Multimedia Tools and Applications (2024) 83:37589–37602
https://doi.org/10.1007/s11042-023-17128-5

Video saliency detection via combining temporal difference and pixel gradient

Xiangwei Lu[1] · Muwei Jian[1,2] · Rui Wang[1] · Xiangyu Liu[1] · Peiguang Lin[1] · Hui Yu[3]

Received: 7 August 2022 / Revised: 24 August 2023 / Accepted: 15 September 2023 /
Published online: 2 October 2023

Abstract
Even though temporal information matters for the quality of video saliency detection, many problems still arise/emerge in present network frameworks, such as bad performance in time-space coherence and edge continuity. In order to solve these problems, this paper proposes a full convolutional neural network, which integrates temporal differential and pixel gradient to fine tune the edges of salient targets. Considering the features of neighboring frames are highly relevant because of their proximity in location, a co-attention mechanism is used to put pixel-wise weight on the saliency probability map after features extraction with multi-scale pooling so that attention can be paid on both the edge and central of images. And the changes of pixel gradients of original images are used to recursively improve the continuity of target edges and details of central areas. In addition, residual networks are utilized to integrate information between modules, ensuring stable connections between the backbone network and modules and propagation of pixel gradient changes. In addition, a self-adjustment strategy for loss functions is presented to solve the problem of overfitting in experiments. The method presented in the paper has been tested with three available public datasets and its effectiveness has been proved after comparing with 6 other typically stat-of-the-art methods.

Keywords Video saliency detection · Temporal difference · Pixels gradient · Edge refinement · Co-Attention

1 Introduction

Video saliency detection aims to recognize interesting zones in dynamic scenes by simulating the attention mechanism of human's eyes. Compared with detecting salient objects in images, video saliency object detection encounters new challenges posed by continuous

✉ Muwei Jian
jianmuweihk@163.com

✉ Hui Yu
Hui.Yu@port.ac.uk

[1] School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China

[2] School of Information Science and Engineering, Linyi University, Linyi, China

[3] School of Creative Technologies, University of Portsmouth, Portsmouth, UK

movement clues and abundant inter-frame information. Differential features of consecutive frames can be used to ensure a better performance in background suppression and edge detection. Video saliency detection has been widely applied in video compression [1], video target tracking [2], video quality assessment [3], video summarization [4], and scene understanding [5–9], image understanding [26–32], action/object recognition [33, 34], etc.

Prevailing video saliency detection methods generally provide macroscopic perspectives and use optical flow and LSTM convolution models to extract temporal characteristics. However, these methods often overlook some interframe details when extracting and integrating features and result in poor continuity of differential features between consecutive frames.

To solve the above-mentioned problem, this paper proposes a static feature extraction network with several static saliency networks working in parallel to extract the information features of video frames and obtain initial temporal information. Interframe differential information between consecutive frames is used to set up a co-attention mechanism. As designed, the co-attention edge module and central module integrate interframe relationships. And the temporal relationships between differential information of consecutive frames are used in assisting learning the temporal and spatial continuity of interframe movements. The detection results have been obviously improved by combining the co-attention module and the pixel gradient-based refinement module. At the same time, we also apply recurrent multi module training optimization and the loss function self-adjustment strategy to further modify the network modal.

The main contributions of the paper are as follows:

(1) A fully convolutional neural network integrating temporal difference and pixel gradient has been designed, which effectively solves the poor performance of temporal spatial coherence and edge continuity.
(2) A co-attention mechanism is set up to integrate temporal differentials to solve the problem of present network models lacking good learning of information continuity;
(3) A method for optimizing image details is proposed based on pixel gradient to rectify the details of salient objects and improve edge continuity and details of the central areas of targets;
(4) A self-adjustment strategy for the loss functions is used to improve the stability of the network structure to avoid overfitting.
(5) Our model has tested on three public datasets, and the experimental results prove its effectiveness.

2 Related work

2.1 Image saliency object detection

Image saliency object detection can be realized by two network models supported by different attention mechanisms: top-down and bottom-up models. The top-down method deals with images under the control of human's consciousness. Yang et al. [10] developed a saliency model combining conditional randomness and visual dictionary, which learned conditional random fields by way of characteristic self-adjustment; and as CRF are output layers, it can learn the dictionary under structural supervision. Gao et al. [11] proposed a bottom-up saliency detector in line with the statistics of characteristics of natural images. Cheng et al. [12] proposed a salient object extraction algorithm based on regional contrast degree, which can calculate both the global contrast difference and spatial weighting consistency. In [35], Hu et al. proposed a novel co saliency detection framework that transforms each image into a semantic superpixel cluster through segmentation, and generates images of different sizes for each input image using the VGG-16 model. The model also explores a new Graph Convolutional Network method to fine-tune multi-scale features, capturing common information between all scale features and private or complementary information for each scale feature. Zhang et al. [37] developed a new deep network framework, DeepACG, which uses Gromov Wasserstein distance to establish dense 4D correlated volumes for all image pixel pairs. This model also designs a semantic aware Common Attention Module (SCAM), which recognizes the semantic classes of foreground collaborative objects and modulates this information into deep representations to achieve the localization of relevant pixels, and the model also utilizes a Contrast Edge Enhancement Module to capture richer contexts and preserve fine-grained spatial information.

2.2 Video saliency object detection

Compared with saliency object detection of images, video saliency detection is more challenging because of involving more complicated detection situations and temporal information. As a response to the problem, Yu et al. [13] proposed a video saliency incorporating spatiotemporal cues and uncertainty weighting, Wang et al. [14] proposed a spatiotemporal saliency detection for video sequences based on the gradient flow field and energy optimization. In addition to the problem in utilization of spatiotemporal information, a lack of massive annotated video datasets has always been a problem. Wang et al. [15] proposed a data enhancement technology which can simulate video training data by using available annotated image data so that the network can learn different salient information and prevent overfitting training videos which are limited in capacity. In [36], Wang et al. proposed a key saliency object re enhancement method based on bottom-up weighted features and top-down semantic knowledge to better simulate human visual attention mechanism in dynamic scenes. This method effectively eliminates noise and redundancy, and provides accurate local spatiotemporal features for saliency detection, further improving the accuracy of detection.

Unlike existing image or video saliency detection methods, this paper presented a new method based on an image detection framework. This method combines the temporal differences of video frames and image saliency detection framework and can capture and utilize the gradient changes of pixels to improve overall visual effects of salient objects.

3 Proposed method

In this paper, we propose a new method to improve the edge continuity and central details of video saliency detection. Our research is composed of three parts: the joint static feature extraction network (SFN), temporal difference co-attention model (TCM), and pixel gradient-based refinement processing module (GRM). Figure 1 is a diagram of the method. SFN aims to extract temporal features on the basis of an image saliency detection method. After SFN extraction of information, TCM integrates temporal difference information via a co-attention mechanism, enhance spatiotemporal consistency between video consecutive frames, and draw attention to both edge and central areas. Then GRM improves details in the edge and central

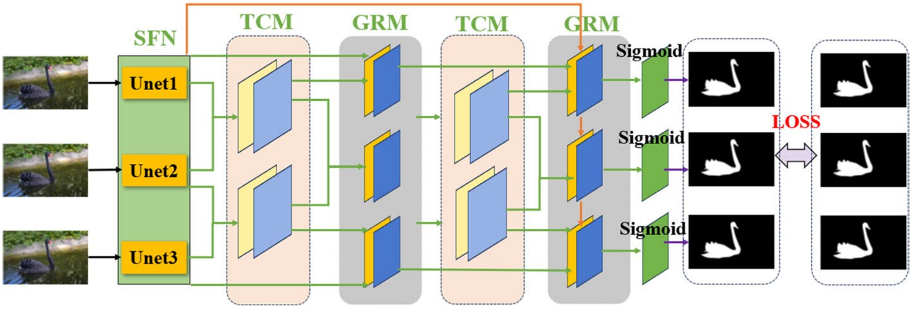


Fig. 1 Diagram of the designed network framework. SFN is a joint encoder and decoder network composed of three image saliency detection modules, whose initial weights are obtained by training using DUTS-TR. TCM and GRM recursively extract and deal with temporal differences of the edge or central areas of high-level semantic information output by SFN via skip-connection

areas by capturing the pixel gradients of initial images. In addition, a self-adjustment strategy is adopted for loss functions to avoid overfitting during training. The network framework of this work is shown in Fig. 1. Specifically, the raw data first enters the SFN network for feature extraction, and then inputs the TCM module for edge and center feature information enhancement. Then, the GRM module improves the detail information of the features, and then enters the TCM module for the second feature information enhancement, as well as the GRM module for the second feature information improvement. Residual connections have been added to supplement the information lost via feature enhancement through convolution, ensuring that modules do not mistakenly learn unimportant information due to information loss and preventing gradient explosion and fitting problem. Then, the sigmoid function is used to binarize the feature information and calculate the loss based on the results. The following section will provide a specific explanation of the module designed in this article.

3.1 Joint static feature extraction network

In view of the excellent feature extraction of VGG16, this paper uses U-Net with a VGG16 encoder as a subnet of the joint static feature extraction network to extract initial features of video frames. Because of lacking a sufficiently large annotated database for video saliency training, this paper uses the database of DUTS-TR to train the U-Net model and copy the weights obtained from training to the subnets for extracting features of the former, middle, and latter frames. It can ensure the highest relevancy of information output at the beginning of overall dynamic network training (Fig. 2).

We have designed unilateral information communication channels for three static extraction networks (U-Net) and used Formulas (1) and (2) to integrate high-level semantic information at the bottom of U-net to make sure middle frames can acquire more common features.

$$\mathrm{f} = (\mathrm{conv5})_{\mathrm{U1}} \times (\mathrm{conv5})_{\mathrm{U3}} \tag{1}$$

$$(\mathrm{conv5})_{\mathrm{U2}} = \mathrm{conv}_{1\times1}\left(\mathrm{cat}\left(\mathrm{f}, (\mathrm{conv5})_{\mathrm{U2}}\right)\right) \tag{2}$$

where f represents the features after fusion of features of the 5th layer; conv5 represents the features of the 5th layer; U1 represents the first sub-net; U2 represents the second

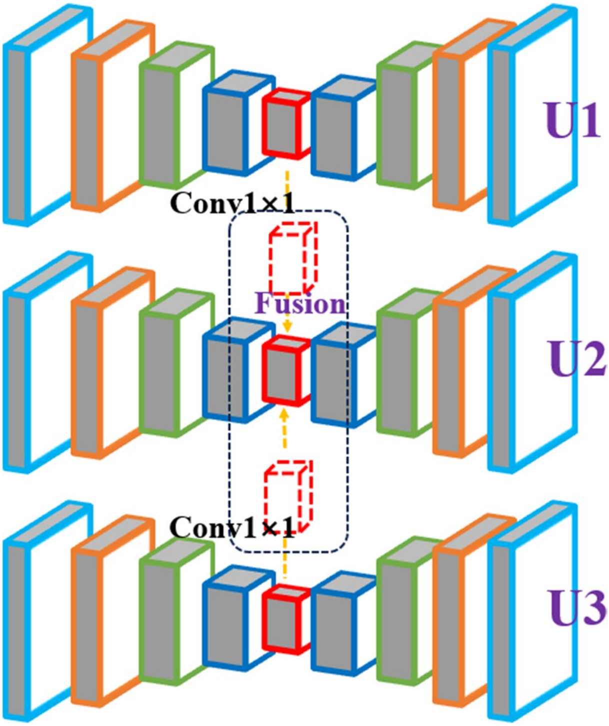


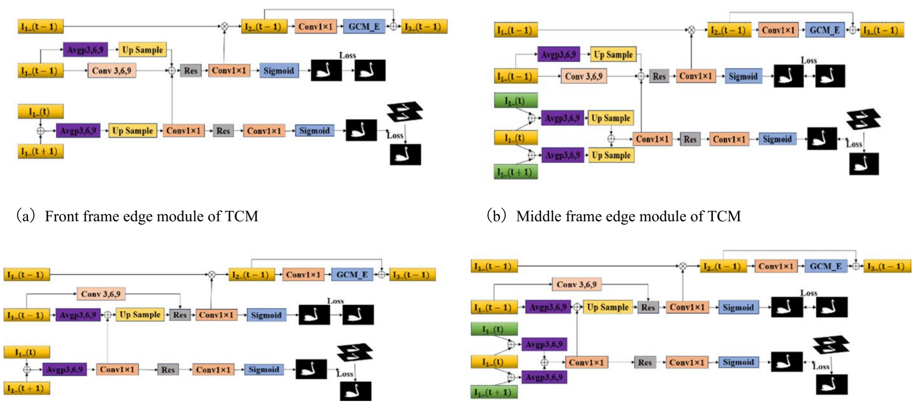
Fig. 2 Joint static feature extraction network

sub-net; U3 represents the third sub-net; cat (*) represents the overlapping of features via channels; conv1×1 (*) represents dimensionality reduction via 1×1 convolution.

3.2 Temporal differential collaborative attention module

Compared with image saliency detection, video saliency detection has its advantage: clues of sequential movements. To improve the performance of video saliency detection, we design a temporal differential co-attention edge module and central module aiming at the edge and center zones of targets, respectively.

As shown in Fig. 3(a), the co-attention edge modules of former frames integrates saliency features of SFN outputs through coordinating inter-frame relationships and then assign weights spatially. This module is composed of two parts: the former frame and the rest. The former frame is trained with annotated GT maps and the rest trained with the union of GT maps to obtain as much edge information as possible. Multi-scale pooling and multi-scale convolution are used for former frames to integrate their information. For the other part, multi-scale pooling and upsampling are used for dimension reduction and integrated information is added to the former frames. Overlapping information is deeply processed and reduced to one dimension via 1×1 convolution blocks, and then is multiplied by I1_(t-1) to put weight on initial data. Similar to Fig. 3(a) and (b) puts together temporal differences between middle frames and former and latter frames and adds them to middle frames. After dimension reduction, information of middle frames is updated and more attention can be paid to their edge zones. The TCM central module is designed following



（a）Front frame edge module of TCM

（b）Middle frame edge module of TCM

（c）Front frame center module of TCM

（d）Middle frame center module of TCM

Fig. 3 Temporal difference co-attention edge module，I1_(t-1) represents the former frame information output by SFN, I1_(t) represents the middle frame information output by SFN, I1_(t+1) represents the latter frame information output by SFN, I2_(t) represents the updated middle frame information, I3_(t) represents the updated middle frame information by the gradient transform module, Avgp3,6,9 represents the 3×3, 6×6, and 9×9 multiscale pooling operations, Conv3,6,9 represents 3×3, 6×6, and 9×9 multiscale convolution operations, Res represents residual block, conv1×1 represents 1×1 convolution block, $\otimes$ represents element-by-element multiplication, and $\oplus$ represents channel and channel concatenate

the similar principle except that the intersection of GT maps of other parts is used rather than their union, and convolution and pooling are conducted before upsampling.

3.3 Pixel gradient-based refinement processing module

As in original images, pixel values are approximate where salient objects locate but are distinct from those of the background along their edges, so pixel gradient can be properly used to set up a barrier to separate the inside from the outside. Once obtaining the profiles of salient objects, the pixel gradient can act as a powerful tool for capturing and utilizing the details of salient objects. In this paper, we use the TCM modules to integrate the location information of edge zones and central zones to support the gradient-based optimization module. As shown in Fig. 4, based on information provided by the TCM edge module, we define the edges of objects through shrinkage first and then expansion. It can help edge optimization get rid of interference by pixel information outside the edge. Meanwhile, as the size of salient objects varies, convolution blocks are designed specifically for multiple-scale pixel expansion. A more balanced approach is used to optimize the information output by TCM central modules. On the basis of shrinkage and expansion, the approach of expansion followed by shrinkage is also used for optimization because besides the problem of hollows, the integrity of edge information needs to be considered when dealing with central information. In addition, to give full play the combined effects of the pixel gradient module and TCM, they are jointly used to recurrently optimize the network structure.

$$s_g = \max(S) + \min(S) - s_g \tag{3}$$

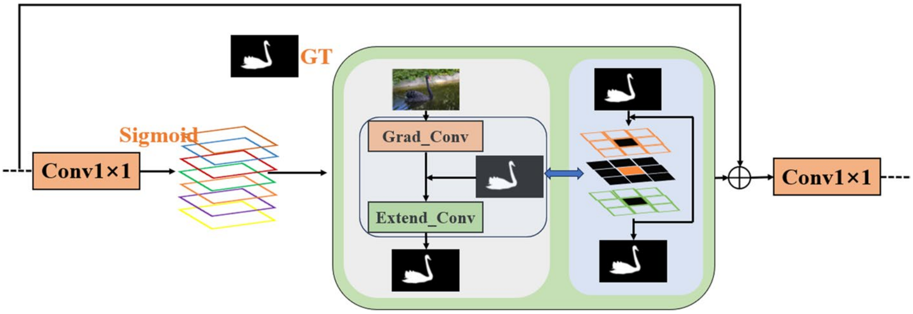


Fig. 4 Pixel gradient refinement module, First, initial saliency maps are obtained after dimension reduction of information from TCM by using 1×1 convolution blocks and then binarization by sigmoid functions. Besides, 8 pixel gradient information maps are acquired for each pixel point at its proximity by using pixel gradient convolution blocks, and these maps are used to expand or shrink the areas of similar pixel values for initial saliency maps. Expansion can be accomplished by using expansive convolution blocks and shrinkage accomplished by expansion after reversing the colors of saliency maps. (as shown in Formula 3)

S_g is gray value, max (S) is the maximum value in the saliency map, and min (S) is the minimum value in the saliency map.

3.4 Loss function self-adjustment strategy

As several TCM modules are involved in our method, which in turn require several loss functions to guide spatial attention, it is hence very necessary to coordinate the relationships between loss functions. Besides, as completely annotated video saliency detection datasets are lacked, overfitting often happens during training. To solve these two problems, a self-adjustment strategy is proposed for loss functions.

The loss function self-adjustment strategy includes testing the generalization performance of network training as shown in Fig. 5 and then adjusts the learning rate and the proportion of loss functions of different modules. In total, 51 binary cross-entropy loss functions are used in this paper, and 51 results obtained corresponding to 51 losses. If 26 of the losses pass the test, the initial learning rate remains unchanged; otherwise, it needs to be 1/10 the initial one. At the same time, the proportion of loss functions should be adjusted according to the number of losses tested qualified: The more qualified losses in a module, the larger proportion it will account for. The formula for the binary cross entropy loss function is expressed as follows:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \bullet \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \bullet \log\left(1 - p\left(y_i\right)\right) \tag{4}$$

4 Experiment

4.1 Datasets

Three commonly used benchmark datasets are used for evaluation in this paper: DAVSOD [16], SegTrackV2 [17], and DAVIS 2017 [18]. The dataset of DAVSOD (Densely

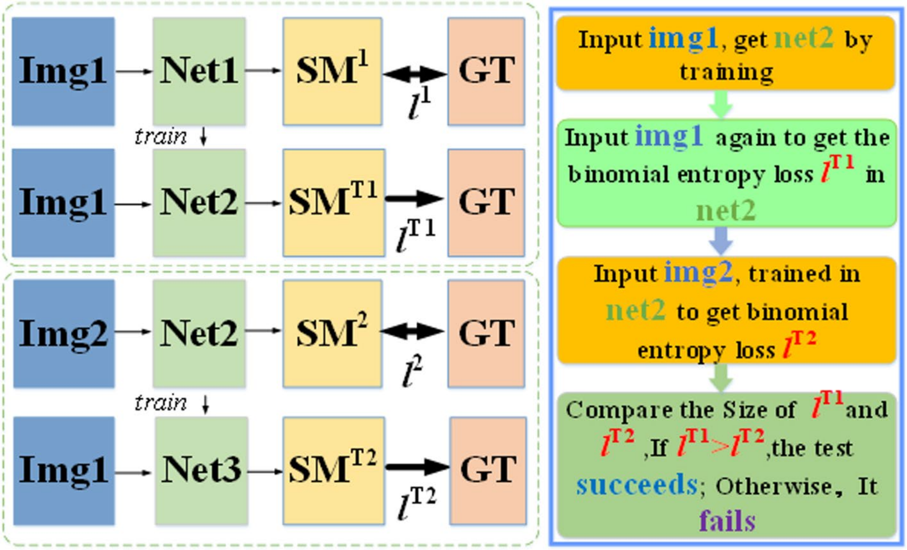

Img1
Net1
SM1
GT
l1
train
Img1
Net2
SMT1
GT
lT1
Img2
Net2
SM2
GT
l2
train
Img1
Net3
SMT2
GT
lT2
Input img1, get net2 by training
Input img1 again to get the binomial entropy loss lT1 in net2
Input img2, trained in net2 to get binomial entropy loss lT2
Compare the Size of lT1 and lT2 ,If lT1>lT2,the test succeeds; Otherwise, It fails

Fig. 5 Flow chart of loss function generalization test

Annotated VSOD) is composed of 226 motion video sequences showing obvious motion displacements of salient objects, and hence, is of great value for evaluating how well motion laws are captured. The dataset of SegTrack V2 includes 14 sequences of video frames featuring birds, animals, vehicles, and mankind, aiming at evaluating the effectiveness of the detection method when external interferences in their motions are severe. DAVIS 2017 (Densely Annotated Video Segmentation 2017) is composed of 90 high-quality, full HD video sequences, 6242 frames in total, and every frame has fully pixel-level annotated groundtruth data. DAVIS 2017 has been so far the most sophisticated dataset that covers all kinds of challenges in the way of segmentation of video objects, such as sheltering, motion ambiguity, and changes of appearance, and can evaluate the detection method from all aspects.

4.2 Evaluation criteria

To comprehensively evaluate the performance of the method, we compare it with 6 popular methods by using three indicators, which are precision-recall (PR) curves, the maximum F-measure [19], and mean absolute error (MAE).

$$\mathrm{F}_{\beta}=\frac{\left(1+\beta^{2}\right) Precision \times \mathrm{Recall}}{\beta^{2} \times Precision \times \mathrm{Recall}} \tag{5}$$

β^2 is used to put weight on precision and recall to reflect their importance, where β^2 is 0.3, indicating precision is more important. P-R curves and F-Measure are used to measure how perfectly saliency maps match GT after binarization. MAE can measure the level of similarity between the primary saliency map and GT by the average absolute difference in pixel values. The three indicators can demonstrate the value of the method proposed in the paper.

Our model is built within the PyTorch framework. To fit the training, we trained the model using Random Gradient Descent (SGD) to 300 epochs with a batch size of 16. This article used a total of 51 binary cross entropy loss functions, resulting in 51 results corresponding to 51 losses. The initial learning rate is set to 0.01, and if 26 losses pass the test, the initial learning rate remains unchanged; Otherwise, it is 1/10 of the initial value. And all experiments in this article were completed on NVIDIA RTX 1080 GPU.

4.3 Performance Comparison

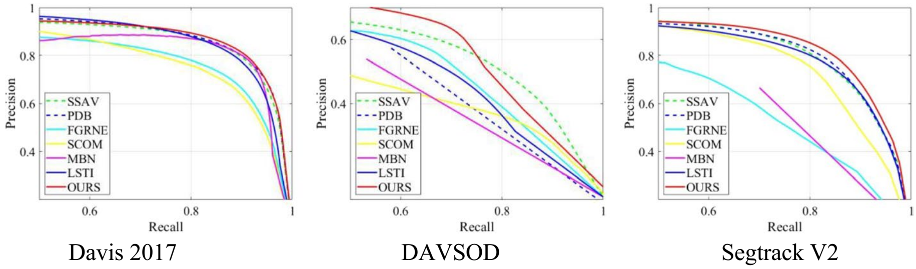
The method in this paper is compared with 8 other advanced detection methods, which are SSAV [20], PDB [21], FGRNE [22], SCOM [23], MBN [24] and LSTI [25]. The superiority of our network model has been highlighted in Figs. 6 and 7, and Table 1. For the sake of fairness, saliency maps used in the paper come from their authors.

4.3.1 Quantitative analysis

To quantitatively analyze the performance of the proposed method, three indicators including PR curves, $\max F_\beta$, and MAE are used in the section for comparing it with other mainstream video saliency detection methods. As indicated by PR curves in Fig. 6, our method has shown an excellent result with its precision and recall being the best among these methods as tested with all three datasets. As shown in Table 1 listing $\max F_\beta$ and MAE of all methods, ours is 0.57%, 0.24%, and 0.34% higher than the second best method in terms of $\max F_\beta$ when being tested with datasets of DAVIS 2017, DAVSOD, and SegTrackV2, respectively; when it comes to MAE, our method is 0.10%, 0.50%, and 0.19% lower. Our method has been proved to be effective as it leads other methods by wide margin in tests with all three different kinds of datasets.

4.3.2 Qualitative analysis

Figure 7 shows comparison of visual results from our network model and other network models. According to this comparison, the images based on our method have more clear-cut edges and more details in its center and are free from central hollows, achieving better visual effects than other methods. Our method has secured satisfactory effects for different kinds of primary images, which either have low contrast between the background or object (the first group), high contrast of the object itself (the second group), complicated background (the third group), or many details in the object (the fourth group).



Davis 2017 DAVSOD Segtrack V2

Fig. 6 P-R Curves in terms of three different databases

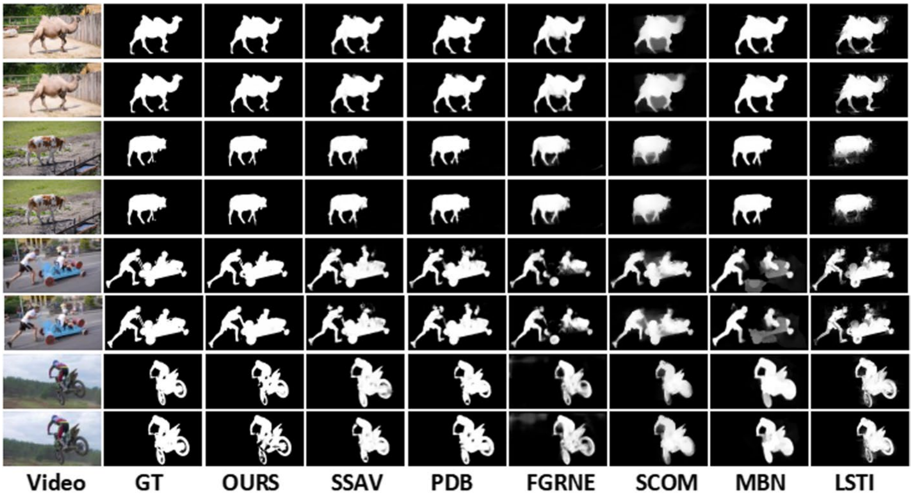


Fig. 7 Comparison of the results of different video saliency detection method

Table 1 Comparison of MAE and maximum F-measure of different methods

Methods	DAVIS 2017		DAVSOD		SegTrackV2	
	$\max F_\beta$	MAE	$\max F_\beta$	MAE	$\max F_\beta$	MAE
SSAV[20]	0.8618	0.0289	0.6229	0.0828	0.8366	0.0239
PDB[21]	0.8645	0.0291	0.5747	0.0962	0.8514	0.0238
FGRNE[22]	0.7970	0.0429	0.6150	0.0853	0.6941	0.0365
SCOM[23]	0.7895	0.0545	0.4943	0.2198	0.8105	0.0294
MBN[24]	0.8576	0.0342	0.5373	0.0996	0.6739	0.0301
LSTI[25]	0.8501	0.0342	0.5856	0.1067	0.8589	0.0253
Ours	0.8702	0.0279	0.6253	0.0778	0.8623	0.0219

4.4 Ablation experiment

In this sector, ablation experiments are conducted to confirm the effectiveness of each module of our model, for which DAVIS 2017 and SegTrackV2 are used.

4.4.1 Effectiveness of GRM

After adding the GRM module, the edge continuity for salient objects has been improved. As shown in Table 2, indicators for both DAVIS 2017 and SegTrackV2 are better than before. But from Fig. 8, it can be found that despite the better detection performance for edges, there are hollows in central areas.

4.4.2 Effectiveness of GRM and TCM

In view of the different results in detecting central and edge areas, TCM is used to draw attention to both the central and edge areas. Then GRM is used to deal with different areas

Table 2 MAE and maxF_β for the ablation experimental method

Configurations	DAVIS 2017		SegTrackV2	
	maxF_β	MAE	maxF_β	MAE
SFN	0.8259	0.0423	0.8100	0.0307
SFN+GRM	0.8384	0.0349	0.8323	0.0242
SFN+GRM+TCM	0.8663	0.0309	0.8585	0.0239
SFN+GRM+TCM+REC	0.8695	0.0287	0.8605	0.0229
SFN+GRM+TCM+REC+Self_Loss	0.8702	0.0279	0.8623	0.0219

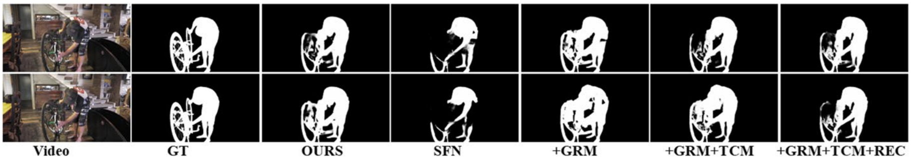


Fig. 8 Visual effect of Ablation Experiment

distinctively to ensure edge continuity and avoid hollows in central areas. As can be seen in Fig. 8, hollows in the central area after applying TCM are much less than before.

4.4.3 Effectiveness of REC and Self_Loss

As shown in Table 2, recurrent optimization and self-adjustment loss functions (Self_Loss) have improved the final outcomes. The recurrent optimization strategy (REC) has exploited the potential of the combination GRM and TCM at maximum, and Self_Loss effectively alleviate overfitting present in training and also help maintain the stability of network structure. As shown in Fig. 8, these strategies have contributed to a better detection performance.

5 Conclusion

This paper proposes a deep learning method for video saliency object detection based on temporal difference and pixel gradient. The detection model is mainly composed of temporal differential co-attention module and pixel gradient refinement module. The co-attention module makes statistics of spatiotemporal information of salient objects, and then the pixel gradient refinement module conducts pixel-level optimization of video frames. And the recurrent optimization strategy is applied to improve the precision of spatiotemporal saliency results. In addition, we propose the loss function self-adjustment strategy to enhance net stability and deal with problems like overfitting. Experiments with several publicly available datasets verify that the designed method can provide better indicators and visual effects.

Current research has demonstrated that our designed common attention mechanism and pixel gradient refinement module can effectively solve the problem of poor continuity of differential features between consecutive frames. However, video saliency object detection still faces various challenges brought by continuous motion clues and rich inter frame information. Thus, the current research has not been able to fully solve all the issues of video saliency detection under complicated scenarios in the wild, which is our future research direction.

Acknowledgments This work was supported by National Natural Science Foundation of China (NSFC) (61976123, 61601427, 61876098); the Taishan Young Scholars Program of Shandong Province; and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Guo C, Zhang L (2009) A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Transactions on image processing 19(1):185–198
2. Wu H, Li G, Luo X (2014) Weighted attentional blocks for probabilistic object tracking. The Visual Computer 30(2):229–243
3. Fan Q, Luo W, Xia Y et al (2019) Metrics and methods of video quality assessment: a brief review. Multimedia Tools and Applications 78(22):31019–31033
4. Götze N, Mertsching B, Schmalz S, et al. (1996) Multistage recognition of complex objects with the active vision system NAVIS
5. Lu X, Yuan Y, Zheng X (2016) Joint dictionary learning for multispectral change detection. IEEE Transactions on cybernetics 47(4):884–897
6. Wang Q, Wan J, Yuan Y (2018) Locality constraint distance metric learning for traffic congestion detection. Pattern Recognition 75:272–281
7. Wang Q, Gao J, Yuan Y (2017) Embedding structured contour and location prior in siamesed fully convolutional networks for road detection. IEEE Transactions on Intelligent Transportation Systems 19(1):230–241
8. Wang Q, Gao J, Yuan Y (2017) A joint convolutional neural networks and context transfer for street scenes labeling. IEEE Transactions on Intelligent Transportation Systems 19(5):1457–1470
9. Wang Q, Wan J, Yuan Y (2017) Deep metric learning for crowdedness regression. IEEE Transactions on Circuits and Systems for Video Technology 28(10):2633–2643
10. Yang J, Yang MH (2016) Top-down visual saliency via joint CRF and dictionary learning. IEEE transactions on pattern analysis and machine intelligence 39(3):576–588
11. Gao D, Vasconcelos N (2007) Bottom-up saliency is a discriminant process 2007 IEEE 11th International Conference on Computer Vision. IEEE, 1-6
12. Cheng MM, Mitra NJ, Huang X et al (2014) Global contrast based salient region detection. IEEE transactions on pattern analysis and machine intelligence 37(3):569–582
13. Fang Y, Wang Z, Lin W et al (2014) Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE transactions on image processing 23(9):3910–3921
14. Wang W, Shen J, Shao L (2015) Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing 24(11):4185–4196
15. Wang W, Shen J, Shao L (2017) Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing 27(1):38–49
16. Brox T, Malik J (2010) Object segmentation by long term analysis of point trajectories European conference on computer vision. Springer, Berlin, Heidelberg, pp 282–295
17. Li F, Kim T, Humayun A, et al. (2013) Video segmentation by tracking many figure-ground segments Proceedings of the IEEE International Conference on Computer Vision. 2192-2199
18. Perazzi F, Pont-Tuset J, McWilliams B, et al. (2016) A benchmark dataset and evaluation methodology for video object segmentation Proceedings of the IEEE conference on computer vision and pattern recognition. 724-732

19. Achanta R, Hemami S, Estrada F, et al. (2009) Frequency-tuned salient region detection 2009 IEEE conference on computer vision and pattern recognition. IEEE, 1597-1604
20. Fan D P, Wang W, Cheng M M, et al. (2019) Shifting more attention to video salient object detection Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8554-8564
21. Song H, Wang W, Zhao S, et al. (2018) Pyramid dilated deeper convlstm for video salient object detection Proceedings of the European conference on computer vision (ECCV). 715-731
22. Li G, Xie Y, Wei T, et al. (2018) Flow guided recurrent neural encoder for video salient object detection Proceedings of the IEEE conference on computer vision and pattern recognition. 3243-3252
23. Chen Y, Zou W, Tang Y et al (2018) SCOM: Spatiotemporal constrained optimization for salient object detection. IEEE Transactions on Image Processing 27(7):3345–3357
24. Li S, Seybold B, Vorobyov A, et al. (2018) Unsupervised video object segmentation with motion-based bilateral networks proceedings of the European Conference on Computer Vision (ECCV). 207-223
25. Wang B, Liu W, Han G et al (2020) Learning long-term structural dependencies for video salient object detection. IEEE Transactions on Image Processing 29:9017–9031
26. Jian M, Lam K-M, Dong J, Shen L (2014) Visual-patch-attention aware saliency detection, IEEE Trans Cybern, pp. 1575–1586
27. Wang Q, Lin J, Yuan Y (2016) Salient band selection for hyperspectral image classification via manifold ranking, IEEE Transactions on Neural Networks and Learning Systems, 1279–1289
28. Han J, Chen H, Liu N, Yan C, Li X (2017) Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics 48(11):3171–3183
29. Cong R, Lei J, Fu H, Lin W, Huang Q, Cao X, Hou C (2019) An iterative co-saliency framework for rgbd images. IEEE Transactions on Cybernetics 49(1):233–246
30. Cong R, Lei J, Fu H, Hou J, Huang Q, Kwong S (2020) Going from rgb to rgbd saliency: A depth-guided transformation model. IEEE Transactions on Cybernetics 50(8):3627–3639
31. Zhang M, Ji W, Piao Y, Li J, Zhang Y, Xu S, Lu H (2020) Lfnet: Light field fusion network for salient object detection. IEEE Transactions on Image Processing 29:6276–6287
32. Li C, Cong R, Kwong S, Hou J, Fu H, Zhu G, Zhang D, Huang Q (2020) Asif-net: Attention steered interweave fusion network for rgb-d salient object detection, IEEE Trans Cybern, pp.1–13
33. Jian M, Qi Q, Dong J et al (2018) Saliency detection using quaternionic distance based weber local descriptor and level priors. Multimed Tools Appl 77:14343–14360
34. Jian M, Wang J, Dong J et al (2020) Saliency detection using multiple low-level priors and a propagation mechanism. Multimed Tools Appl 79:33467–33482
35. Hu R, Deng Z, Zhu X. Multi-scale Graph Fusion for Co-saliency Detection. Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 7789–7796
36. Wang Z, Zhou Z, Lu H, Jiang J et al (2020) Global and local sensitivity guided key salient object re-augmentation for video saliency detection. Pattern Recognition 103:107275
37. Zhang K, Dong M, Liu B et al. (2021) DeepACG: Co-Saliency Detection via Semantic-aware Contrast Gromov-Wasserstein Distance. the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13703-13712
38. Wang Y, Wang R, Fan X, Wang T, He X (2023) Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 10031-10040

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xiangwei Lu is pursuing his Master's degree supervised by Prof. Muwei Jian, at the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China. His research interests include image processing, computer vision, and saliency object detection.

Muwei Jian received the PhD degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, in October 2014. He was a Lecturer with the Department of Computer Science and Technology, Ocean University of China, from 2015 to 2017. Currently, Dr. Jian is a Professor and Ph.D Supervisor at the School of Computer Science and Technology, Shandong University of Finance and Economics. His current research interests include human face recognition, image and video processing, machine learning and computer vision. Prof. Jian holds 3 granted national patents and has published over 60 papers in refereed international leading journals/conferences such as IEEE Trans. on Cybernetics, IEEE Trans. on Multimedia, IEEE Trans. on Circuits and Systems for Video Technology, Pattern Recognition, Information Sciences, Signal Processing, ISCAS, ICME and ICIP. Prof. Jian was actively involved in professional activities. He has been a member of the Program Committee and Special Session Chair of several international conferences, such as SNPD 2007, ICIS 2008, APSIPA 2015, EEECS 2016, ICTAI2016, ICGIP 2016, ICTAI 2017 and ICTAI 2018. Dr. Jian has also served as a reviewer for several international SCI-indexed journals, including IEEE Trans., Pattern Recognition, Information Sciences, Computers in Industry, Machine Vision and Applications, Machine Learning and Cybernetics, The Imaging Science Journal, and Multimedia Tools and Applications. Currently, he serves as an Associate Editor of IET Computers & Digital Techniques and the Journal of image and Graphics (JIG).

Hui Yu is Professor with the University of Portsmouth and an Industrial Fellow of the Royal Academy of Engineering, UK. He worked at the University of Glasgow and Queen's University Belfast before joining the University of Portsmouth in 2012. His research interests include vision, creative computing and AI with applications to 4D facial and affective analysis, human-machine interaction, VR/AR, video analysis and classification. He serves as an Associate Editor of the IEEE Transactions on Human-Machine Systems and the IEEE Transactions on Computational Social Systems journal.