**Title:**

# Improved Salient Object Detection Using Hybrid Convolution Recurrent Neural Network

**Author Details:**

Dr. Nalliyanna V. Kousik,
School of Computing Science and Engineering,
Galgotias University, NCR Delhi, India-201307
Mobile: +91-9787928813
**Email:** nvkousik@galgotiasuniversity.edu.in

Mr. Yuvaraj Natarajan,
Department of Computer Science and Engineering,
St. Peter's Institute of Higher Education and Research,
Tamil Nadu, India-600 054
Mobile: +91-7904565698
**Email:**yraj1989@gmail.com

Mr. RajanA. Raja,
Department of Electronics and Communication Engineering,
B.S. Abdur Rahman Crescent Institute of Science and Technology,
Tamil Nadu, India-600048.
Mobile: +91-9842766513
**Email:**arshathraja.ru@gmail.com

Dr. Suresh Kallam,
Department of Computing Science & Engineering,
Sree Vidyanikethan Engineering College, Tirupati, India-517102
Mobile: +91-9966322466
**Email:** sureshkallam@gmail.com

Dr. Rizwan Patan
Department of Computing Science & Engineering,
Velagapudi Ramakrishna Siddhartha Engineering College,
Vijayawada, India-520007
Mobile: +91-9700266476
**Email:** prizwan5@gmail.com

Prof. Amir H. Gandomi
Professor of Data Science,
Faculty of Engineering & Information Technology
University of Technology Sydney,
Ultimo, NSW 2007, Australia
**Email:** gandomi@uts.edu.au

# Improved Salient Object Detection Using Hybrid Convolution Recurrent Neural Network

NalliyannaV. Kousik[1], Yuvaraj Natarajan[2], R. Arshath Raja[3], Suresh Kallam[4], Rizwan Patan[5], Amir H. Gandomi[6,*]

*[1]School of Computing Science and Engineering, Galgotias University, Greater Noida, Gautam Budh Nagar, Uttar Pradesh, India.*

*[2]Research Scholar, Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Tamil Nadu, India.*

*[3]Research Scholar, Department of Electronics and Communication Engineering, B.S.AbdurRahman Crescent Institute of Science and Technology, Tamil Nadu, India.*

*[4]Department of Computing Science & Engineering, Sree Vidyanikethan Engineering College, Tirupati, India*

*[5]Department of Computing Science & Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India*

*[6]Faculty of Engineering & Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia*

Emails: nvkousik@galgotiasuniversity.edu.in, yraj1989@gmail.com, arshathraja.ru@gmail.com, sureshkallam@gmail.com, prizwan5@gmail.com, gandomi@uts.edu.au

**Abstract:** Salient object detection is a critical and active field that aims at the detection of objects in a video, however, it draws increased attention among researchers. With increasing dynamic video data, the performance of saliency object detection method has been degrading with conventional object detection methods. The challenges lie with blurry moving targets, rapid movement of objects and background occlusion or dynamic background change on foreground regions in video frames. Such challenges result in poor saliency detection. In this paper, we design a deep learning model to address the issues, which uses a novel framework by combining the idea of Convolutional Neural Network (CNN) with Recurrent Neural Network (RNN) for video saliency detection. The proposed method aims at developing a spatiotemporal model that exploits temporal, spatial and local constraint cues to achieve global optimization. The task of finding the salient objects in benchmark dynamic video datasets is then carried out by capturing the temporal, spatial and local constraint features with the Convolution Recurrent Neural Network (CRNN). The CRNN is evaluated on benchmark datasets against conventional video salient object detection methods in terms of precision, F-measure, mean absolute error (MAE) and computational load. The experiments reveal that the CRNN model achieves improved performance than other state-of-the-art saliency models in terms of increased speed and reduced computational load.

**Keywords:** Convolution Recurrent Neural Network, Salient Object Detection, Spatial Features, Temporal Features, Local Constraint Cues

## 1. Introduction

The video salient object detection emphasizes objects that draws attention in a video, and this has emerged as an active part of research due to its increasing attention on applications like object tracking, object segmentation, and action recognition.

Different video salient object detection Cheng et al. (2015), Zouet al. (2015), Li, G., & Yu, Y. (2016), Zhu et al. (2014), Li, G., & Yu, Y. (2015), Girshick, R. (2015) approaches have been developed so far on still image benchmark datasets. These methods underperform if they are applied to dynamic videos due to four different phenomenon: the first phenomenon is that the blurred vision of moving targets in video frames that degrades the performance of salient objects appearance. Secondly, occlusion of salient objects partially or totally by background regions in video frames. Thirdly, the rapid or arbitrary movement of objects determines poor saliency performance. Finally, rapid or arbitrary change of background regions during the salient object movement that causes difficulty in salient object extraction. These four phenomena ensure that the detection of salient objects is difficult in video frames.

It is well known that information from a moving object is essential and intuitive for supporting the salient video object estimation. Utmost all existing approaches calculate the salient object motion using optical flow since it is highly sensitive to variation in illumination and changes in localization. The localization and illumination frequently occur in videos, and this affects the stability of salient objection motion estimation. Specific methods operated on utilizing contour detection in fields of color frames and optical flow, however, it suffers from limited scalability in extracting the contour regions from the cluttered backgrounds. Hence, the exclusion of dynamically changing irrelevant background poses a significant difficulty in modeling a salient object motion estimation. Hence, it needs a complete address in the field of video saliency. On the other hand, the salient object detection suffers mostly from high computational load due to the accommodation of huge visual scenes. Further, it suffers mostly from time efficiency, which acts as a bottleneck for video saliency algorithm in detecting or estimating the salient objects.

To resolve the aforementioned limitation, spatiotemporal optimization is used that exploits spatial and temporal cues with a local constraint for optimizing the video saliency. This leads to object detection using a local constraint that supports the temporal and spatial cues to achieve the task of global optimization. In this regard, the proposed work aims at the usage of Objectness to analyse the motion energy of the foreground object in a video frame for evaluating the foreground region probability.

In existing approaches, the distinctive regions are detected using Convolutional neural networks (CNNs) Girshick, R. (2015), Ren et al. (2015), Li et al. (2016) Zhao et al. (2015), Long et al. (2015). However, it poses a serious problem due to the larger availability of labelled training data.

This makes the system more time-consuming and complex. Many deep learning models are introduced to address the concerns in video saliency detection; however, CNN (Girshick, R. (2015), Ren et al. (2015), Li et al. (2016) Zhao et al. (2015), Long et al. (2015)) is the most important of all. The convolutional layer, designed with multiple layers, is responsible for convolving the regions of local images independently. The responses from this layer are combined based on the region coordinates. The feature responses are then summarised using pooling layer, which uses a pooling kernel size and a fixed stride to process the responses. This confined setting does not take into account the adjacent regions during computation, which offers disadvantages. For instance, if the convolution or pooling is performed on the bottom right regions in an image, the features in this region remain with regardless of the appearance on the top left region. Hence, the contextual dependencies are not captured that leads to poor representation of an image.

To read all the regions in an image, the cost of the computation and usage of resources using CNNs may surmount, and it fails in tolerating the structure variance. An ideal network is pretended to have a boosting memory that should get a hold on the scanned regions of an image and its spatial correlations. The RNN serves such a purpose that models the dependencies among the regions in an image in a time sequence manner. The feedback connection in RNN and a hidden layer may retain its previous state inputs, and hence the correlations are found between the regions in an image, even if the sequences states are different. This has motivated the present study to develop a deep learning model to generate labelled training data.

This enables the video data to be accessed easily and generated rapidly closer to realistic video motion sequences.

The proposed method avoids the challenges of CNNs by replacing it with CRNN in many video processing applications with dynamic video saliency detection. The proposed method computes the maps of dynamic video saliency models by considering the temporal and spatial information of video frames. The CRNN video saliency model produces spatiotemporal saliency by exploring the dynamic and static video saliency information. The CRNN is adopted to predict the pixel-wise video saliency. The static saliency is exploited and encoded in CRNN learning stage through transfer and tuning of video frame classification Simonyan et al. (2014). The dynamic saliency is learned using labelled data that includes both natural and human-generated data via supervised learning. Finally, the detection process of static video saliency is integrated with the detection process of dynamic video saliency, and thus, it produces a final estimation of spatiotemporal video saliency. The deep learning video saliency model is considered to be computationally efficient than other video saliency models.

The proposed video saliency model is said to be effective and efficient that reduces computational load and time efficiency. These objective functions are achieved by capturing the temporal saliency via CRNN from the video frame pairs. The CRNN is designed using a novel technique that combines both Convolutional Neural Network (CNN) with Recurrent Neural Network (RNN). This architecture has reduced the pitfalls of analyzing a video frame in faster response time than conventional CNN. The design of CRNN is carried out in such a way that it reduces the computational load in analyzing the video frames with an aggregation of layers in both CNN and RNN. The reduction in computational load is the usage of separate modules for static and dynamic saliency, where module 1 (CNN) and module 2 (RNN), is used respectively. The proposed method is evaluated on FBMS dataset Long et al. (2015), and it shows accurate salient maps than existing methods at 26fps. Hence, it could be regarded that the proposed method is effective in terms of both accuracy and speed.

The main contribution of the work is threefold:

- The authors investigate the design of CNN with RNN to form CRNN for saliency prediction (pixel-wise) for video salient object detection in dynamic scenes.

- The authors propose a training scheme using video data generated synthetically from image datasets. It further encodes both dynamic and static video salient information into CRNN.

- The proposed method is effective and efficient than existing deep video saliency models over dynamic scenes.

List of the endeavour advances in the proposed field:

- We have utilized the advancement in Salient object detection with deep learning architecture including CNN and RNN.
- The CRNN video saliency model produces spatiotemporal saliency by exploring the dynamic and static video saliency information.
- The CRNN is adopted to predict the pixel-wise video saliency.
- The static saliency is exploited and encoded in CRNN learning stage through transfer and tuning of video frame classification.
- The dynamic saliency is learned using labelled data that includes both natural and human-generated data via supervised learning.
- We used two different modules to address the problem one is static and the other is dynamic that comprises of CNN and RNN to process each image, respectively.

The outline of the paper is given below: section 2 provides the related works. Section 3 discusses the proposed framework for video salient static and dynamic information. Section 4 provides the experimental results and discussions. Section 5 concludes the paper with future work.

## 2. Related works

In this section, we discuss various deep learning strategies adopted to improve video saliency detection. It is seen from several studies that CNNs are used for video saliency detection. One such method uses residual motion and pixel color values to extract the saliency features with normalized energy of residual motion map Chaabouni et al. (2019). The deep neural network with regression nets is also used to reduce time consumption Xi et al. (2019). A method in Wang et al. (2018) obtains both the spatial and temporal characteristics using deep learning hybrid spatiotemporal saliency feature extraction framework. A technique in Ding et al. (2019). uses CNN with color saliency network, depth saliency network and saliency fusion network for saliency detection in RGBD images and stereoscopic images.

Further, FCN with deep CNN referred Hoseinzad, & Haratizadeh (2019) is used for pixel-wise salient object detection that learns the multi-level features using different convolution layers (CLs), of CNNs Zhang (2019), Kasinathan et al. (2019). In Lin et al. (2019) CNN is used to construct the auto-encoder for video saliency detection. In Zhang et al. (2018), joint optimization is carried out for segment-level saliency prediction. However, most techniques used do not focus on both temporal and spatial characteristics, and no separate framework is provided for both. Hence, the proposed method concentrates on proposing a framework that extracts the temporal and spatial characteristics of moving objects in video saliency detection. The proposed work addressed the following issues:

- To resolve the limitation, spatiotemporal optimization is used that exploits spatial and temporal cues with a local constraint for optimizing the video saliency.
- This leads to object detection using a local constraint that supports the temporal and spatial cues to achieve the task of global optimization.
- The proposed work aims at the usage of Objectness to analyze the motion energy of the foreground object in a video frame for evaluating the foreground region probability.
- To address or mitigates the challenges, we segregated the framework into static and dynamic ones.
- The dataset obtained is set to address the four problems as it is elusive of blurring, occlusion, rapid movement and background changes.
- Yes, the test videos are known to challenge those aspects of the problem.
- The study is aimed to solve the 4 problems using the video datasets.

## 3. Proposed Method

In the proposed method, we adopt the process of training and testing of video saliency network using CRNN method. The approach is used to generate video data from image datasets, and the annotated video sequences are associated with video data to train the CRNN video saliency network. This section discusses the proposed CRNN that acts as a video saliency model.

The overview of CRNN is initially dealt before discussing the details of the proposed system. The video frames are fed to CRNN at a high level, and the network produces saliency maps successively in which pixels with bright values indicates higher values in video saliency network. The image and video sequence are used to train the CRNN, and in general, it is used for training the spatiotemporal saliency.

Figure 1 shows the CRNN architecture that operates as a video saliency model. Inspired by human visual perception Mital et al. (2013), the proposed study uses both dynamic signals and static signals of saliency that tends to contribute to video saliency. The proposed CRNN module is designed into two different modules that take into account the spatial characteristics and temporal characteristics of a visual scene.

In the first module, the static saliency feature is captured from the single image frame, which is regarded as the input. This module adopts CRNN to generate saliency estimate, and optimal pre-training modules are used for the saliency estimation on large datasets. This module is trained richly to acquire static video saliency information using the interesting object's data from image saliency benchmarks.

In the second module, the static saliency and video frame pair are used taken as input from the first module. It then generates dynamic video saliency features. This module is trained using real labelled and synthetic video data.

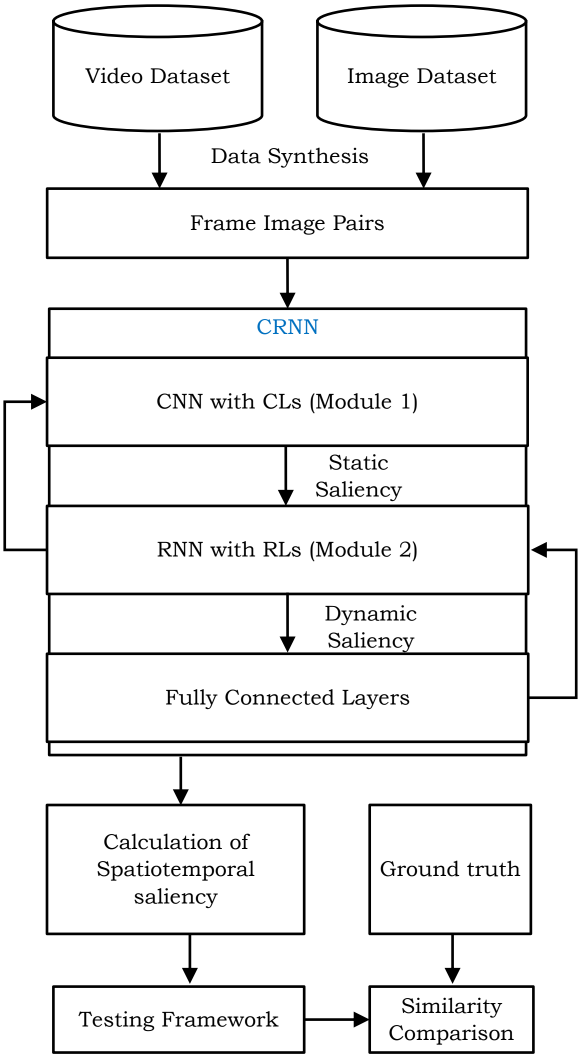


Figure 1: Proposed CRNN Spatiotemporal Framework

### 3.1. CRNN

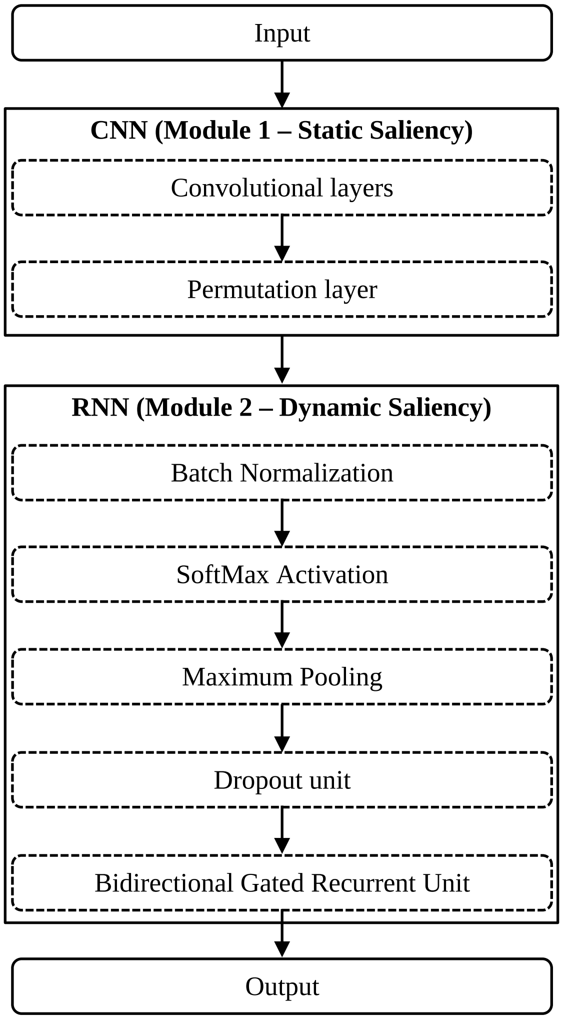


Figure 2: CRNN framework for video saliency

In Figure 2, CNN layers have first five layers that are of the convolutional and pooling layers. The CNN extracts the salient object region in both static and dynamic video saliency. Second is the RNN layer,

1) the CNN output is converted parallel into four different spatial sequences that use4 directional scanning's, namely, left-to-right, right-to-left, top-to-bottom, and bottom-to-top. A scanning window of varying size is used in each sequence, i.e. size of a row or a column. The scanning is continued on the non-overlapping window until the entire video frame sequence is covered.

2) The local temporal and spatial dependencies in each scanning window is focused on a sequential video frame segment (or feature vector), and the RNN shared unified weights are thus updated as W→ (left-to-right), W← (right-to-left), W↓ (top-to-bottom), and W↑ (bottom-to-top). The updated weights are processed sequentially on the window by window. Finally, FC layers are used to father the outputs of 4 directional RNN feature vector, and it relates to two fully connected layers.

CRNN has 5 CLs, one recurrent layer (RL), and two fully connected layers (FCLs). The CLs are used to learn the static and dynamic saliency of the videos, where the operation of which is like the first five layers of the seven-layered Alex-Net Krizhevsky et al. (2012). The RL is used to learn the temporal and spatial dependencies of video frames. The FCL finally gathers the output from RL, and then it learns the video saliency representation. For the purpose of classification, the N-way SoftMax layer is adopted.

### *3.1.1. Convolutional Neural Network*

From Figure 2, the left part uses 5 CLs to learn the static and dynamic saliency of the videos from the input video frames. With increasing convolutional layers, video object saliency in motion is extracted with more robustness. The 5th CL after training using ImageNet extracts the objects in motion from the moving foreground objects. Also, it extracts additionally the objects that are static in foreground objects of the input video frames. Depending on the static and dynamic saliency provided by the CLs, the temporal and spatial dependencies are learned appropriately by RLs.

- *Convolutional layers* perform the convolution operation, which is a linear operation with the multiplication of a set of weights with the array of input video data (i.e. pixels of the image from each video frame). A filter is used to detect the saliency from the input image by scanning the entire image, where this process is referred to as translation invariance.

Since the operation of filtering multiplied with input array on multiple times results in the formation of a two-dimensional feature map.

- *Permutation layer* confirms the saliency detected by the convolutional layer, where the permutation-based representation lists the reference objects in order. The similarity is estimated between the reference object and the ground truth using the two corresponding permutations rather than using a distance function. It tends to provide more closest relation than the one represented using the distance function.

### *3.1.2. Recurrent Neural Network*

With the output (reference saliency or objects) obtained from 5th CL, the RLs from RNN is built to learn the temporal and spatial video saliency dependencies in video frames. The RNN performs the regular operations that include:

- *Batch normalization* reduces the number of hidden layers required for processing the input features from CNN. It adjusts the scaling and activation function to speed up the learning process.
- *SoftMax Activation* outputs a vector that represents the probability distributions of a list of potential outcomes.
- *Maximum pooling* is a pooling operation that calculates the maximum or largest value in each patch of each feature map, where it is processed by convolutional and permutation layer.
- *Dropout unit* ignores a certain number of hidden units during the forward pass, where the individual nodes are dropped with probability 1-*p* such that a reduced network is left;

incoming and outgoing edges to a dropped-out node are also removed on the other hand, it may also be added. These operations are carried out to avoid overfitting.

- Bidirectional Gated Recurrent Unit is used to resolve the gradient problem and makes the operates faster and efficient.

The video frames in RNN are converted first to one dimension image sequences of regions through directional scanning of a two-dimensional video frame. The sequence of regions is then utilized for training the RLs to learn the temporal and spatial video saliency objects in motion.

The use of feedback loops increases the learning strategy since the network can remember past data. However, this leads to a limitation of memory space since it increases with sufficient video data. The relevant functions are updated iteratively, and the past sequences are remembered ideally. This proves meaningful observation between the past and present video saliency objects. The relationship between the regions in the video frames at different temporal and spatial positions are thus learned using RLs, and these dependencies are called as temporal and spatial dependencies, respectively.

### *3.1.3. Fully Connected Layers*

Other than general RNN, the proposed system does not use any intermediate labels, rather it uses video frame labels. The study uses FCLs to collect the data from hidden units of RLs, and it is then connected with the final image label. The study uses 2 FCL and one SoftMax layer like Alex-Net Krizhevskyet al. (2012).

The backpropagation is used to transmit the global representation from FCL to RLs for improved learning that improves the temporal and spatial video saliency. Similarly, RLs sends the output feedback back to CLs to learn the static and dynamic video salient objects in motion.

In case of rapid motions for video saliency detection, various shapes of the target are considered challenging during the saliency detection. Hence, the saliency framework using CRNN is provided with paths over each frame to reach the target. The saliency segments the interested target regions based on the above process, and that locates the objects in motion. The saliency detection of objects in motion is shown in the framework of Figure 3. In figure 4 results of a salient motion object detection using the CRNN training stage framework, based on the input video datasets, the frames vary dynamically. In comparison, with state of the art methods, the static and dynamic features are processed effectively to limit all four problems. It is further understood from the study that videos lesser than 240p frames produces inaccurate saliency or object detection, which is the core limitation of the work.

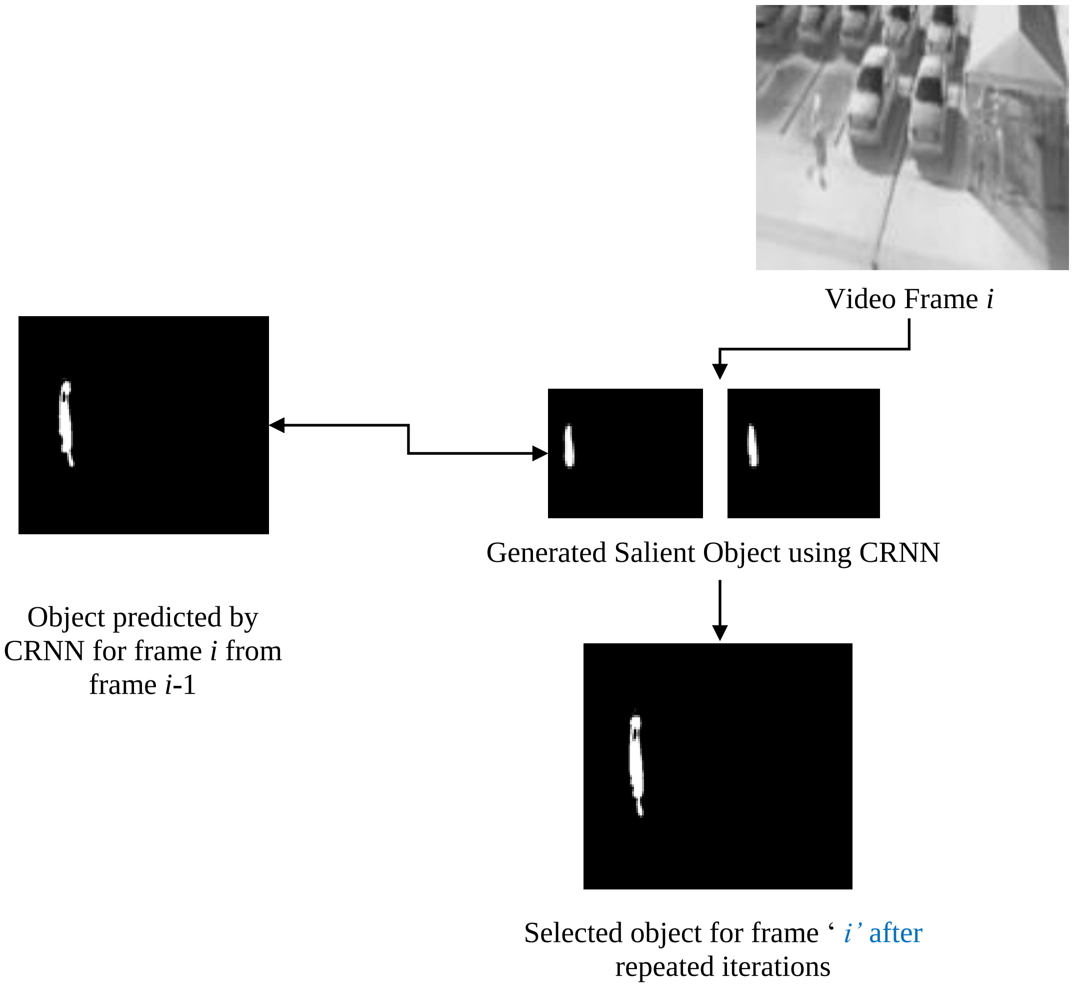
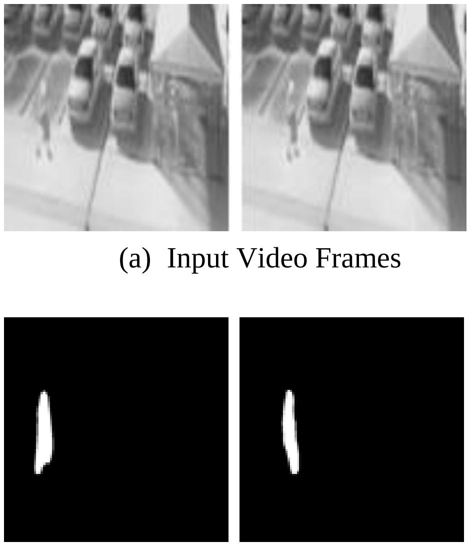
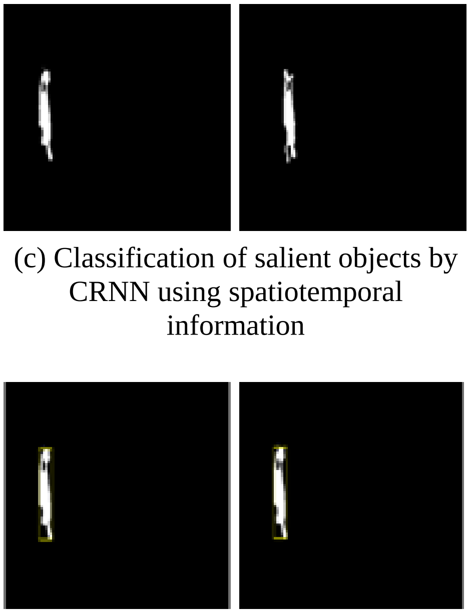

Video Frame *i*
Object predicted by CRNN for frame *i* from frame *i*-1
Generated Salient Object using CRNN
Selected object for frame ‘ *i’* after repeated iterations

Figure 3: Framework of salient motion object detection using CRNN training stage framework


(a) Input Video Frames

(b) Initial classification of salient objects using CRNN



(c) Classification of salient objects by CRNN using spatiotemporal information

(d) The final classification of salient objects using CRNN training after repeated iterations

Figure 4: Results of a salient motion object detection using CRNN training stage framework using an input video frame

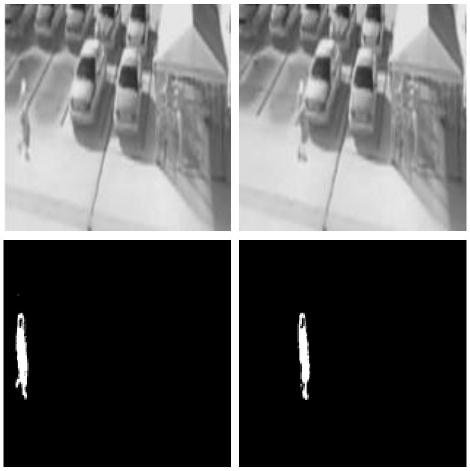
## 4. Results and Discussions

The proposed method is evaluated on INO analytics dataset Mangale, S., & Khambete, M. (2015), and it shows accurate salient maps than existing methods at 26fps. The proposed method is compared with existing CNN and local estimation and global search by deep Network (LEGS) Wang et al. (2015). The reason for choosing LEGS method is that it obtains the spatial and temporal characteristics using deep learning hybrid spatiotemporal saliency feature extraction framework, which is more similar to the present study.

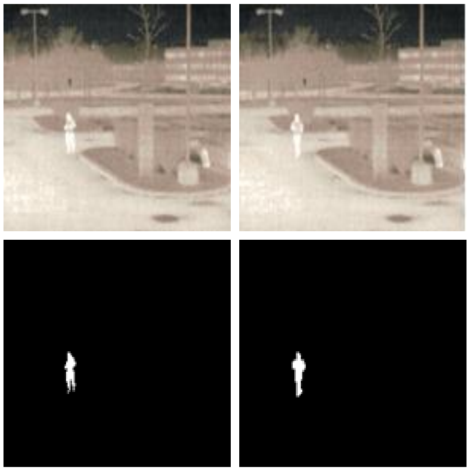
The simulation is carried out on Nvidia Geforce TITAN X GPU and Intel Xeon E7 12 cores CPU with 64GB memory.

### 4.1.Qualitative Analysis

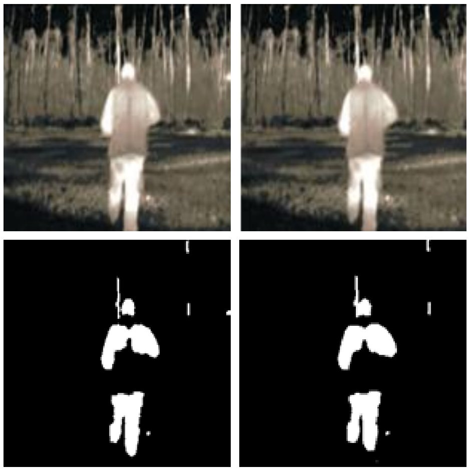
Qualitative results of the proposed system are given in this section, where the first row shows the original video frames and the second row shows the classified video frames. Figure 5(a) – Figure 5(h) shows the results of classified output using CRNN. The first row shows the original video frames and the second row shows the classified results using CRNN.



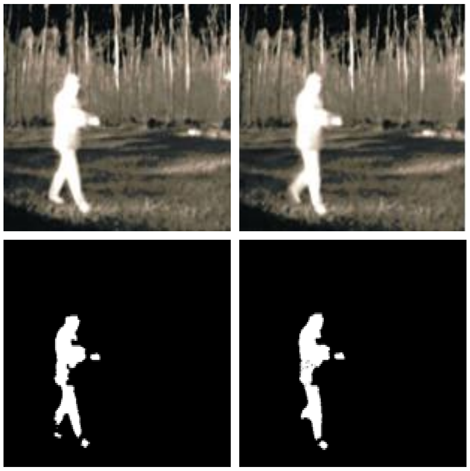
(a) Video saliency detection using CRNN on a close person (cp) dataset on various frames



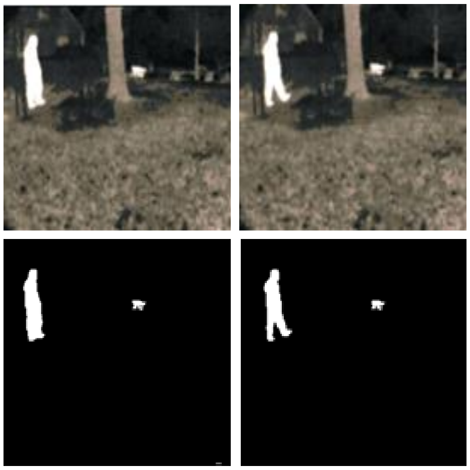
(b) Video saliency detection using CRNN on trees and runners (tr) dataset on various frames



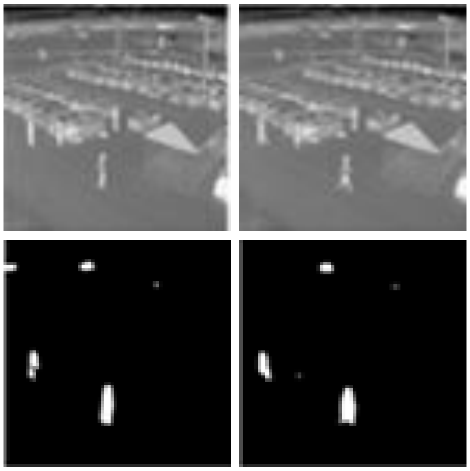
(c) Video saliency detection using CRNN on irw1 dataset on various frames



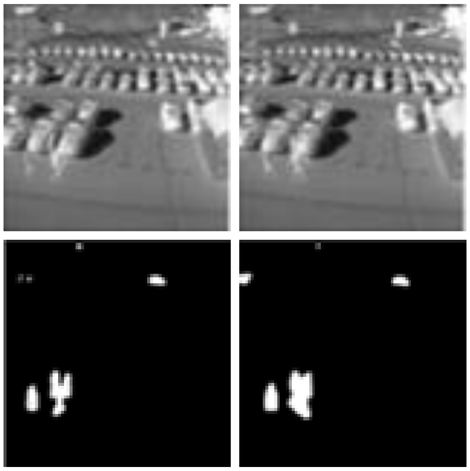
(d) Video saliency detection using CRNN on irw2 dataset on various frames



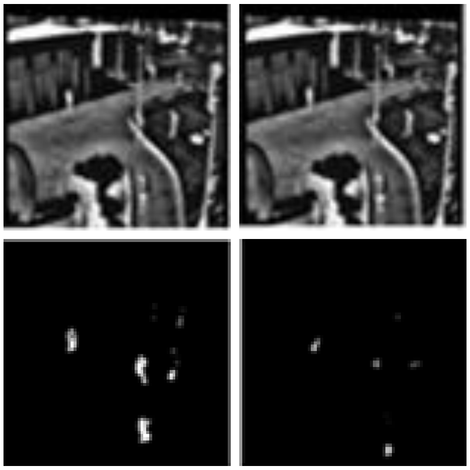
(e) Video saliency detection using CRNN on irw3 dataset on various frames



(f) Video saliency detection using CRNN on parking snow(ps) dataset on various frames



(g) Video saliency detection using CRNN on group fight(gf) dataset on various frames



(h) Video saliency detection using CRNN on OCTBVS 2a video dataset on various frames

Figure 5: Classification results of single/multiple video salient objects with actual video frames on the first row and classified video frames on the second row

Further, the motion estimation results are shown in Figure 6 and Figure 7, where the results of tracking a single salient object are given in Figure 6, and multiple salient objects are given in Figure 7. The CRNN is used accurately to detect the salient objects, and by repeated learning using the training labelled datasets, the proposed method classifies or detects the salient objects accurately. Figures 6 Figure 7 show that the CRNN accurately detects the salient objects with reduced error than other methods, which is expressed in terms of precision and recall results.

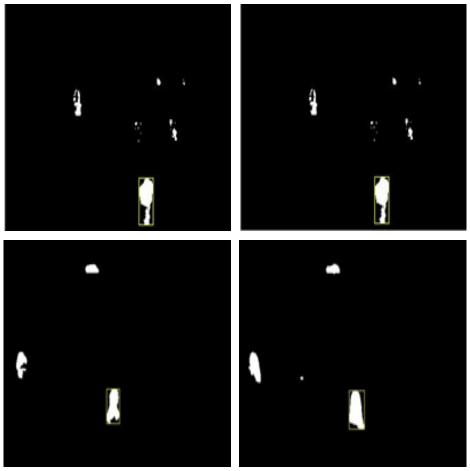


Figure 6: Tracking results of a single object

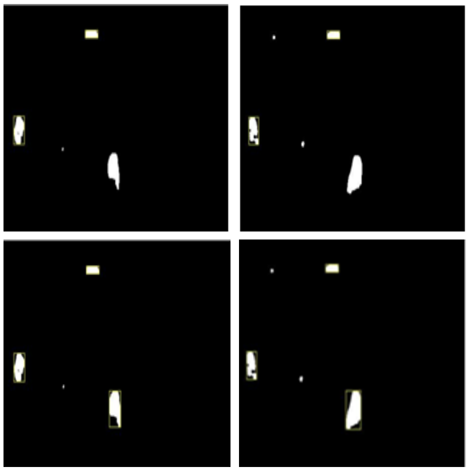


Figure 7: Tracking the results of multiple objects

The image saliency using CRNN avoids difficulties like lack of information from inter-frames and hand-crafter feature. The CRNN effectively captures the salient objects in a better way, thus proving the power of CRNN in saliency detection. The proposed study considers both dynamic and static saliency information to capture the foreground salient objects in all test cases. This even makes CRNN detect the moving salient objects correctly.

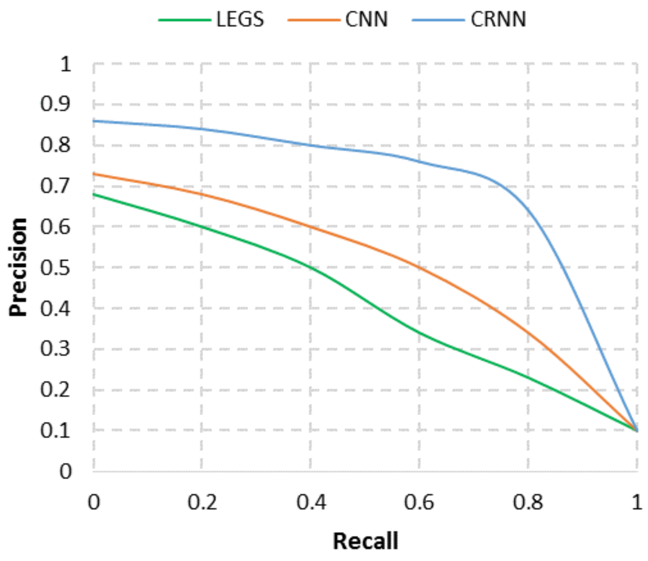
### 4.2. Quantitative Results

The quantitative evaluation is carried out in terms of three different measures, namely precision-recall (PR) curves, F-measure and MAE.

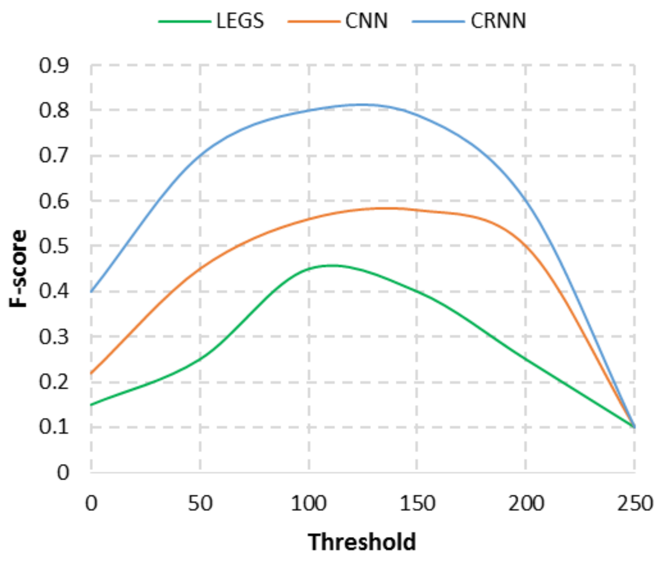
Initially, precision-recall (PR) curves are employed for evaluation. Precision is defined as the percentage of correctly assigned salient pixels, and recall is defined as the fraction of correctly detected salient pixels in comparison with ground truth salient pixels. In this proposed study, we calculate overall PR curves that take into account 10 test cases, i.e. different motion INO video dataset. Further, we present the F-measure sequence values along the graph of PR-curves with the threshold range varied between 0 and 255. The mean absolute error (MAE) is used as another measure to consider true negative saliency assignments. MAE is defined as the average per-pixel difference between detected saliency probability map and ground truth map. The MAE is averaged across all the video frames and then with all the videos in the test set.

Figure 8 (a) shows the PR curves between proposed and existing deep learning models for video saliency detection. Figure 8 (b) shows the F-measure between proposed and existing deep learning models for video saliency detection. Figure 9 (a) shows the MAE between proposed and existing deep learning models for video saliency detection.

The result shows that the proposed method outperforms other methods on INO video dataset. Our video saliency CRNN method attains improved precision rates than existing methods that demonstrate the saliency maps obtained by the proposed method is precise, and it is considered to be responsive to an actual video saliency information. The F-score achieves a better rate than other exiting methods. It is further concluded that the proposed method has the lowest MAE than other methods.

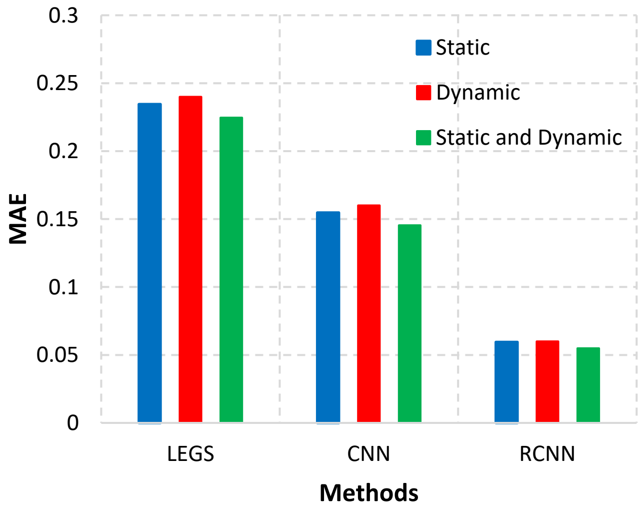


a) PR curves

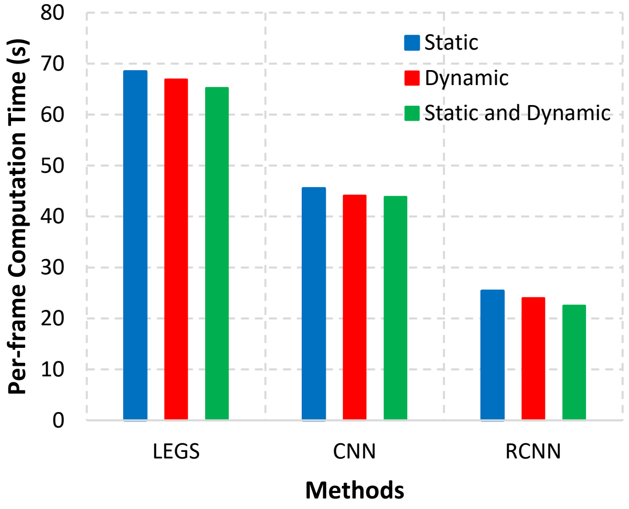


b) F-measure

Figure 8: a) PR curves and b) F-measure two are proposed towards existing deep learning models for video saliency detection



a) MAE



b) Computational Load

Figure 9: a) MAE and b) Computational Load are two proposed models over existing deep learning models for video saliency detection

For quantitative analysis, we found that the fusion of static and dynamic model increases the performance of saliency detection. The existing method fails in the detection of video object saliency with the inclusion of either static saliency network or dynamic saliency network. The static saliency information is used for training the dynamic model that increases the accuracy of prediction than the other methods. It is concluded from the above study that with the reduction of training data, the performance reduces and vice versa, which shows that the proposed method using CRNN is a data-driven model.

The computational load is measured between proposed and other methods, as shown in Figure 9 (b); itis seen that the speed of operation using the proposed method is higher than those of other methods. It is seen that the proposed method avoids the major bottleneck of run time efficiency with reduced computational time. Most of the time, the video saliency suffers from the computation on edge or motion information. The results are given in Figure 9 that includes static dynamic and both static and dynamic between the proposed and other existing methods. It is seen that the proposed method with the static and dynamic model has reduced MAE and computational cost than static or dynamic ones, and with other conventional methods. The results of reduced computational time are due to the reduced number of networks in CNN to process the input data, where RNN eventually carries out the processes that are to be handled by CNN. However, this remains as increasing complexity in existing CNNs, which consumes more time for getting the required output. The operations carried out by convolution layer in CNN has been reduced, and a fair output is sent by CNN to be processed by RNNs.

## 5. Conclusions

In this paper, a CRNN spatiotemporal optimization is proposed for video saliency object detection. The study offers improved saliency detection with CRNN that removes the potential backgrounds effectively. The valid regions are detected with an objectness measure that supports saliency propagation. This video saliency model is effective in capturing the temporal saliency from the frame pairs, which reduces the computational load and improves the time efficiency on dynamic video scenes. The study further reports that the static video saliency offers increased computational load than the dynamic, and static and dynamic modes, but the MAE is lesser in static than in dynamic saliency. This is due to the fact that with increasing network size for the detection of static and dynamic saliency, a downgrade is reported with increasing threshold range. It is finally noticed that the proposed method is effective in terms of both accuracy and speed than the existing deep video saliency models over static and dynamic scenes. The increased network, in turn, increases the computational load that degrades the F-score and possibly affects the system performance. Further limitation includes the marginal reduction of MAE and computational load than CRNN with static or dynamic video saliency separately.

## 6. Future Work

The proposed model can improve and may carry these changes towards future developments,

- The CRNN can be improved with a reduced number of networks in its hidden layer that should not possibly increase the computational load.
- Optical flow estimation may be integrated with CRNN internal structure that can increase the speed of operations of individual layers. This possibly an reduce the number of layers required to perform the deep learning operation.
- We would extend the CRNN learning to deep CRNN that enables the hidden layers to learn complex dependencies. In such cases, the number of hidden layers should lesser than the actual ones.
- The CRNN is limited by the use of limited video dataset for testing since the study has produced a limited number of labels during the training phase. In future, the researchers may tend to work on large videos with increased resolution and size, such that it generates more labelled data and improves the pattern of matching or identifying the saliency in videos.

## References

Chaabouni, S., Benois-Pineau, J., & Amar, C. B. (2019). Chabonet: Design of a deep cnn for prediction of visual saliency in natural video. *Journal of Visual Communication and Image Representation*, *60*, 79-93.

Chen, Z., Xu, Z., Yi, W., Yang, X., Hou, W., Ding, M., &Granichin, O. (2019). Real-time and multimodal brain slice-to-volume registration using CNN. Expert Systems with Applications, 133, 86-96.

Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. M. (2015). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(3), 569-582.

Ding, Y., Liu, Z., Huang, M., Shi, R., & Wang, X. (2019). Depth-aware saliency detection using convolutional neural networks. *Journal of Visual Communication and Image Representation*.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

Hoseinzade, E., & Haratizadeh, S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Systems with Applications, 129, 273-285

Kasinathan, G., Jayakumar, S., Gandomi, A. H., Ramachandran, M., Fong, S. J., &Patan, R. (2019). Automated 3-D Lung Tumor Detection and Classification by an Active Contour Model and CNN Classifier. Expert Systems with Applications. 134, 112-119.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5455-5463).

Li, G., & Yu, Y. (2016). Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 478-487).

Lin, X., Tang, Y., Tianfield, H., Qian, F., & Zhong, W. (2019). A Novel Approach to Reconstruction based Saliency Detection via Convolutional Neural Network Stacked with Auto-encoder. *Neurocomputing*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Mangale, S., & Khambete, M. (2015, May). Moving object detection using visible spectrum imaging and thermal imaging. In *2015 International Conference on Industrial Instrumentation and Control (ICIC)* (pp. 590-593). IEEE.

Mital, P., Smith, T. J., Luke, S., & Henderson, J. (2013). Do low-level visual features have a causal influence on gaze during dynamic scene viewing?. *Journal of Vision*, *13*(9), 144-144.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wang, L., Lu, H., Ruan, X., & Yang, M. H. (2015). Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3183-3192).

Wang, Z., Ren, J., Zhang, D., Sun, M., & Jiang, J. (2018). A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. *Neurocomputing*, *287*, 68-83.

Xi, X., Luo, Y., Wang, P., & Qiao, H. (2019). Salient object detection based on an efficient End-to-End Saliency Regression Network. *Neurocomputing*, *323*, 265-276.

Zhang, L., Fang, X., Bo, H., Wang, T., & Lu, H. (2018). Deep multi-level networks with multi-task learning for saliency detection. *Neurocomputing*, *312*, 229-238.

Zhang, Q., Lin, J., Zhuge, J., & Yuan, W. (2019). Multi-level and Multi-scale Deep Saliency Network for Salient Object Detection. *Journal of Visual Communication and Image Representation*, 59, 415-424.

Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1265-1274).

Zhu, W., Liang, S., Wei, Y., & Sun, J. (2014). Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2814-2821).

Zou, W., Liu, Z., Kpalma, K., Ronsin, J., Zhao, Y., &Komodakis, N. (2015). Unsupervised joint salient region detection and object segmentation. *IEEE Transactions on Image Processing*, *24*(11), 3858-3873.