Visualization of Salient Object With Saliency Maps Using Residual Neural Networks

RUBINA RASHID¹, SAQIB UBAID², MUHAMMAD IDREES³, RIDA RAFI⁴, AND IMRAN SARWAR BAJWA⁵

¹Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
²Department of Computer Science and Engineering, Khawaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
³Department of Computer Science and Engineering, UET Lahore, Lahore 39161, Pakistan
⁴Department of Computer Science and Engineering, University of Sargodha, Lahore 40100, Pakistan

Corresponding author: Imran Sarwar Bajwa (imran.sarwar@iub.edu.pk)

ABSTRACT Visual saliency techniques based on Convolutional Neural Networks (CNNs) exhibit an excessive performance for saliency fixation in a scene, but it is harder to train a network in view of their complexity. The imparting Residual Network Model (ResNet) that is more capable to optimize features for predicting salient area in the form of saliency maps within the images. To get saliency maps, an amalgamated framework is presented that contains two streams of Residual Network Model (ResNet-50). Each stream of Reset-50 that is used to enhance the low-level and high-level semantics features and build a network of 99 layers at two different image scales for generating the normal saliency attention. This model is trained with transfer learning for initialization that is pretrained on ImageNet for object detection, and with some modifications to minimize prediction error. At the end, the two streams integrate the features by fusion at low and high scale dimensions of images. This model is fine-tuned on four commonly used datasets and examines both qualitative and quantitative evaluation metrics for state-of-the-art deep saliency model outcomes.

INDEX TERMS Saliency maps, CNNs, ResNet-50, global semantics features.
saliency systems may have limited power while using CNN, it takes so much time to train a desired model in CNN, so the detection but difficult to train due to its complexity. Normally, to reformulate the layers as gaining knowledge for image neural networks (CNNs) have a sequence of breakthroughs related background as shown in Fig. 2. Although, convolutional images due to different viewpoints (camera viewpoint) illu- tions and high image dimensions for getting better diversity and high image dimensions for getting better saliency. The robustness of the saliency framework can be enhanced by using these key features.

Thus, we suggest a two-modality framework to get conceptual components from crude image pixels progressively, which has richer prior information for a better saliency prediction as this model has learned and how to identify images from ImageNet [33] dataset. It is a case of transfer learning where features are learned on one job and reused for another with or without fine-tuning. The transfer learning paradigm is considered important typically for smaller saliency datasets [21]. For image identification, Residual Network Model (ResNet) [9, 43]. Commonly, visual saliency models use a multiscale configuration for improving accuracy, which integrates the information at low and high image scales [11]. This improves the saliency detection performance of our model, which finds out the tiny salient regions and the center of large salient regions in high and low scales, respectively [11].

In addition, there are various prediction models that make saliency maps based on: the probability distribution of the position of the eye fixation on the image [11], low-level features such as multiscale contrast, color spatial distribution to describe a salient object locally and regionally, high-level features such as “objectness information” [2]. Since then, models of saliency have emerged to fixate the most prominent regions by snubbing the less significant part, but still there are many opportunities to get better due to its complexity, having many different object types, having large dissimilarity of multiple objects in a scene [12], variations exist in images due to different viewpoints (camera viewpoint) illum- inations, different object pose, partial occlusions and unre- lated background as shown in Fig. 2. Although, convolutional neural networks (CNNs) have a sequence of breakthroughs to reformulate the layers as gaining knowledge for image detection but difficult to train due to its complexity. Normally, it takes so much time to train a desired model in CNN, so the saliency systems may have limited power while using CNN when known and obvious objects are not present within the image [2]. For the solution of this problem, the Residual Network Model (ResNet) [9], one of the deep CNN models, is used to carry strong semantic features within the image. In addition to this, a feature significantly describes the particular attribute of the object, some commonly used features are size, color, and shape. The primary objective is to process a saliency outline geographically to the level of saliency for visual consideration. Thus, we suggest a two-modality framework to get conceptual components from crude image pixels progressively, which has richer prior information for a better saliency prediction as this model has learned and how to identify images from ImageNet [33] dataset. It is a case of transfer learning where features are learned on one job and reused for another with or without fine-tuning. The transfer learning paradigm is considered important typically for smaller saliency datasets [21]. For image identification, Residual Network Model (ResNet) [9, 43]. Commonly, visual saliency models use a multiscale configuration for improving accuracy, which integrates the information at low and high image scales [11]. This improves the saliency detection performance of our model, which finds out the tiny salient regions and the center of large salient regions in high and low scales, respectively [11].

In addition, there are various prediction models that make saliency maps based on: the probability distribution of the position of the eye fixation on the image [11], low-level features such as multiscale contrast, color spatial distribution to describe a salient object locally and regionally, high-level features such as “objectness information” [2]. Since then, models of saliency have emerged to fixate the most prominent regions by snubbing the less significant part, but still there are many opportunities to get better due to its complexity, having many different object types, having large dissimilarity of multiple objects in a scene [12], variations exist in images due to different viewpoints (camera viewpoint) illum- inations, different object pose, partial occlusions and unre- lated background as shown in Fig. 2. Although, convolutional neural networks (CNNs) have a sequence of breakthroughs to reformulate the layers as gaining knowledge for image detection but difficult to train due to its complexity. Normally, it takes so much time to train a desired model in CNN, so the saliency systems may have limited power while using CNN

FIGURE 2. Show different variation in images (a) Cluttering, (b) Camera view point, (c) Illumination, (d) Occlusion, (e) Pose, (f) Scaling, (g) Unrelated background.
design of our visual saliency model. Section IV mentions the
details of model training and proceeds with the investigation
of our model evaluation. Section V discusses the final results.
Finally, we end up with conclusions in Section VI.

II. RELATED WORK
The most obligate goal is to discuss the most recent research
strategies about CNNs saliency models that foresee the like-
lihood circulation area of the eye prediction over the image.
Saliency maps have different intensity for each pixel and each
pixel has its place on the most salient object. In spite of the
fact that these strategies accomplish better execution than
conventional models depending on visual saliency. In [2],
Jia et al. made an improved saliency method with multiple
layers of CNN to study visual elements named as EML-NET
that acquired encouraging results after merging the com-
fortable prior information, which discovered the results of
convolution by means of CNN model on the comprehensive
saliency dataset. It can be utilized further to expand the scal-
ability performance that turns into more thought-provoking
for getting features from several layers. In [3], the researchers
have proposed a framework and built on two equally trained
CNN models, one trained model was generated for top-down
visual saliency, and the other trained model was exploited
for classification. In addition, the authors collected the eye
look map dataset by means of Tobbi T60 visual tracker and
evaluated the performance in two forms: visual map and
enhanced classification accuracy. Furthermore, a comparison
has been shown between Inception VGG-19 and SalClassNet
classifiers.

In [1], Feng et al. computed a comprehensive spontaneous
CNN architecture that captured the global and local contrast
features information based on different scales, which could
successfully spot the salient region within the images. More-
over, comparative results with ten state-of-the-art architec-
tures have been exposed. In [4], the authors made a design:
to extract multifarious semantic features, to study end-to-
end pixel-wise visual saliency at different scales while con-
sidering only the global perspective by utilizing link layers
through large receptive fields. In addition, key factors were
included: massive deepness, dissimilar size, kernels working
in parallel to pinpoint the saliency, greater receptive fields
for global context, center bias for pattern outline identification
reliant on location. The proposed network in [8] contained
two end-to-end CNN fixation streams, one stream was pre-
trained on human visual guesstimate on eye tracking data and
the other was pretrained on an image identification dataset
named as semantic stream which was figured out semantic
signals from the input images. Furthermore, these two CNN
streams merged to form a module like inception block with
convolution and deconvolution layers to notice a complex
prominent element. The authors in [13] presented a deep
CNN based attentional push approach for saliency prediction.
This model contained two pathways: a saliency way fed by
the whole image to implant fixation method for computing
the augmented maps, a push way fed by 2-D cropped actor
head image to guess the gaze scene actors. Followed by a
trivial conven that merged and generated the saliency. In [15],
a ConvLSTM model built on LSTM that iteratively fixated
different locations in the images to refine feature predic-
tion and learnt the earlier saliency maps made by Gaussian-
function. In addition to this, ConvLSTM learnt center bias
without mixing the prediction features manually.

In [17], Cao et al. establish a simple and end-to-end CNN
network that identifies input features with fewer para-
ters for the production of visual saliency maps. Moreover,
the authors performed widespread experiments for the selec-
tion of high quality features at low layer, middle layer,
and high layer. The major motivation of this design was to
select input features that enabled the network to improve
the results and showed major similarities with the contrast
evidence which was presented in ground truth masks. In [23],
Ji et al. suggest a new encoder decoder CNN framework by
acquainted multidimensions spatial-wise and channel-wise
devoted layers. These attention layers united the perspective
information related to features at varying scales and then
finally produced the saliency. In addition to this, the structure
was designed to get visual saliency maps with accurate side
way edge information. This structure showed effective results
on various datasets. In [25], Monroy et al. extend CNN
architecture by transfer learning to predict the 2D and omni-
directional images (ODIs) saliency in an accurate manner.
In this pipeline, the generated visual maps were very close
to the ground truth.

However, in [34], the authors suggested a novel way to
extend the 2D prediction method by applying on cube face
images for 360degree images. This method used CNN based
fusion approach which has been trained on CMP images and
a new loss function. In [30], the researchers introduced two
new datasets: one was odd one-out (O^3) images, the second
was a psychophysical pattern (P^3). These two datasets were
used to evaluate the capacity of visual saliency algorithms for
finding single target. Furthermore, the effect of an architec-
ture based on CNN saliency model training was investigated
on these forms of datasets and did not find an odd one-
out target ability for major improvement. A pyramid feature
attention network (PFAN) was proposed in [38] that enhanced
the contextual and spatial features by using a novel CNN.
It contained four modules named as: a context-aware pyra-
mid feature extraction (CPFE), channel-wise attention (CA),
spatial attention (SA), and edge preservation (EP). CPFE was
designed to acquire rich context features at the multiscale
level, but other modules were utilized to generate feature
maps and then applied fusion for the saliency regions.

Moreover, in [28], the authors presented a novel deep
spatial contextual computational saliency method named as
deep spatial contextual long-term recurrent convolutional net-
work (DSCLRCN) that inevitably learns local features from
the input images in parallel. Then, it acquired long-term spa-
tial interactions between global context and the overall scene
context to conclude the saliency maps. In [29], the authors
proposed a multiresolution convolutional neural network
TABLE 1. A comparison study of CNNs models used for visual saliency, which are identified as a benchmark and developed recently.

| Paper | Year | Model | N-Layers | Datasets | Loss function | Evaluation Metrics |
|-------|------|-------|----------|----------|---------------|-------------------|
| MGCC[1] | 2020 | VGG-16 | 4*3+4+1 | ECSSD, HKU-IS, PASCAL-S DUT-OMRON | Binary cross entropy | Precision-recall (P-R) curves, F-measure, Mean Absolute Error (MAE) |
| Two stream fixation-semantici[8] | 2017 | CNN | 9+(2*2)+(3*3)+2+1 | DUT-OMRON, PASCAL-S, ECSSD, HKU-IS | Cross-entropy loss | Max F-Measure, Mean Absolute Error (MAE) |
| EML-NET[2] | 2020 | CNN | - | SALICON, MIT1003, MIT300, CAT2000 | Combined loss | CC, NSS, AUC-Judd, SiM, EMD, AUC-Borji, sAUC, KLD |
| Attentional Push[13] | 2017 | CNN/VGG G(16) | 14+4 | GazeFollow, SALICON, Fixation and non-fixation locations with multiresolution and utilized input images as raw pixels. At the end, Top-down and bottom-up features were integrated in the last layer to predict visual saliency. | GazeII in [21] used the initial deep features from the VGG-19 model, which was used for the image identification model. It trained on fixation and non-fixation locations with multiresolution and utilized input images as raw pixels. At the end, Top-down and bottom-up features were integrated in the last layer to predict visual saliency. | Euclidean loss | AUC, NSS, CC |
| Multi-scale attention (MCA-CNN)[23] | 2018 | CNN | - | ECSSD, PASCAL-S, DUT-OMRON, HKU-IS | Binary cross entropy | Precision-recall curve, MAE, F-measures |
| SalClustNet[3] | 2018 | CNN | 13+1 | SalDogs, POET, SalBirds, SalFleurs, CUB-200-2011, Oxford Flower-102 | Cross-entropy loss | NSS, s-AUC, CC |
| Saliency Attentive Model (SAM)[15] | 2018 | LSTM, VGG G-16, ResNet-50 | 16+50 | SALICON, MIT300, CAT2000 | Combination of multiple saliency metrics | AUC, CC, NSS, sAUC |
| GNet, SNet[17] | 2018 | CNN | 16+3 | HKU-IS, MSRA, ECSSD | Squared loss | MaxF, AveF, MAE |
| SalNet360[25] | 2018 | CNN | 74 | SALICON | Euclidean Loss | KL, CC, NSS, AUC |
| DeepFix[4] | 2017 | CNN | (2*2)+(3*3)+1 | SALICON, CAT2000, MIT300, PASCAL-S, ONE, PIFRIM | Euclidean loss | EMD, NSS, CC, AUC Judd, AUC-Borji, sAUC |
| DeSaliency Models Detect Odd-One-Out Target[30] | 2020 | CNN | - | Psychophysical patterns (P), Odd-One-Out (O), SALICON | Global saliency index (GSI), Saliency ratio (SR) | - |
| Pyramid feature attention network (PFAN)[38] | 2019 | VGG-16+CNN | - | DUTS-test, ECSSD, HKUIS, PASCAL-S, DUT-OMRON, DUTS | Cross-entropy loss | wFp and MAE |
| DSCLR[28] | 2016 | CNN+LSTM | 13, 49 | SALICON, MIT300, MIT1003 | NSS loss | sAUC, AUC, NSS, CC |
| Multi-level CNN[5] | 2016 | CNN | 13 | SALICON, MIT300 | Square error loss | EMD, NSS, CC, AUC Judd, AUC-Borji, sAUC |
| GAZE II[21] | 2016 | VGG-19+Readout Network | 19+4 | SALICON, MIT1003, MIT300 | Probability distribution, softmax | AUC, sAUC |
| HD-CNNs[40] | 2015 | CNN | - | CIFAR100, ImageNet | Multinomial logistic loss | Top-1 and top-5 errors |

(Mr-CNN), which was a predicted eye fixation computational framework that learned two types of features simultaneously from input images. It was trained on fixation and non-fixation locations with multiresolution and utilized input images as raw pixels. At the end, Top-down and bottom-up features were integrated in the last layer to predict visual saliency.

Gaze II in [21] used the initial deep features from the VGG-19 model, which was used for the image identification model. It trained on fixation and non-fixation locations with multiresolution and utilized input images as raw pixels. At the end, Top-down and bottom-up features were integrated in the last layer to predict visual saliency.

III. DESIGN OF VISUAL SALIENCY MODEL

In this section, we will discuss the important factors of Residual Neural Network and the details of our two-stream visual saliency model architecture.

A. RESIDUAL NEURAL NETWORK

First, we introduce the basic structure of the Residual Network Model (ResNet) [9], that is, an excellent feed-forward deep intensely arrangement of interconnected convolutional layer blocks which has the power to learn and highlight features from input data. Then describe the proposed two-stream framework in Section 3.2. Most of the CNNs designs such as AlexNet [18], VGG-16 [35], and GoogLeNet [36] are comparatively ‘Shallow’ for generating saliency map, but the Residual Network builds its deep architecture based on the popular CNN model for saliency presentation [2].
The Residual Network Model (ResNet-50) [9] is the varietal way and the deepest ever presented for classification in vision. It won the 1st place on the tasks of ImageNet [33] detection, ImageNet localization, COCO detection, and COCO segmentation in ILSVRC & COCO 2015 competition [9]. A deeper network has demonstrated the degradation problem by loss of information which affects the training accuracy [2], [9]. Therefore, for faster training and to construct a really deep network, two approaches are initially introduced in ResNet [9]: one is a stack building block of similar connecting shapes and the other is a new skip connection approach [9], [10].

These building blocks are known as “Residual Units” in [10] that optimiz Residual Network Model (ResNet) [9] than plane deep learning models. Due to feed forward network identity mapping in the form of short-cut connection which skips some layers and adds results to the tiled layer output. Skip connection is an information compensation strategy which intuitively collects prior layer information with equal scale that compensates current layer features [2].

**B. TWO STREAM VISUAL SALIENCY NETWORK**

Inspired by “Salicon” presented in [11], which was a pioneer effort to train a model using DNNs for “Visual Saliency Prediction”. The main concentration is how to make visual saliency prediction for useable applications. The architecture can be useful to calculate the saliency prediction within the images based on ResNet-50 [9], which is pretrained on ImageNet [33] for object classification. Therefore, a trainable end-to-end two-stream ResNet-50 [9] framework is proposed to address the fixation map problem and permits to learn the parameters for back propagation of the pretrained ResNet [9] for optimizing a saliency. Several ways are there to accomplish the integration of two stream data going from initial fusion to the later one [39].

However, to achieve this, we design a very simple two-stream architecture with 49 convolutional layers in each stream that will be fused at the end for capturing extracted features to reduce the semantic gap for saliency maps.

Consequently, we have one ResNet-50 [9] to generate RH streams and another ResNet-50 [9] to generate RL streams. The detailed architecture is displayed in Fig. 3. These two streams are fed by two input images with three dimensions “1000 x 800 x 3” and “500 x 400 x 3”. The first two measurements record the spatial area of the responsive field of the neuron, and the third one lists the layouts for which the neuron is tuned [11]. The neurons are tuned to detect the same patterns because of these two streams that share the same filters but at a different scale. This model contains 99 “Convolutional” layers in total, two “Max Pooling” layers, and one “Concatenate” layer. Firstly, Reset-50 is employed to get the initial features by initializing the first 30 layers from the pretrained ResNet-50 [9] on ImageNet [33] dataset. Then, we modify some parameters in it to record the saliency measurements. The proposed system explains the parameters of the model architecture as shown in Fig. 4. One “Max Pooling” layer is used after the first convolutional layer with pad = ‘0’ and stride = ‘3’. When RGB images are resized as high (1000 x 800 x 3) and low (500 x 400 x 3) scales, respectively, we indicate the neural reactions of these two streams after the second last convolutional layer of “Conv5” block with dimensions of “35 x 48” and “17 x 24”, but both streams are taken third dimension as 2048 at this level.

Note that RH has half the spatial resolution of RH at the second last convolutional layer of “Conv5” block. Next, the output of low-scale residual network is resized by upsampling to “35 x 48” with a linear interpolation to match the same spatial resolution of high-scale residual network. Combine the responses of two-scale residual networks for creating the maps of saliency. The last “Max Pooling” layer is used with stride = ‘1’ and pad = ‘0’ to denote the global features. Then, we introduce the last convolutional layer to learn the global visual contrast information. This convolutional layer is used as a single filter with “0” padding and stride ≈ “0”, that identifies whether the reactions in the last layer relate to the salient region in the form of accurate saliency maps. This layer generates the resolution of “37 x 52”. At the end, we resized the ground truth maps to meet the size of our network output.

**IV. EXPERIMENTAL DETAILS**

The extensive experiments demonstrated that typically all saliency algorithms did not show adequate singleton target in natural images. Therefore, our framework can be simply extended to have a variety of previous knowledge for visual saliency detection. All investigations are conducted on four commonly used datasets, containing ECSSD [31], HKU-IS [20], PASCAL-S [37], and DUT-OMRON [14].

| Model Name | AUC-Judd | SIM | EMD | AUC- Borji | sAUC | CC | NSS | KL |
|------------|---------|-----|-----|-----------|------|----|-----|----|
| DeepGaze II [21] | 0.88 | 0.46 | 3.98 | 0.86 | 0.72 | 0.52 | 1.29 | 0.96 |
| DeepFix [4] | 0.87 | 0.67 | 2.04 | 0.80 | 0.71 | 0.78 | 2.26 | 0.63 |
| Salicon [11] | 0.87 | 0.60 | 2.62 | 0.85 | 0.74 | 0.74 | 2.12 | 0.54 |
| DISLRCN [28] | 0.87 | 0.68 | 2.17 | 0.79 | 0.72 | 0.80 | 2.35 | 0.95 |
| SaGAN [6] | 0.86 | 0.63 | 2.29 | 0.81 | 0.72 | 0.75 | 2.04 | 1.07 |
| Deep Multi-Level [5] | 0.85 | 0.59 | 2.63 | 0.75 | 0.70 | 0.67 | 2.65 | - |
| Deep CNN [32] | 0.83 | 0.52 | 3.31 | 0.82 | 0.69 | 0.58 | 1.51 | 0.81 |
| WPSAM [22] | 0.80 | 0.45 | 4.22 | 0.78 | 0.62 | 0.51 | 1.36 | 1.00 |
| MC-CNN [29] | 0.79 | 0.48 | 3.71 | 0.75 | 0.69 | 0.48 | 1.37 | 1.08 |
| EML-Net [2] | 0.88 | 0.68 | 1.84 | 0.77 | 0.70 | 0.79 | 2.47 | 0.84 |
| CNN-VLM [17] | 0.79 | 0.43 | 4.55 | 0.79 | 0.71 | 0.44 | 1.18 | 1.06 |
A. MODEL TRAINING

We implement the proposed model in PyCaffe by using ResNet-50 [9], pre-trained on ImageNet [33] as a basic model to extract early features. The most common four datasets are employed for further training on the high and low scale dimensions of the input images. In training, we fine-tune by using training images to determine the learning weights with a momentum of 0.9 and a weight decay of 0.0005 on four different datasets separately until the training loss converges. Training has been running for 80 epochs with real ground truth fixation masks for fine-tuning. Fine-tuning of ResNet-50 [9] model for visual saliency up to 80 epochs is shown in Fig. 5. The learning rate of the first 30 convolutional layers is set to 0, but the learning rate of rest of the convolutional layers is set to 0.0001. In addition, network parameters are optimized using “Adam optimizer” with a batch size of 16. The visual saliency detection can be considered as a binary prediction problem; thus we utilize binary cross entropy as the loss function. We prepared the system in PUCIT, a NVIDIA Titan GPU with 12GB memory, and it took different time spans for the four datasets upon the system utilized.

B. DATASETS

As more models have been proposed in the writing, more datasets have been acquainted with further saliency discovery models, but the reality is that more datasets are required in the literature. There are some widely used datasets which play an important role for the most prominent object visualization. Different benchmarks used various datasets for assessing remarkable visual saliency for salient objects and for performance evaluation of saliency generation models. In this work, we evaluate the proposed visual saliency model by using the most persuasive datasets, including ECSSD [31], HKU-IS [20], PASCAL-S [37], DUT-OMRON [14] that are commonly used in many earlier works in the field of remarkable saliency fixation.

PASCAL-S [37] dataset contains 850 validation sets of natural images with ground truth of full segmentation from PASCAL VOC 2010 dataset, which has 8 free-viewing viewers for exploring the images. For each input image, the individual was asked to identify a salient object by clicking with no time limit and there were also no constraints on the number of objects one can choose.

HKU-IS [20] dataset contains 4447 complex images which consists of many disconnected salient objects having diverse spatial locations. This dataset is thought-provoking for similar background and foreground looks.

ECSSD [31] dataset contains the most challenging 1000 images with diversified patterns in both foreground and background. It is a structurally complex new scene dataset, which contains challenging natural images for saliency detection and corresponding ground truth masks. Five helpers produced the ground truth mask.

DUT-OMRON [14] dataset contains complex 5168 images with pixel wise ground truth masks of salient objects. It is a diverse dataset which consists of sample images of side length 400 pixels.

C. EVALUATION METRICS

In this section, we discuss three criteria which are used for performance evaluation of our proposed model, i.e., Maximum F-measure (MaxFβ), mean absolute error (MAE), precision, and recall curve (PR).
FIGURE 4. Architecture of two streams ResNet-50 [9] used for visual saliency prediction with fusion.

FIGURE 5. Fine-tuning of ResNet-50 [9] model for visual saliency up to 80 epoches.

PR-curve is used to measure the estimated saliency map with the threshold ranging from 0 to 255. The visual saliency map can be changed into a binary map. Then, its precision and recall can be obtained by comparing the machine-generated saliency map with the ground truth masks. By doing these comparisons at each threshold value produces P-R curves for the four mentioned datasets.

F-measure is used to measure the harmonic mean of average precision and recall. It is based on pixel-wise error and can evaluate the overall performance [12].

Mean Absolute Error is used to represent the average absolute difference of the estimated saliency map and the ground truth saliency map. It often snubs structural similarities [1].

To validate the efficiency of our model, we perform several experiments, in which we find that our rich hierarchical model explore the representative potential at pixel and semantic level for learning visual saliency strategies which can be utilized for recovering local details. The visual saliency finding results from the above experiments show that visual saliency maps generated by benefiting saliency optimization process with better quality [45]. Commonly, the accuracy and superior performance of the model have been improved after multilevel feature fusion between low and high scales and performed analysis on ECSSD [31], HKU-IS [20], PASCAL-S [37] and DUT-OMRON [14] datasets. DUT-OMRON [14] is the largest and more challenging dataset among the four datasets, which has a difficult scenario due to the large number of complex scenes to identify the best performance of a model [8], [14]. PR curve demonstrates the clear and comparative small range distribution of precision and recall points by using a binary cross entropy loss function [43]. As a result, the produced saliency maps are sensitive to binary thresholds which produce smooth PR curves. It demonstrates the higher performance and better PR curves especially on ECSSD dataset. However, SMD [41] drops faster compared to the nearest ELD [19] and MSI-CNN [26] methods on all used datasets. According to our observation, multi-scale fusion strategy plays an important role on a model performance, and the quantitative results can be further improved by different factors such as: number of layers, image dimension, and hyper-parameter values. Fig. 6 shows the comparative results of our method in terms of PR curves on four commonly used datasets.

V. RESULTS DISCUSSION

We performed a fair comparison of our proposed method with eight state-of-the-art saliency fixation approaches including...
MGCC [1], FSN [8], MSA-CNN [23], SMD [41], ELD [19], MSI-CNN [26], JLSD [27], CNET + PNET [24], BENDer#1 [43] on four commonly used datasets. We choose these methods because they are based on CNN, identified as a benchmark, and developed recently. As illustrated in Table 3, we can see that our model achieves significant performance after two scales fusion of residual global features from each stream than the other 7 methods, but JLSD [27] proved to be equally better with our method. From Table 3 and Fig. 6, we can see that our model achieves considerable performance over all four datasets, including the lowest MAE and the highest Maximum F-measure ($\text{MaxF}_\beta$) after two scale fusions of residual global features from each stream. We observe that our trained model measured stable values of max F-measure but showed unstable in terms of MAE. In addition, it has lower MAE values on PASCAL-S [37] and ECSSD [31] dataset which is considered better and at the second position in terms of MaxF$_\beta$ on all four datasets compared to the other seven models but JLSD [27] and BENDer#1 [43] are very close to our model. Moreover, our method shows encouraging results in terms of overall performance metrics.

**Qualitative Comparison:** We selected ResNet-50 [9] with more generalization ability due to the large number of operations connected to depth features for getting improved performance [44]. One frame produces pixel level visual saliency maps and the other frame produces full resolution semantic level visual maps [42]. By using ResNet-50 [9], we get improved results in the form of visual maps. Such improvement caused by the fusion of two streams results that highlight the fixation and semantic level, but still it is a big challenge to get the best quality saliency scores. Fig. 7 shows a comparison between our model’s saliency map and six other saliency prediction model’s results which are provided by the concerning authors. These state-of-the-art models are Gaze-II [21], Salicon [11], LSTM (SAM) [28], Deep CNN [32], ELM [40] and Mr-CNN [29]. Our model can predict the correct salient
TABLE 3. Performance of the proposed method and other 6 state-of-the-art approaches on four commonly used datasets. Red, blue, and green indicate the best, the second best, and the third best results in terms of maximum F-measure (↑ means larger), MAE (↑ means smaller) and “-” represents no reported.

| DATASETS | DUT-OMRON | PASCAL-S | HKU-IS | ECSSD |
|----------|-----------|----------|--------|-------|
| METHODS  | MaxF ↑ MAE ↓ | MaxF ↑ MAE ↓ | MaxF ↑ MAE ↓ | MaxF ↑ MAE ↓ |
| MCC[1]   | 0.726     | 0.074    | 0.808  | 0.109 |
| FSN[8]   | 0.768     | 0.066    | 0.827  | 0.095 |
| MSA-CNN[23] | 0.747     | 0.103    | 0.808  | 0.104 |
| SMD[41]  | 0.624     | 0.166    | 0.690  | 0.201 |
| ELID[19] | 0.760     | 0.092    | 0.771  | 0.126 |
| MSN[26]  | 0.713     | 0.091    | 0.790  | 0.117 |
| JLSDD[27] | 0.832     | 0.052    | 0.851  | 0.067 |
| CNET+PNEL[24] | 0.718     | 0.114    | 0.790  | 0.134 |
| BENDER[143] | 0.7098    | 0.0673   | 0.8642 | 0.078 |
| Others   | 0.831     | 0.072    | 0.853  | 0.066 |

VI. CONCLUSION

Recently, visual saliency map generation can be considered a useful study for image and video applications. Hence, to predict the visual saliency detection system, we fine-tune the ResNet-50 [9] model that is pretrained on ImageNet [33]. We perform various experiments on rearchitect ResNet-50 [9] model in the form of two streams. These two streams are fed by input images at low and high scales that prompt a saliency identification. In the future, this model can be tested with more number of layers to get significant results.

CONFLICT OF INTEREST

None of the authors have a conflict of interest related to the research and results presented in this paper.

DATA AVAILABILITY STATEMENT

The datasets used in the experiments and discussed in the paper will be available if required.

FUNDING STATEMENT

No funding is involved in this research.

REFERENCES

[1] W. Feng, X. Li, G. Gao, X. Chen, and Q. Liu, “Multi-scale global contrast CNN for salient object detection,” Sensors, vol. 20, no. 9, p. 2656, May 2020.
[2] S. Jia and N. D. B. Bruce, “EML-NET: An expandable multi-layer NETwork for saliency prediction,” Image Vis. Comput., vol. 95, Mar. 2020, Art. no. 103887.
[3] F. Murabito, C. Spannamato, S. Palazzo, D. Giordano, K. Pogorelov, and M. Riegler, “Top-down saliency detection driven by visual classification,” Comput. Vis. Image Understand., vol. 172, pp. 67–76, Jul. 2018.
[4] S. S. K. Krishnaveni, K. Ayush, and R. V. Babu, “DeepFix: A fully convolutional neural network for predicting human eye fixations,” IEEE Trans. Image Process., vol. 26, no. 9, pp. 4446–4456, Sep. 2017.
[5] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “A deep multi-level network for saliency prediction,” 2016, arXiv:1609.01064. [Online]. Available: https://arxiv.org/abs/1609.01064
[6] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O’Connor, J. Torres, E. Sayrol, and X. Giro-I-Nieto, “SalGAN: Visual saliency prediction with generative adversarial networks,” 2017, arXiv:1701.01081. [Online]. Available: https://arxiv.org/abs/1701.01081
[7] B. Kim, J. Seo, S. Jeon, J. Koo, J. Choe, and T. Jeon, “Why are saliency maps noisy? Cause of and solution to noisy saliency maps,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 4149–4157.
[8] X. Chen, A. Zheng, J. Li, and F. Lu, “Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic CNNs,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1050–1058.
[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–777.
[10] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis. Springer, 2016, pp. 630–645.
[11] X. Huang, C. Shen, X. Boix, and Q. Zhao, “SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 262–270.
[12] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, “Depth-induced multi-scale recurrent attention network for saliency detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7254–7263.
[13] S. Gorji and J. J. Clark, “Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2510–2519.
[14] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2013, pp. 3166–3173.
[15] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “Predicting human eye fixations via an LSTM-based saliency attentive model,” IEEE Trans. Image Process., vol. 27, no. 10, pp. 5142–5154, Oct. 2018.
[16] H. Kato and T. Harada, “Visual language modeling on CNN representations,” 2015, arXiv:1511.02872. [Online]. Available: https://arxiv.org/abs/1511.02872
[17] F. Cao, Y. Liu, and D. Wang, “Efficient saliency detection using convolutional neural networks with feature selection,” Inf. Sci., vol. 456, pp. 34–49, Aug. 2018.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. 2012, pp. 1097–1105.
[19] G. Lee, Y.-W. Tu, and J. Kim, “Deep saliency with encoded low level features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 660–668.
[20] G. Li and Y. Yu, “Visual saliency detection based on multiscale deep CNN features,” IEEE Trans. Image Process., vol. 25, no. 11, pp. 5012–5024, Nov. 2016.
[21] M. Kümmerer, T. S. A. Wallis, and M. Bethge, “DeepGaze II: Reading fixations from deep features trained on object recognition,” 2016, arXiv:1610.01563. [Online]. Available: https://arxiv.org/abs/1610.01563
[22] A. Lahiri, S. Roy, A. Santara, P. Mitra, and P. K. Biswas, “WEPASM: Weakly pre-learnt saliency model,” 2016, arXiv:1605.01101. [Online]. Available: https://arxiv.org/abs/1605.01101
[23] Y. Ji, H. Zhang, and Q. M. Jonathan Wu, “Salient object detection via multi-scale attention CNN,” Neurocomputing, vol. 322, pp. 130–140, Dec. 2018.
[24] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, “Multi-source weak supervision for saliency detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5074–5083.

 object region even under complex scenes, human structure, and animals. It can also detect a significant regions when cluttering and unrelated background is present in the images. It can be observed that our model may be indicated as the most prominent object in a better manner but also achieves encouraging results for different size objects in images.
[25] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, “SalNet360: Saliency maps for omni-directional images with CNN,” *Signal Process., Image Commun.*, vol. 69, pp. 26–34, Nov. 2018.

[26] K. Huang and S. Gao, “Image saliency detection via multi-scale iterative CNN,” *Vis. Comput.*, vol. 36, no. 7, pp. 1–13, 2019.

[27] Z. Yu, Y. Zhuge, H. Lu, and L. Zhang, “Joint learning of saliency detection and weakly supervised semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7223–7233.

[28] N. Liu and J. Han, “A deep spatial contextual long-term recurrent convolutional network for saliency detection,” 2016, arXiv:1610.01708. [Online]. Available: https://arxiv.org/abs/1610.01708

[29] I. S. Bajwa, M. N. Asghar, and M. A. Naem, “Automated detection of early tropical cyclones formation in satellite images,” *Pakistan J. Sci.*, vol. 69, no. 4, pp. 417–421, Dec. 2017.

[30] S. Bukhari, M. R. Anjum, and S. I. Bajwa, “Chaos image encryption followed by the steganography technique,” *Sindh Univ. Res. J.*, vol. 49, no. 1, pp. 113–118, Mar. 2017.

[31] Q. Yan, L. Xu, J. Shi, and J. Jia, “Hierarchical saliency detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*. Jun. 2013, pp. 1155–1162.

[32] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. O’Connor, “Shallow and deep convolutional networks for saliency prediction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Jun. 2016, pp. 598–606.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[34] I. Djemai, S. Fezza, W. Hamidouche, and O. Deforges, “Extending 2D saliency models for head movement prediction in 360-degree images using CNN-based fusion,” 2020, arXiv:2002.09196. [Online]. Available: https://arxiv.org/abs/2002.09196

[35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Jun. 2015, pp. 1–9.

[37] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, “The secrets of salient object segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[38] T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Jun. 2019, pp. 3085–3094.

[39] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, “Deep learning for light field saliency detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. Oct. 2019, pp. 8838–8848.

[40] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, “HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. Dec. 2015, pp. 2740–2748.

[41] Y. Ji, H. Zhang, Z. Zhang, and M. Liu, “CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances,” *Inf. Sci.*, vol. 546, pp. 835–857, Feb. 2021.

[42] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, “Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion,” 2021, arXiv:2103.11832. [Online]. Available: https://arxiv.org/abs/2103.11832

[43] K. Huang, T. Xu, Y. Zhang, C. Pan, J. Hao, and X. Li, “Salient object detection on hyperspectral images in wireless network using CNN and saliency optimization,” *Ad Hoc Netw.*, vol. 112, Mar. 2021, Art. no. 102369. * * *