Hindawi
Scientific Programming
Volume 2022, Article ID 8148972, 7 pages
https://doi.org/10.1155/2022/8148972

Research Article

Deep Neural Network-Based Sports Marketing Video Detection Research

Longcheng Xu,[1] Deokhwan Choi ,[2] and Zeyun Yang [2]

[1]*Physical Education Department of Jiangsu Vocational College of Finance and Economics, Huaian, Jiangsu 223000, China*
[2]*Physical Education Department of Woosuk University, Jeonju, Jeollabuk-do 55340, Republic of Korea*

Correspondence should be addressed to Deokhwan Choi; y553149602@163.com

Received 5 January 2022; Accepted 11 February 2022; Published 19 March 2022

Academic Editor: Baiyuan Ding

With the rapid development of short video, the mode of sports marketing has diversified, and the difficulty of accurately detecting marketing videos has increased. Identifying certain key images in the video is the focus of detection, and then, analysis can effectively detect sports marketing videos. The research of video key image detection based on deep neural network is proposed to solve the problem of unclear and unrecognizable boundaries of key images for multiscene recognition. First, the key image detection model of the feedback network is proposed, and ablation experiments are conducted on a simple test set of DAVSOD. The experimental results show that the proposed model achieves better performance in both quantitative evaluation and visual effects and can accurately capture the overall shape of significant objects. The hybrid loss function is also introduced to identify the boundaries of key images, and the experimental results show that the proposed model outperforms or is comparable to the current state-of-the-art video significant object detection models in terms of quantitative evaluation and visual effects.

1. Introduction

Vision is the main way humans receive information from the outside world, and according to research in the field of neuroscience, about 10^8 to 10^9 bytes of data enter the human eye every second [1]. This is because of the selective role of the visual attention mechanism, which allows the visual system to selectively ignore irrelevant information and pay attention to relevant information, just like separating the grains of wheat from the husk. In this Internet era where the amount of data is exploding, how to get the information of people's concern from the huge amount of information in a labor and material-saving way has gained a lot of attention. Therefore, introducing attention mechanisms into data processing tasks and prioritizing the allocation of data processing resources to more critical information can help improve the efficiency of processing information [2–6].

In 1998, Borji and Itti [7] proposed the first computational model of visual saliency based on Koch et al.'s theory and the classical feature integration theory of cognitive psychology [8] and the pointing search model [9], whose algorithmic process contains three main steps: extraction of three primary visual features: color, luminance, and orientation. Three types of key features are computed at multiple scales using central-peripheral contrast (key feature extraction); the feature maps are normalized and then synthesized (feature fusion), and the key targets in the images are labeled using the WTA mechanism. The algorithm has had a significant impact on subsequent research on computational models of visual criticality in the field of computer vision, especially since mainstream criticality detection algorithms used a similar framework before deep learning techniques were used on a large scale.

Early image salient object detection models [10] were mainly based on a bottom-up approach using different underlying visual features, such as color, edges. Since salient object detection is closely related to the human eye attention detection task and both model the human visual attention mechanism, the early salient object detection models also borrowed some basic theories of the human visual attention mechanism, including the classical contrast assumption, center-surround assumption. For example, both

assumptions were used by Liu et al. [11] and Achanta et al. [12], and a similar assumption was used by Cheng et al. [10], who considered color contrast information on both local and global scales, and the algorithm was concise and straightforward and received wide attention from the academic community. In addition, Yan et al. [13] proposed to complete the apparently consistent image representation at different scales by over-segmenting the image at different scales and to extract and fuse the salient features at different scales for optimization to obtain the final salient object detection results. Visual center bias is also a commonly used hypothesis based on human attentional mechanisms [13]. The hypothesis is based on the phenomenon that the human visual system has a tendency to assign higher attentional weights to the center of the scene when observing the scene. After that, the popular hypothesis is the background prior hypothesis, which was proposed by Wei et al. [14] in 2012. Unlike the center-periphery hypothesis and the visual center shift hypothesis, which attempt to define "what is more likely to be the salient region," this hypothesis attempts to define "what is more likely to be the background." This assumption is based on the observation that in most scenes, the parts around the edges of the image have a higher probability of belonging to the background. This assumption can be considered as a further development of the visual center bias assumption. Before the large-scale application of deep learning techniques, the background prior assumption was the most effective assumption in the field of saliency detection, and the majority of high-performing models [15–19] were based on this assumption. These works focus on how to further improve the accuracy of the background prior assumption and how to apply more advanced one-class classifiers. By the background prior assumption, which is equivalent to obtaining a class of (background) samples, the problem can be considered as a one-class classifier giving only one class of samples.

With the great success of deep learning techniques in image classification problems, the focus of research in the field of significant object detection has gradually shifted to deep learning-based models. Slightly earlier work used deep learning features as a more effective key representation and trained using fully convolutional neural networks. Lee et al. [20] used depth features as high-level information and Gabor-filtered response and color histogram as bottom-level features to fuse different levels of significant information for significant prediction. These models achieve better performance but have some drawbacks, such as the large number of parameters and loss of spatial information due to the use of fully connected layer-based classification networks and the high computational cost of these algorithms due to the need for significant/insignificant classification of each superpixel or target object alternative.

With the rise of fully convolutional neural networks, in recent years, significant object detection efforts based on deep learning have used or adapted full convolutional neural networks for pixel-level critical prediction. There is some work [21] inspired by the pixel-level semantic segmentation task, proposing the fusion of features from different neural network layers for critical object detection. Since the shallower layers of deep neural networks can retain more fine-grained underlying visual features, and the deeper layers can extract higher-level, semantic-level features, the fusion of features from different neural network layers can retain the original underlying spatial information and obtain higher-level semantic information. Currently, the main research focus of the work on significant object detection based on deep learning techniques is to explore more efficient network structures that can retain more spatial details. Wang et al. [22] proposed an ASNet model for detecting visually salient objects by means of visual attention prior. The model treats visual attention as a high-level understanding of the whole scene, which is learned through higher-level neural network layers, and the salient object detection task is considered as a more fine-grained, object-level saliency detection, with visual attention providing top-down guidance. The ASNet model is based on a stacked convolutional long and short-term memory neural network, which has a unique recurrent structure that can iteratively optimize saliency detection results. This work provides a deeper understanding of the visual attention mechanism and reveals the correlation between salient object detection and human eye attention detection. As a whole, the deep learning-based salient object detection model achieves much better performance than traditional models [23–26].

In response to the current research status, this paper investigates video salient object detection based on deep neural networks as follows, extracting richer spatial saliency information and better capturing the overall shape of salient objects. In this paper, an attention feedback network-based video salient object detection model is proposed. To further obtain clearer bounds, a new hybrid loss function is introduced in this paper based on the video salient object detection model and the attentional feedback network.

2. Deep Neural Networks

2.1. Convolutional Neural Network. When people read or watch a video, they perceive and understand the current content based on the text or images they have already observed before and do not completely forget what they have observed before, and their brain goes blank to understand the content that follows. Traditional neural networks cannot predict salient information in later frames based on the salient object regions in the previous video frames. The emergence of recurrent neural networks makes the network memorable, and its network structure is shown in Figure 1. Assuming that $\{X_t\}_{t=0}^{t}$ is a set of inputs with $(t+1)$ time steps and $\{H_t\}_{t=0}^{t}$ is the corresponding output of the network, network N receives at time step t not only X_t but also the value of the first $(t-1)$ value of the hidden state at a time step, that is, the network processes the current input with reference to the previous memory.

However, when the video sequence is long, the interval between the current video frame to be processed and the related video frame may be large, and at this time, the RNN may lose the memory of distant video frames due to problems such as gradient disappearance. To address the problem of long-term dependence, Hochreiter et al. [27]

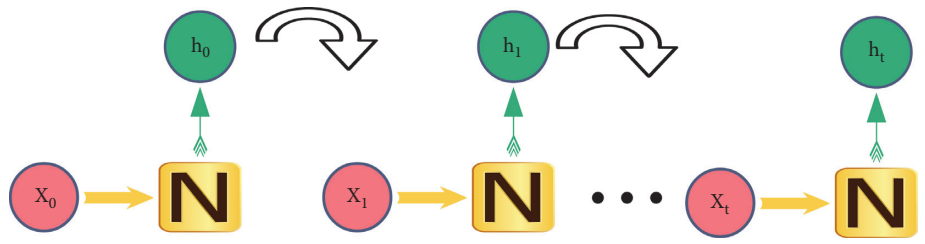

Figure 1: Network structure.

proposed a long-term and short-term memory network, as shown in Figure 2, where the contents of the three stages indicate the forgetting phase, updating state phase, and output phase, respectively.

All three stages contain a sigmoid layer that maps the input information to between [0, 1] and then selectively filters the useful information and forgets the useless information by a per-bit multiplication operation.

The forgetting stage is used to filter the useful information and forget the useless information. The current input is x_t, connecting x_t with the hidden state h_{t-1} of the previous moment, denoted as J_t, and ※ denotes the connection operation, as shown as follows:

$$h_t \text{※} x_t = J_t. \tag{1}$$

The sigmoid layer is then used to map J_t to between [0, 1] to obtain the output gate f_t, where W_f and b_f denote the weight and bias vector of the network layer, respectively, and σ denotes the sigmoid operation, as shown as follows:

$$f_t = \sigma\left(W_f \cdot J_t + b_f\right). \tag{2}$$

Then, the corresponding element multiplication operation $(\therefore)$ is performed with the cell state C_{t-1}, thus selectively filtering the useful information and forgetting the useless information, and the cell state at this point is noted as C_t^*.

$$C_t^* = f_t \therefore C_{t-1}. \tag{3}$$

The update cell state phase allows the control cell state to selectively absorb relevant information from J. J_t passes through the sigmoid layer and generates the input gate i_t.

$$i_t = \sigma\left(W_i \cdot J_t + b_i\right). \tag{4}$$

The information obtained by multiplying the feature obtained by J_t after the tanh layer with the corresponding element of it is the information added to the cell state, and the new cell state C_t^* is obtained by adding this information to the C_t obtained in the forgetting phase by bits.

$$C_t = C_t^* + i_t \therefore \tanh\left(W_c \cdot J_t + b_c\right). \tag{5}$$

The output phase controls what information is output at the current moment. J_t is inputted into the sigmoid layer to get an output gate O_t.

$$O_t = \sigma\left(W_o \cdot J_t + b_o\right). \tag{6}$$

Let O_t and the current cell state C_t be multiplied bitwise by the features obtained through the tanh layer to obtain the output at the current moment H_t.

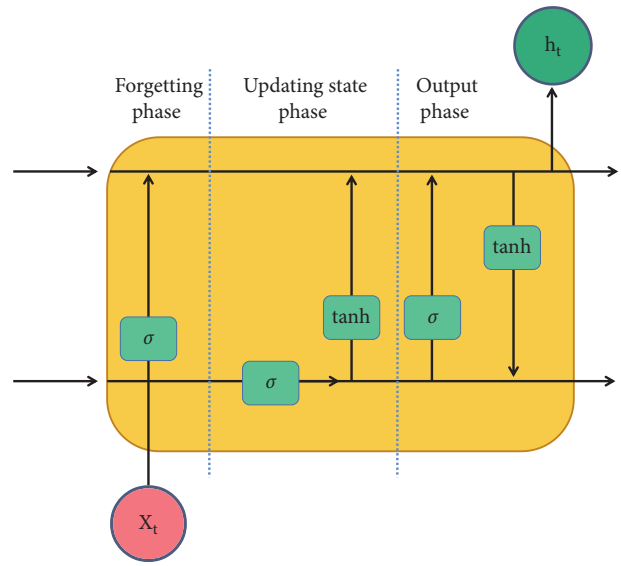

Figure 2: Memory network.

$$H_t = O_t \therefore \tanh\left(C_t\right). \tag{7}$$

2.2. Loss Function. When performing pixel-level salient object detection, it can be viewed as a binary partitioning problem, where pixels belonging to the salient object are labeled as 1 and pixels belonging to the background are labeled as 0. Assume that y_i denotes the label of sample x_i, the desired output, and $\overline{y}$ denotes the probability value of $y_i = 1$ for a given sample x_i.

$$\overline{y}_i = P\left(y_i = 1 | x_i\right). \tag{8}$$

$1 - \overline{y}$ denotes the probability value of $y_i = 0$ given sample x_i.

$$1 - \overline{y}_i = P\left(y_i = 0 | x_i\right). \tag{9}$$

When x_i occurs, the probability of y_i occurrence can be expressed by $P(y_i \mid x_i)$. From the perspective of maximum likelihood, $P(y_i|x_i)$ can be expressed in the following form.

$$P\left(y_i | x_i\right) = \overline{y}^{y_i} \cdot \left(1 - \overline{y}\right)^{1 - y_i}. \tag{10}$$

When the real mark $y_i = 0$, 1, and take the logarithm operation. Since the smaller the value of the loss function, the more favorable it is, and the log takes a negative value, and the loss function is calculated as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N} -[y_i \log \overline{y} + (1 - y_i)\log(1 - \overline{y})]. \quad (11)$$

2.3. Feedback Network. In order to reduce the loss of necessary visual criticality information due to repetitive stride and pooling operations and to learn richer static criticality information, AFNet is used as the main skeleton of the static criticality module. Stimuli in Figure 3 show the input image frames, and the encoding and decoding networks consist of five convolutional blocks of VGG16 (denoted as E^i and D^i, respectively, $i \in \{1, 2, 3, 4, 5\}$), where the information transfer between the corresponding convolutional blocks is controlled by the attention feedback module.

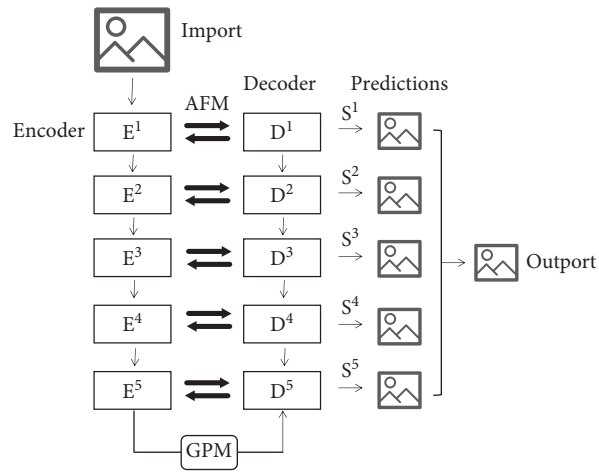

Figure 3: Feedback network model.

3. Design of Deep Neural Network

3.1. Feedback Network Detection Model. The NHM model is proposed to capture richer spatial criticality information and thus better capture the overall shape of key images. The NHM model uses the attentional feedback network as the backbone of the static criticality module to reduce the loss of visually critical information caused by scale-space issues and to guide the correct fusion of multiscale features from coarse to fine scales. The multiscale feature maps extracted from the five decoding blocks of the attentional feedback network are then fused and fed to the pyramidal expansion convolution module to retain more spatial visual critical information. After that, the time-critical information is captured using a key object transfer-aware convolutional long short-term memory network in consideration of attention-aware transfer, and finally, the parameters of the model are optimized by gradually reducing the value of the loss function through continuous iterations. The algorithm is divided into three parts: extraction of multiscale spatial features, integration of spatio-temporal critical information, and loss minimization.

To mitigate the negative effects such as the loss of visual information generated by the scale-space problem, the backbone of the static criticality detection module consists of AFNet and PDC modules connected together. AFNet as a novel codec forms the design of a fully convolutional network, its encoding and decoding network consists of five convolutional blocks, and E^i and D^i denote the encoder and decoder blocks, respectively, where $i \in \{1, 2, 3, 4, 5\}$, indicating that E^i and D^i each contain five convolutional blocks, where each layer of the encoder block transmits its critical information through the feedback module in AFNet to the corresponding decoder block. The feedback module uses a two-step iterative learning approach, where the time steps are denoted by $i \in \{1, 2\}$, which helps to correct inaccurate predictions generated in the previous network by simulating a feedback mechanism that multiplies the ternary map pixel by pixel with the obtained feature map, thus helping to capture the overall shape of the key object. Facing the global spatial criticality detection problem, AFNet uses the global perception module to overcome the problem that the fully connected operation ignores local information and generates redundant data. A multiscale segmentation strategy is used to divide the feature map into 4, 16, and 36 parts, which are then stacked and reorganized for global convolution operation to make full use of the global and intraregional saliency information.

The key image in the dynamic scene is detected directly by the image key object detection model. The key object detection can only detect the spatial differences of color contrast, direction contrast, brightness contrast, and so on. However, in dynamic scenes, the temporal factor is usually used as an important clue for the criticality detection. Second, detection only on each individual frame without reference to the criticality information contained in previous frames may be highly incoherent, because the target and background may differ significantly in appearance in different frames, which will lead to incoherent detection results between frames. Finally, video content often contains significant redundancy, as consecutive video frames require enough similar content to provide a smooth viewing experience. Simply ignoring content redundancy can lead to higher computational costs. Therefore, VSOD needs to consider both temporal and spatial saliency information, so a dynamic saliency detection module is used to integrate temporal and spatial saliency information. In order to better simulate the perceptual function of the human visual system, temporal saliency information is learned, and the process of attentional perceptual transfer is captured, and this paper uses SSLSTM as a dynamic saliency detection module, which combines the powerful spatio-temporal feature extraction capability of ConvLSTM with the attentional transfer mechanism.

Deep neural networks gradually optimize the network by iteratively minimizing the loss function. The loss function measures the difference between the value predicted by the model and the true value, and the weights of the network are updated by gradient descent.

$$L = \sum_{t=1}^{T} \left(l(I_t) \cdot l^{Att}(A_t, F_t) + l^{\mathrm{VSDD}}(S_t, M_t) \right). \quad (12)$$

The meaning of each symbol is shown in Table 1, because the video significant object detection dataset contains relatively

Table 1: Compare other model.

Metric		LNSM	MBNM	PDBM	SSAV
ViSal	S	0.993	0.883	0.884	0.933
	MAE	0.025	0.01	0.026	0.022
	Max f	0.993	0.873	0.866	0.929
DAVSOD	S	0.779	0.627	0.67	0.714
	MAE	0.14	0.149	0.097	0.082
	Max f	0.68	0.51	0.575	0.593
UVSD	S	0.825	0.688	0.891	0.85
	MAE	0.053	0.069	0.008	0.015
	Max f	0.701	0.54	0.853	0.791
VOS	S	0.86	0.732	0.794	0.808
	MAE	0.122	0.089	0.061	0.063
	Max f	0.779	0.66	0.728	0.732

few human eye focus annotations, so l_t to indicate whether the dataset contains human eye focus annotations, when the dataset does not contain human eye focus annotations, the loss function at this time does not contain the $l_t A_t$ term, the error will not be back-propagated. The meaning of each symbol is shown in Table 1. Since the video important target detection data set contains relatively few eye focus annotations, it is used to indicate whether the data set contains eye focus annotations. When the data set does not contain eye focus annotations, the loss function at this time does not contain ItAt term, and the error will not be propagated back.

3.2. Loss Function Design. A novel hybrid loss function is proposed based on the boundary enhancement loss, and the function consists of the loss L^a of the predicted attention-perception feature map, the loss L^v of the final key object prediction result, and the loss L_b^v of the final predicted target boundary.

$$L = \omega_1 \cdot \left(L^a + L^v\right) + \omega_2 \cdot L_b^v, \tag{13}$$

where ω_1, ω_2 are used as the learning rate parameters for object-level loss and object-boundary loss of the control target, respectively, and let $\omega_1 : \omega_2 = 1 : 10$ to emphasize the learning of the target boundary.

The dataset used for part of the training does not contain human eye focus annotations, so the predicted loss L^a of the perceptual attention feature map can be divided into two parts: loss L_f^a calculated using human eye focus annotations and loss L_m^a calculated using salient object annotations.

$$L^a = \delta(1) \cdot L_f^a + (1 - \delta(1)) \cdot L_m^a, \tag{14}$$

$L^a = L_f^a$ when $\delta(1) = 0$, $L^a = L_m^a$ when $\delta(1) = 1$. The final key object prediction results are denoted by S_t. That is, the loss L^v can be calculated.

When $\delta(1) = 0$, $L^a = L_f^a$; When $\delta(1) = 1$, $L^a = L_m^a$. S_t is used to represent the prediction result of the final key object, and M_t represents the object level annotation of the key object. The loss L^v can be calculated as follows:

$$L^v = \sum_{t=1}^{T} \left(L^{VSO\ D}(S_t, M_t)\right). \tag{15}$$

The average pooling operation $\overline{P}$ can be used to extract smooth boundaries. Suppose it is necessary to extract the boundary $B(X)$ of the image X and take the absolute value after making a difference between X and $\overline{P}(X)$. The final predicted target boundary loss L_b^v is as follows:

$$L_b^v = \sum_{t=1}^{T} \left(L^{VSOD}(B(S_t), B(M_t))\right). \tag{16}$$

On the basis of NHM, a mixed loss function for capturing clear boundaries is added. The loss function is based on the boundary enhancement loss and is composed of the attention perception feature map predicted by the model, the prediction results of key images, and the prediction results of key image boundaries. The model is recorded as LNSM.

4. Experiments and Results

4.1. Experimental Design. The experiments were run on an Nvidia GTX1080TI GPU. The experiments in this paper were implemented using the Python language on Caffe's deep learning framework, and Matlab was used for quantitative evaluation of performance. The training set of DAVIS, DAVSOD, and FBMS and the validation set of DAVSOD were also used to train the proposed model, where the weights of the network model were initialised by the AFNet model, and video was processed per batch, and the number of time steps for the conLSTM network layer processing was set to 3. The training process was set up as follows: first, the static key model was pretrained with a base learning rate of 10^{-9}; then, the entire model was trained by setting the learning rate of the dynamic key module to 10^{-8} and the learning rate of the static key module to 10^{-10}; finally, the static key module weights were fixed, and the dynamic key module was fine-tuned with the learning rate set to 10^{-10}. The LNSM module was trained using 32 hours and 64 k iterations.

4.2. Compare Other Model. In this paper, the proposed LNSM is compared with four advanced video critical object detection models, MBNM, PDBM, and SSAV, on datasets created specifically for the VSOD task (the entire dataset for ViSal and UVSD, a test set for VOS, and a simple test set for DAVSOD), and the experimental results of the quantitative evaluation are shown in Table 1. It can be seen from Table 1 that the three indexes of the model proposed in this paper are better than other models on DAVSOD and ViSal datasets. Especially on the simple test set of DAVSOD, the *f*-value index and average absolute error based on pixel error and the structural index measuring the overall structural difference have improved the performance by 0.06, 0.03, and 0.064, respectively, compared with SSAV; advanced performance has also been achieved on other datasets. Moreover, ViSal is the first test benchmark especially designed for video key object detection; DAVSOD dataset takes into account the transfer of visual attention and its selectivity when labeling and can represent the real attention behavior of the human visual system in dynamic scene. These two

datasets are very representative. The experimental results show that the LNSM model has good performance for creating datasets especially for VSOD and DAVSOD datasets that mark key images according to human eye concerns.

5. Conclusion

This paper focuses on key image detection based on deep neural networks to complete the detection of sports marketing videos. For the detection of multiple scenes, a feedback network-based video off-image detection model and a hybrid loss function are proposed to solve the detection problem of key images. The LNSM model proposed in this paper is compared with the quantitative evaluation and visualisation results of the three state-of-the-art models on six representative datasets. The quantitative results demonstrate that LNSM outperforms other advanced models in all three evaluation metrics on the DAVSOD and ViSal datasets and achieves advanced performance comparable to other models on widely used datasets.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] K. Koch, J. McLean, R. Segev et al., "How much the eye tells the brain," *Current Biology*, vol. 16, no. 14, pp. 1428–1434, 2006.

[2] V. R. S. Mani, A. Saravanaselvan, and N. Arumugam, "Performance comparison of CNN, QNN and BNN deep neural networks for real-time object detection using ZYNQ FPGA node," *Microelectronics Journal*, vol. 119, Article ID 105319, 2022.

[3] X. Liang, A. M. Javid, M. Skoglund, and S. Chatterjee, "Decentralized learning of randomization-based neural networks with centralized equivalence," *Applied Soft Computing*, vol. 115, Article ID 108030, 2022.

[4] M. Jian, J. Wang, H. Yu, and G.-G. Wang, "Integrating object proposal with attention networks for video saliency detection," *Information Sciences*, vol. 576, pp. 819–830, 2021.

[5] F. B. Chen, X. L. Wang, X. Li, Z. R. Shu, and K. Zhou, "Prediction of wind pressures on tall buildings using wavelet neural network," *Journal of Building Engineering*, vol. 46, Article ID 103674, 2022.

[6] J. Pilarz, I. Polishuk, and M. Chorążewski, "Prediction of sound velocity for selected ionic liquids using a multilayer feed-forward neural network," *Journal of Molecular Liquids*, vol. 347, Article ID 118376, 2022.

[7] A. Borji and L. Itti, "State-of-the-Art in Visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[8] M. Carrasco, "Visual attention: the past 25 years," *Vision Research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[9] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: bottom-up versus top-down," *Current biology : CB*, vol. 14, no. 19, pp. R850–R852, 2004.

[10] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[11] T. Liu, J. Sun, N. N. Zheng, X. O. Tang, and H. Y. Shum, "Learning to detect a salient objec," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, Minnesota, June 2007.

[12] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, Miami, Florida, June 2009.

[13] Q. Yan, L. Xu, J. P. Shi, and J. Y. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1155–1162, Long Beach, California, June 2013.

[14] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proceedings of the 12th European conference on Computer Vision*, pp. 29–42, Florence, Italy, October 2012.

[15] C. Yang, L. H. Zhang, H. C. Lu, X. Ruan, and M. H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173, Portland, Oregon, USA, June 2013.

[16] B. W. Jiang, L. H. Zhang, H. C. Lu, C. Yang, and M. H. Yang, "Saliency detection via absorbing Markov chain," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1165–1672, Sydney, NSW, Australia, April 2013.

[17] Y. Qin, H. C. Lu, Y. Q. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 110–119, Boston, MA, USA, June 2015.

[18] J. M. Zhang, S. Sclaroff, Z. Lin, X. H. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80fps," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1404–1412, Santiago, Chile, December 2015.

[19] W. C. Tu, S. F. He, Q. X. Yang, and S. Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2334–2342, Las Vegas, Nevada, USA, June 2016.

[20] Z. Chen, C. Wu, Z. Huang et al., "Dangerous driving behavior detection using video-extracted vehicle trajectory histograms," *Journal of Intelligent Transportation Systems*, vol. 21, no. 5, pp. 409–421, 2017.

[21] P. P. Zhang, D. Wang, H. C. Lu, H. Y. Wang, and X. Ruan, "Amulet: aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE Int'l Conf. on Computer Vision*, pp. 202–211, USA, 2017.

[22] W. G. Wang, J. B. Shen, X. P. Dong, and A. Borji, "Salient object driven by fixation prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1711–1720, USA, June 2018.

[23] X. Wang, Y. Zhang, and C. Ning, "A novel visual saliency detection method for infrared video sequences," *Infrared Physics & Technology*, vol. 87, pp. 91–103, 2017.

[24] M. Li, E. Zhang, L. Wu, and J. Duan, "A multistage and multiresolution deep convolutional neural network for inverse halftoning," *Expert Systems with Applications*, vol. 191, Article ID 116358, 2022.

[25] Xu Zhang, Z Yuanyuan, and L Shaoyuan, "Bayesian neural network with efficient priors for online quality prediction," *Digital Chemical Engineering*, vol. 2, Article ID 100008, 2021.

[26] B. S. Kronheim, M. P. Kuchera, and H. B. Prosper, "TensorBNN: Bayesian inference for neural networks using TensorFlow," *Computer Physics Communications*, vol. 270, Article ID 108168, 2022.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.