortunately, the corresponding saliency maps shown in Fig. 14(d) are satisfactory when compared with the ground truths shown in Fig. 14(b). The reason behind this lies in the sufficient usage of spatial information and temporal information in our model, in which they complement each other. Therefore, we can make a conclusion that the proposed model is robust and effective to videos with both scenes including stationary objects and slowly moving objects.

Lastly, in our model, the features originated from the five blocks are all utilized to obtain the saliency map. Thus, in this part, we perform some comparisons to show the performance gain by adding features from each block, as shown in Fig. 13. "Our-Sn" means that our model takes the integrated features originated from blocks $\{1, 2, ..., n-1, n\}$. For example, Our-S3 denotes that our model takes the integrated features from block1, block2 and block3. According to Fig. 13, our model, which takes all the integrated features from all blocks, achieves the best performance in terms of all the aforementioned three metrics. This clearly verifies the effectiveness of our model, and also indicates the rationality of the model design.

4.3. Failure examples and analysis

Although our model is able to achieve good results in most cases, it is still difficult for our model to efficiently deal with some challenging scenarios. Two examples are shown in Fig. 15. The first case presented in Fig. 15(a) show that a gymnast with black trousers (*i.e.* the salient object) is performing on a pommel horse. Note that, the audiences exhibit cluttered background in this video, and the heterogeneous salient object also shows non-rigid deformations. Our model only highlights part of salient object, and some

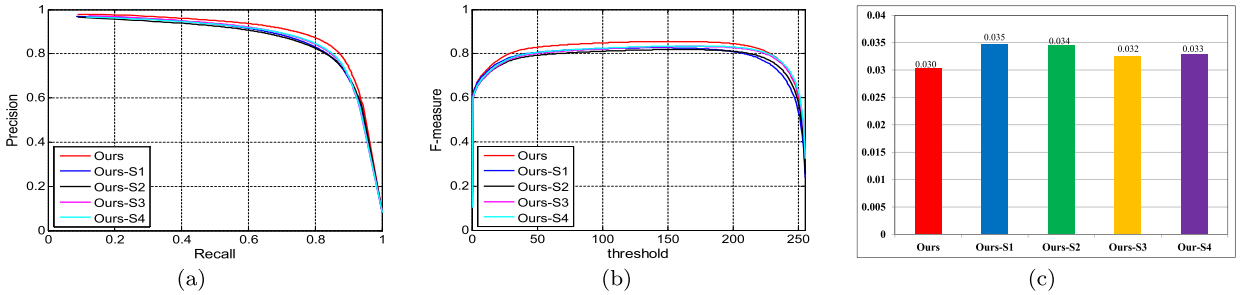

Fig. 13. Quantitative comparisons for the effects of five-level blocks in our model on DAVIS dataset: (a) PR curves, (b) F-measure curves, and (c) MAE values.

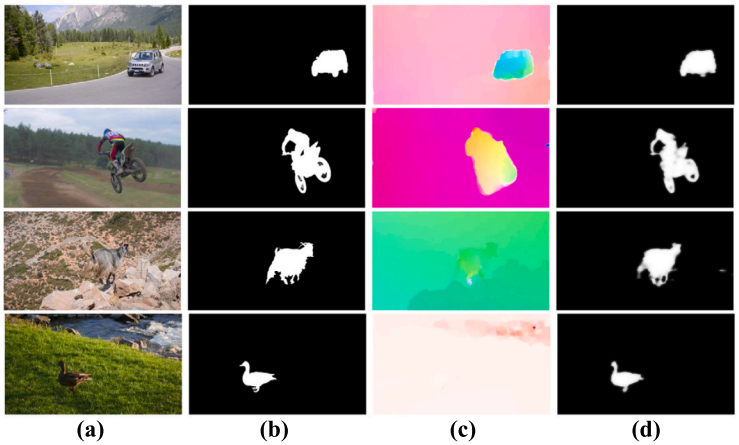

Fig. 14. Qualitative results about the effect of temporal information. (a): Input video frames, (b): binary ground truths, (c): optical flow images, (d): saliency maps generated by our model.

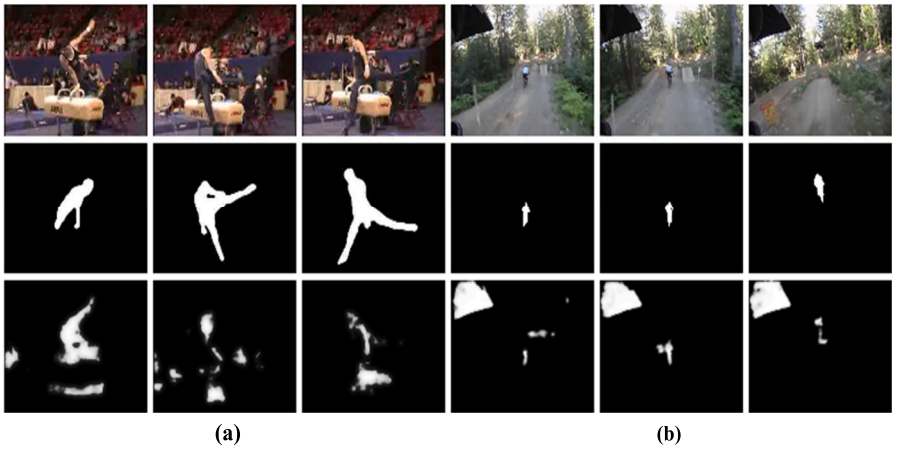

Fig. 15. Failure examples. (a): case 1, (b): case 2. Here, the top row is the input video frames, the second row is the ground truths, and the third row is our results.

background regions are highlighted falsely. The reason behind this lies in that the extracted spatial and temporal information are inaccurate under such complicated scenes, *e.g.* low contrast background with confusing visual appearance, cluttered background, and non-rigid deformation. The second case in Fig. 15(b) present the situation where a camera is with severe jitter, and the cyclist is the salient object to be detected. It can be seen that our model fails to highlight the salient object uniformly, and the background is not effectively suppressed. It can be attributed to two aspects including the inaccurate optical flow images and the small salient object, which leads to the poor characterization of temporal information.

4.4. Computation cost

We present and analyze the computation cost of our model in this part. Our model is implemented by Matlab R2014a platform together with the Caffe toolbox. The configurations of our deployed PC machine include an i7-4790K CPU (32 GB RAM) and a single NVIDIA Titan XP GPU (with 12 GB memory). The training process costs nearly 48 hours under this configuration. In the test phase, the optical flow computation takes 33.542 seconds for a video frame with the video resolution of 480×854. The generation of saliency map only needs 0.111 seconds. Therefore, it can be easily found that the computation of optical flow is the bottleneck of our model in terms of computation cost. To accelerate the runtime of our model, we can adopt the recent deep learning based optical flow computation method [15], which takes 5.45 seconds for a 480×854 video frame.

5. Conclusion

This paper presented a novel end-to-end spatiotemporal integration network (STI-Net) for popping-out salient objects in videos, which includes three key steps including feature aggregation, saliency prediction, and saliency fusion. The key advantage of our model lies in the sufficient utilization of spatial and temporal information. Firstly, in the feature aggregation step, the interaction between spatial and temporal deep features provides an effective description for salient objects in videos. Secondly, in the saliency prediction step, the incorporation of the boundary cues originating from spatial and temporal domains is conductive to generate accurate prediction results. Finally, in the saliency fusion step, taking into account original information, such as the current frame and optical flow image, allows for the correction of errors in the prediction results. Extensive experiments are carried out on two challenging video datasets, and the experimental results show that the proposed spatiotemporal saliency network presents a comparable performance when compared with the state-of-the-art saliency models, demonstrating the effectiveness of our model. In future work, we would deploy the efficient optical flow computation method into the entire network in an effective way and promote the advancement of video saliency detection.

CRediT authorship contribution statement

Xiaofei Zhou: Conceptualization, Writing – original draft. **Weipeng Cao:** Funding acquisition, Investigation, Writing – review & editing. **Hanxiao Gao:** Data curation, Validation. **Zhong Ming:** Formal analysis, Software, Supervision. **Jiyong Zhang:** Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (61901145, 62106150), in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grants GK229909299001-009, and in part by the Hangzhou Dianzi University (HDU) and the China Electronics Corporation DATA (CECDATA) Joint Research Center of Big Data Technologies under Grant KYH063120009.

References

[1] M.A. Alsmirat, Y. Jararweh, I. Obaidat, B.B. Gupta, Automated wireless video surveillance: an evaluation framework, J. Real-Time Image Process. 13 (3) (2017) 527–546.
[2] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[3] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M.J. Black, R. Szeliski, A database and evaluation methodology for optical flow, Int. J. Comput. Vis. 92 (1) (2011) 1–31.
[4] T. Brox, J. Malik, Large displacement optical flow: descriptor matching in variational motion estimation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 500–513.
[5] C. Chen, S. Li, Y. Wang, H. Qin, A. Hao, Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion, IEEE Trans. Image Process. 26 (7) (2017) 3156–3170.
[6] P. Chen, J. Lai, G. Wang, H. Zhou, Confidence-guided adaptive gate and dual differential enhancement for video salient object detection, in: IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.
[7] M.M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S.M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[8] C. Choi, T. Wang, C. Esposito, B.B. Gupta, K. Lee, Sensored semantic annotation for traffic control based on knowledge inference in video, IEEE Sens. J. 21 (10) (2021) 11758–11768.

[9] D.P. Fan, W. Wang, M.M. Cheng, J. Shen, Shifting more attention to video salient object detection, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2019, pp. 8554–8564.
[10] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: International Conference on Computer Vision, ICCV, IEEE, 2015, pp. 1026–1034.
[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2016, pp. 770–778.
[12] M.S. Hossain, G. Muhammad, W. Abdul, B. Song, B.B. Gupta, Cloud-assisted secure video transmission and sharing framework for smart cities, Future Gener. Comput. Syst. 83 (2018) 596–606.
[13] Q. Hou, M.M. Cheng, X. Hu, A. Borji, Z. Tu, P.H. Torr, Deeply supervised salient object detection with short connections, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2017, pp. 3203–3212.
[14] C.R. Huang, Y.J. Chang, Z.X. Yang, Y.Y. Lin, Video saliency map detection by dominant camera motion removal, IEEE Trans. Circuits Syst. Video Technol. 24 (8) (2014) 1336–1349.
[15] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: evolution of optical flow estimation with deep networks, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2017, pp. 1647–1655.
[16] L. Itti, P. Baldi, A principled approach to detecting surprising events in video, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2005, pp. 631–637.
[17] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[18] M. Jian, J. Wang, H. Yu, G.G. Wang, Integrating object proposal with attention networks for video saliency detection, Inf. Sci. 576 (2021) 819–830.
[19] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: a discriminative regional feature integration approach, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2013, pp. 2083–2090.
[20] Y. Jiao, X. Wang, Y.C. Chou, S. Yang, G.P. Ji, R. Zhu, G. Gao, Guidance and teaching network for video salient object detection, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 2199–2203.
[21] Y.Y. Ke, T. Tsubono, Recursive contour-saliency blending network for accurate salient object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2940–2950.
[22] Z. Khan, B. Latif, J. Kim, H.K. Kim, M. Jeon, Densebert4ret: deep bi-modal for image retrieval, Inf. Sci. 612 (2022) 1171–1186.
[23] F. Li, T. Kim, A. Humayun, D. Tsai, J.M. Rehg, Video segmentation by tracking many figure-ground segments, in: International Conference on Computer Vision, ICCV, IEEE, 2013, pp. 2192–2199.
[24] Y. Li, S. Li, C. Chen, A. Hao, H. Qin, A plug-and-play scheme to adapt image saliency deep model for video data, IEEE Trans. Circuits Syst. Video Technol. 31 (6) (2020) 2315–2327.
[25] Y. Li, G. Liu, Q. Liu, Y. Sun, S. Chen, Moving object detection via segmentation and saliency constrained rpca, Neurocomputing 323 (2019) 352–362.
[26] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[27] Z. Liu, J. Li, L. Ye, G. Sun, L. Shen, Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation, IEEE Trans. Circuits Syst. Video Technol. 27 (12) (2017) 2527–2542.
[28] Z. Liu, W. Zou, O. Le Meur, Saliency tree: a novel saliency detection framework, IEEE Trans. Image Process. 23 (5) (2014) 1937–1952.
[29] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.
[30] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2016, pp. 724–732.
[31] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, M. Jagersand, U2-net: going deeper with nested u-structure for salient object detection, Pattern Recognit. 106 (2020) 107404.
[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015, pp. 1–14.
[33] H. Song, W. Wang, S. Zhao, J. Shen, K. Lam, Pyramid dilated deeper convlstm for video salient object detection, in: European Conference on Computer Vision, ECCV, Springer, 2018, pp. 715–731.
[34] S. Song, Z. Jia, J. Yang, N. Kasabov, Salient detection via the fusion of background-based and multiscale frequency-domain features, Inf. Sci. 618 (2022) 53–71.
[35] Y. Tang, X. Wu, Saliency detection via combining region-level and pixel-level predictions with cnns, in: European Conference on Computer Vision, ECCV, Springer, 2016, pp. 809–825.
[36] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE Trans. Image Process. 24 (11) (2015) 4185–4196.
[37] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Trans. Image Process. 27 (1) (2018) 38–49.
[38] W. Wang, J. Shen, R. Yang, F. Porikli, Saliency-aware video object segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2018) 20–33.
[39] H. Wen, X. Zhou, Y. Sun, J. Zhang, C. Yan, Deep fusion based video saliency detection, J. Vis. Commun. Image Represent. 62 (2019) 279–285.
[40] Z. Wu, L. Su, Q. Huang, Decomposition and completion network for salient object detection, IEEE Trans. Image Process. 30 (2021) 6226–6239.
[41] S. Xingjian, Z. Chen, H. Wang, D.Y. Yeung, W.K. Wong, W.c. Woo, Convolutional lstm network: a machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems, 2015, pp. 802–810.
[42] M. Xu, P. Fu, B. Liu, J. Li, Multi-stream attention-aware graph convolution network for video salient object detection, IEEE Trans. Image Process. 30 (2021) 4183–4197.
[43] X. Xu, W. Liu, L. Yu, Trajectory prediction for heterogeneous traffic-agents using knowledge correction data-driven model, Inf. Sci. 608 (2022) 375–391.
[44] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, Z. Luo, Dynamic context-sensitive filtering network for video salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2021, pp. 1553–1563.
[45] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: aggregating multi-level convolutional features for salient object detection, in: International Conference on Computer Vision, ICCV, IEEE, 2017, pp. 202–211.
[46] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2015, pp. 1265–1274.
[47] X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: a simple gated network for salient object detection, in: European Conference on Computer Vision, 2020, pp. 35–51.
[48] J. Zhou, J. Gan, W. Gao, A. Liang, Image retrieval based on aggregated deep features weighted by regional significance and channel sensitivity, Inf. Sci. 577 (2021) 69–80.
[49] X. Zhou, Z. Liu, C. Gong, L. Wei, Improving video saliency detection via localized estimation and spatiotemporal refinement, IEEE Trans. Multimed. 20 (11) (2018) 2993–3007.
[50] X. Zhou, Z. Liu, K. Li, G. Sun, Video saliency detection via bagging-based prediction and spatiotemporal propagation, J. Vis. Commun. Image Represent. 51 (2018) 131–143.