Learning to Predict Eye Fixations via Multiresolution Convolutional Neural Networks

Nian Liu, Junwei Han, *Senior Member, IEEE*, Tianming Liu, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Eye movements in the case of freely viewing natural scenes are believed to be guided by local contrast, global contrast, and top-down visual factors. Although a lot of previous works have explored these three saliency cues for several years, there still exists much room for improvement on how to model them and integrate them effectively. This paper proposes a novel computation model to predict eye fixations, which adopts a multiresolution convolutional neural network (Mr-CNN) to infer these three types of saliency cues from raw image data simultaneously. The proposed Mr-CNN is trained directly from fixation and nonfixation pixels with multiresolution input image regions with different contexts. It utilizes image pixels as inputs and eye fixation points as labels. Then, both the local and global contrasts are learned by fusing information in multiple contexts. Meanwhile, various top-down factors are learned in higher layers. Finally, optimal combination of top-down factors and bottom-up contrasts can be learned to predict eye fixations. The proposed approach significantly outperforms the state-of-the-art methods on several publically available benchmark databases, demonstrating the superiority of Mr-CNN. We also apply our method to the RGB-D image saliency detection problem. Through learning saliency cues induced by depth and RGB information on pixel level jointly and their interactions, our model achieves better performance on predicting eye fixations in RGB-D images.**

***Index Terms*— Contrast, convolutional neural network (CNN), eye fixation prediction, RGB-D, saliency detection.**

I. Introduction

TREMENDOUS visual information comes into our eyes every day. However, our brain could not process all these visual data due to our limited energy. Fortunately, the visual system enables human to selectively focus attention on some distinctive visual scenes and ignore trivial ones. This selection mechanism is called *visual attention*, which usually consists of two forms: bottom-up saliency-driven attention and top-down task-driven attention [1]. The former attention depends on the visual stimuli themselves and happens at the early stage of scene viewing. The latter is guided by specific task or human prior knowledge and is slow acting.

Visual saliency can rapidly allocate our processing resources to locate our eye fixations on distinctive visual scenes, thus making human to be efficient in further high-level processing, e.g., scene understanding and object detection. In computer science community, researchers have developed a number of computational saliency models [2]–[8] to estimate human eye fixation locations by resorting to computer vision methods in the latest decades of years. These saliency prediction models also benefit many other engineering applications, e.g., image and video compression [9], image and video retargeting [10], [11], and object recognition [12].

Biological evidence indicates that locations popping out from their local neighboring regions are more probably to be attended by human. Inspired by this evidence, most traditional approaches typically model saliency by measuring the distinctiveness of each location, which consists of three components: early feature extraction, feature contrast inference, and contrast integration. For early feature extraction, most traditional works utilized biological plausible low-level features, e.g., intensity, color, and orientation used in [2], subbands of the steerable pyramid, and color probabilities used in [3]. As we can see, all these features are hand-crafted, which require proper and enough domain-specific knowledge. However, lots of fields in human visual attention still remain unexplored, and thus, the hand-crafted features may not generalize well for various visual scenes. Although many models tried to learn various early features via machine learning algorithms, for example, sparse coding [5], [14], [15], autoencoder [54], [55], and ICA [13], these shallow models are very difficult to learn high-level semantic concepts and complex patterns because of their limited capabilities to mine hidden representations.

Another key component in saliency modeling is contrast inference over extracted features. In terms of the context used for contrast calculation, most previous works can be categorized into two schools: local contrast-based methods and global contrast-based methods. Local contrast-based methods

Manuscript received December 25, 2015; revised May 9, 2016 and October 21, 2016; accepted November 4, 2016. Date of publication November 29, 2016; date of current version January 16, 2018. This work was supported by the National Science Foundation of China under Grant 61473231 and Grant 61522207. This work was presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2015 [53]. *(Corresponding author: Junwei Han.)*

N. Liu and J. Han are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

T. Liu is with the Department of Computer Science, University of Georgia, Athens, GA 30602 USA (e-mail: tliu@cs.uga.edu).

X. Li with the State Key Laboratory of Transient Optics and Photonics, Center for OPTical IMagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@opt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2628878

measure the saliency of each image location via computing the contrast against its local surroundings. Itti *et al.* [2] designed the "difference of gaussians" operator across multiple scales to calculate the "center-surround difference" as local contrast, which has become the most influential theory. Following this idea, many other works inferred local contrast via different mechanisms, e.g., self-information in [13], Kullback–Leibler divergence in [16], and sparse coding residual in [15]. Global contrast-based methods measure saliency by referring to the whole image. There are also various mechanisms used in previous works to compute global contrast, e.g., graph computations in [17], spectrum residual over Fourier transform in [18], and whitened feature decomposition responses in [6]. However, we can see that most of the previous works rely on hand-designed mechanisms to calculate contrast, which may not be universally fit for a large amount of images with diverse visual scenarios.

The last component for detecting saliency is to combine different contrast features to calculate the final saliency maps. Itti *et al.* [2] combined three contrast maps with predefined linear weights. Zhao and Koch [19] proposed a least square technique to obtain optimal weights for different contrasts. Judd *et al.* [3] trained an Support Vector Machine (SVM) to combine various features. To fuse local contrast and global contrast, Borji and Itti [5] investigated different integration schemes, including addition, multiplication, maximization, and minimization. However, all these integration schemes are either naïve arithmetic or simple shallow models, which could not handle complex natural scenes well.

Although most traditional works mainly consider visual attention in free viewing natural scenes as contrast-based saliency modeling, it has been found that high-level factors (e.g., objects [20], [21], actions [22], and events) also direct eye movements due to the human prior knowledge [23], [24]. Hence, it is reasonable to fuse bottom-up saliency and high-level cues to yield a better model to predict eye fixations. A few previous works have considered this idea. For example, Cerf *et al.* [25] combined saliency information with face detection. Judd *et al.* [3] and Borji [4] incorporated more high-level factors, such as faces, texts, cars, and humans, into saliency information. Although they have achieved good performance via integrating top-down factors, they only considered a small number of predefined high-level factors by utilizing specific object detectors. Thus, there is still much room for improvement by investigating more top-down factors.

From the discussions above, we can conclude that the most previous saliency detection works adopt *ad hoc* features to compute contrast via carefully designed mechanisms. Then, various saliency cues, including local contrast, global contrast, and specifically chosen top-down factors, are heuristically integrated to infer the saliency value of each location. However, natural scenes are usually sophisticated; thus, human-designed features and contrast inference mechanisms, ad hoc top-down factors, and simple integration schemes could not adapt to all types of natural scenes well. Thus, all these problems indicate the urgent demand of designing a novel integrated saliency model to enhance all the components involved in the eye fixation prediction problem. As shown in Fig. 1, in the first row, the effigy in the middle of the image mainly differs from the surrounding walls in terms of shape and structure, instead of color or texture features used in traditional methods. Therefore, if we can learn proper features to discriminate it from its surroundings, we can successfully highlight it, as shown in Fig. 1(c). In the second row, the detection of the salient traffic sign usually suffers from the distraction of the complex backgrounds, which results in false positives in traditional methods, even in the method in Fig. 1(f) which combines local and global contrasts heuristically. However, we can get rid of the distraction if we combine the local and global contrasts well, as shown in Fig. 1(c). In the third row, the head of the animal is a semantic top-down factor which attracts human attention most in the image. However, all traditional methods fail in detecting it, including the method in Fig. 1(g) which just incorporates some specifically chosen high-level factors. Thus, we can efficiently highlight it if we can learn various top-down factors from large-scale data, as shown in Fig. 1(c). In the fourth row, the image contains various saliency cues, including top-down factors (face, human, and text) and highly contrasted regions in the cluttered backgrounds. Only by integrating these saliency cues intrinsically we can obtain the correct saliency map, as shown in Fig. 1(c).

In this paper, we propose a novel computational framework using a multiresolution convolutional neural network (Mr-CNN) to learn low-level features, saliency information, top-down cues, and their integration from the scratch. Specifically, we treat the fixation prediction problem as a pixelwise classification problem. For each pixel, we supply three multiresolution image regions with different contexts as the essential information over which the saliency value of the pixel can be inferred via mimicking the classic "center-surround difference" mechanism. We feed these three image regions into the three streams of the Mr-CNN. The first convolutional network (CNN) stream models the property of the center region of the pixel, the second CNN stream models the surrounding context of the pixel, and the third CNN stream models the global context of the whole image. Then, these three streams are combined to infer the saliency value of the pixel. The Mr-CNN is trained in an end-to-end manner, with the input of raw image pixels and their corresponding eye fixations as labels. Benefiting from its supervised training mechanism and hierarchical architecture, the proposed Mr-CNN can infer saliency relevant factors in convolutional layers, rather than relying on hand-crafted features. These features are learned with progressively increasing complexity in different layers, and thus can model the intrinsic hierarchical structure of image representation, making it efficient and discriminative. In higher layers, the Mr-CNN can automatically learn various high-level factors because of the deep architecture, rather than using specifically chosen object detectors in previous works. When the features come from all the center region, the surrounding context, and the global context are combined, local and global contrasts can be learned intrinsically and efficiently under the supervision of large-scale training data, instead of using a variety of human-crafted mechanisms in previous approaches. Finally, the subsequent two fully connected layers and the output layer learn to integrate local contrast, global contrast,

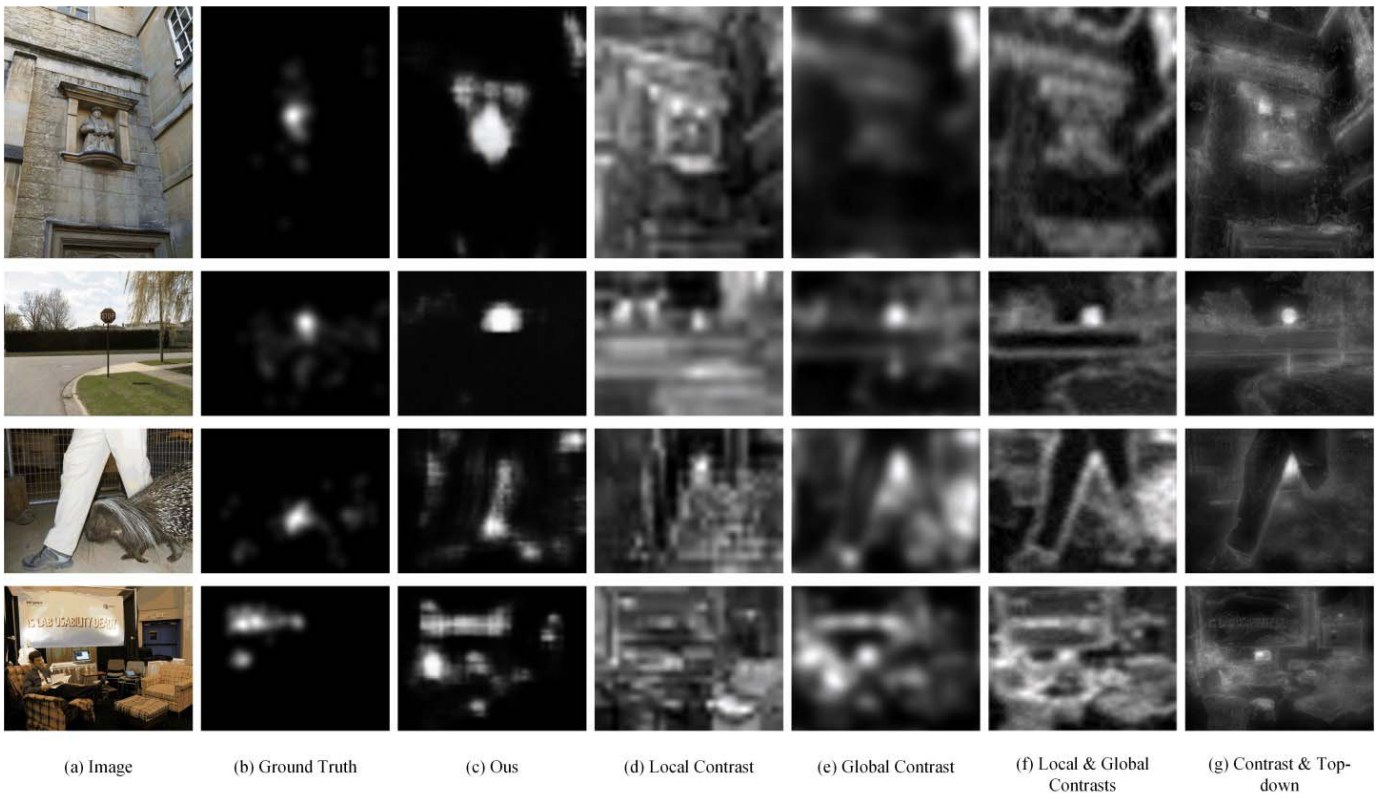

Fig. 1. Generated saliency maps of different types of methods, including (a) image, (b) GT, (c) our proposed method, (d) a local contrast-based method, (e) global contrast-based method, (f) local and global contrasts combined method, and (g) a method that integrates contrast and top-down factors.

and top-down cues to predict eye fixations, rather than utilizing simple integration schemes or shallow models in traditional methods.

On the other hand, human visual systems operate in real 3-D environments. Depth cues are generally used by human visual attention mechanisms. Therefore, it is important and interesting to study the problem of eye fixation prediction in RGB-D images acquired by new sensing technologies, such as the Microsoft Kinect. A small number of previous works have explored the integration of depth information in saliency detection. Ouerhani and Hügli [26] directly extended the approach of Itti *et al.* [2] by adding a depth conspicuity map to the intensity, color, and orientation channels. Then, the four conspicuity maps were fused over multiple scales to obtain the final saliency map. Lang *et al.* [27] learned a depth prior from 3-D eye fixation data to generate depth saliency maps first, and then fused them with other RGB channel-based saliency maps by simple summation or multiplication. Ciptadi *et al.* [28] extracted layout and shape features from the depth channel and then integrated them with color features to calculate the saliency value of each pixel. Above-mentioned previous works mainly adopted simple depth-induced saliency cues (the depth conspicuity map used in [26] and the depth prior used in [27]) or human designed features (the layout and shape features used in [28]). None of them tried to mine more informative saliency cues from the depth information. In addition, the works of [26] and [28] fused the depth information in the middle stage, i.e., they first extracted saliency cues from the depth information, and then combined them with saliency cues induced from the RGB information to infer saliency jointly. The work of [27] fused the depth information in the late stage. It first yielded two different saliency maps using depth information and RGB information separately. Then, these two saliency maps were fused to obtain the final saliency map. However, none of these previous works explored to fuse depth and RGB information in the early stage, i.e., combining them immediately on the pixel level. This strategy can be used to mine more informative saliency cues which combine depth and RGB information jointly, and incorporate complicated joint interactions between them.

Thus, we propose to apply Mr-CNN to address the problem of RGB-D image saliency detection via exploring more informative depth-induced saliency cues and performing the early fusion. We directly feed the raw RGB images and depth maps into the Mr-CNN as images with four channels. By jointly learning various saliency cues from the RGB and depth information from the start, the Mr-CNN can learn to predict more accurate eye fixations than traditional models.

Lately, deep neural networks have been introduced into the visual saliency detection task by some works. Shen *et al.* [29] adopted unsupervised feature learning techniques to mine informative high-level semantic concepts from salient image regions via multiple convolutional sparse coding layers. Next, they used a linear SVM to optimally fuse these features to estimate saliency values. Similarly, Vig *et al.* [30] first utilized hierarchical neuromorphic net-

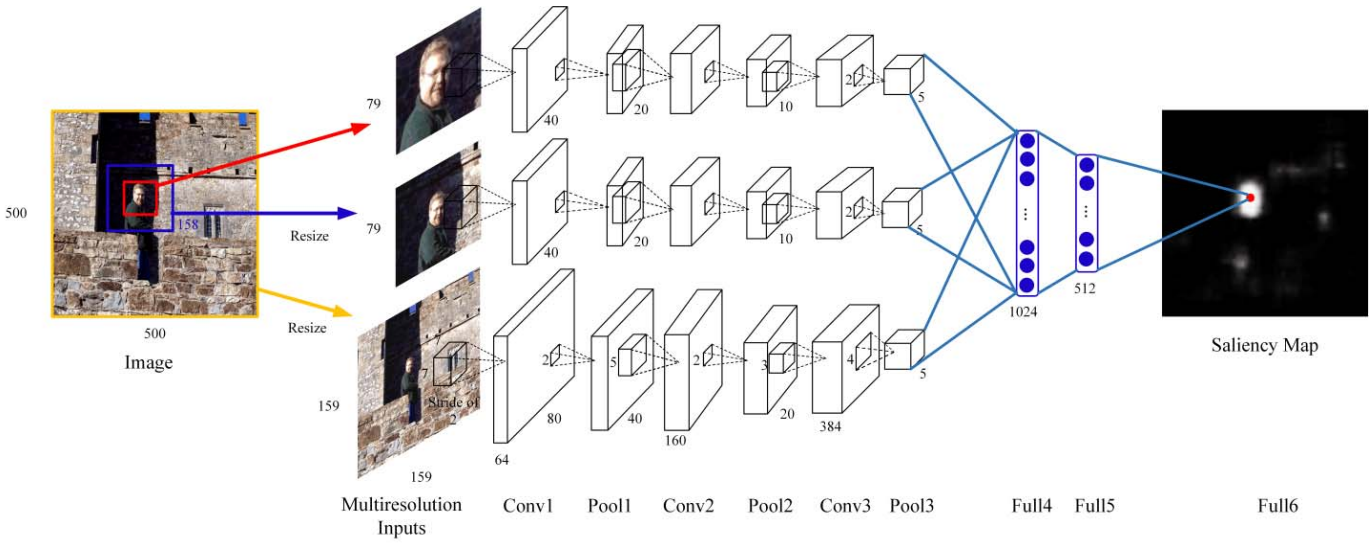

Fig. 2. Architecture of our proposed Mr-CNN model. Here, convolutional layer, max-pooling layer, and fully connected layer are abbreviated as "Conv," "Pool," and "Full," respectively. The sizes of the original image, feature maps, fully connected layers, convolution kernels, and pooling windows are indicated in this figure. Best viewed in color.

works to learn optimal features in a biologically plausible way, and then, a linear SVM was adopted to predict the saliency of each image location. Nevertheless, only the learning of deep features was considered by the above two works, while the important contrast mechanism was totally ignored. On the contrary, we utilize the multiresolution architecture to learn the local and global contrasts explicitly. Lin *et al.* [31] adopted a group of adaptive convolutional filters to obtain both low-level and mid-level features. Then, the local contrast was calculated by center-surround difference over the obtained features. However, their model did not take into account top-down factors while we develop an integrated network to fuse low-level bottom-up saliency and top-down high-level semantics. Besides, Kümmerer *et al.* [32] learned a linear combination of features from different layers of AlexNet [33] to predict eye fixations. Kruthiventi *et al.* [34] learned a fully CNN based on VGG net [35] with global context and adopted location biased convolution to incorporate center-bias for eye fixation prediction. Huang *et al.* [36] learned multiscale CNNs based on deep architectures (AlexNet [33], VGG net [35], and GoogLeNet [37]) and linearly integrated their features to learn to detect saliency. Although these three models achieve very good results, different from our method, they heavily rely on deep models pretrained on large-scale data of ImageNet [38].

We briefly summarize the contributions of this paper as fourfolds.

1) By adopting the proposed Mr-CNN in eye fixation prediction problem, we simultaneously learn early features, local contrast, global contrast, top-down factors, and their integration intrinsically, rather than depending on hand-crafted features, computing mechanisms, and heuristic integration schemes.
2) We visualize hierarchical features learned from the Mr-CNN. The showed various features demonstrate that our Mr-CNN can learn both low-level features related to bottom-up saliency and high-level top-down factors.
3) The proposed method is evaluated on seven eye-tracking benchmark data sets and outperforms other ten state-of-the-art models by a large margin.
4) We apply our method to the RGB-D image saliency detection problem. Through automatically and jointly learning saliency cues from the RGB information and the depth information, the proposed method achieves better results compared with previous works.

The rest of this paper is organized as follows. Section II elaborates the proposed Mr-CNN model. Section III reports the experimental results on seven eye fixation benchmarks and the application to the RGB-D image saliency detection problem. Finally, we draw the conclusions in Section IV.

II. Proposed Approach

In this section, we describe the proposed approach in detail. As shown in Fig. 2, the architecture of the model is mainly built on an Mr-CNN. We first briefly review the basic concepts about CNN; then, we go into the depth of the proposed Mr-CNN and demonstrate the application of it on saliency detection.

A. Brief Review of CNN

Generally, a CNN [39] consists of several convolutional layers, pooling layers, and fully connected layers. Typically, the input image is first inputted into a convolutional layer; then, several pooling layers and convolutional layers connect over it alternately to extract the hierarchical features of the input image, subsequently two or three fully connected layers follow on to do classification.

Considering a CNN with L layers, we represent the hidden state of layer l as $\mathbf{H}_l$, where $l \in \{1, \ldots, L\}$. Following this notation, we use $\mathbf{H}_0$ to represent the input image additionally.

For each convolutional layer or fully connected layer l, we have two parts of learnable parameters. The first part is the weight matrix $\mathbf{W}_l$ that connects layer l and layer $l-1$, and the second part is the bias term vector $\mathbf{b}_l$.

For a convolutional layer l, each neuron is only connected with a local region on $\mathbf{H}_{l-1}$ and $\mathbf{W}_l$ is shared among all spatial locations. In detail, the 2-D convolution operation is first performed on $\mathbf{H}_{l-1}$ with the convolution kernels $\mathbf{W}_l$; then, the convolution responses and the bias term $\mathbf{b}_l$ are added up. Subsequently, a pointwise nonlinear function *Actv* is typically adopted to activate the neuron responses, thus obtaining the resultant feature maps $\mathbf{H}_l^C$. The whole convolution operation can be represented as

$$\mathbf{H}_l^C = Actv(\mathbf{H}_{l-1} * \mathbf{W}_l + \mathbf{b}_l) \tag{1}$$

where $*$ denotes the convolution operation.

Next, a max-pooling layer is followed to choose the dominant features over nonoverlapping windows at each location per feature map. We denote the pooling window at location p as $\text{pw}(p)$, thus the max pooling process can be formulated as

$$\mathbf{H}_{l,p}^P = \max_{q \in \text{pw}(p)} (\mathbf{H}_{l-1,q}). \tag{2}$$

When several convolutional layers and pooling layers are stacked alternately in depth, we can extract hierarchical features with larger and larger receptive fields, thus getting low-level features in lower layers and high-level semantic features in higher layers. The extracted convolutional features still keep the spatial layout and are location invariant, thus subsequently they are further integrated into a 1-D global feature vector by two or three fully connected layers. For a fully connected layer l, the weight matrix $\mathbf{W}_l$ connects all the nodes of $\mathbf{H}_{l-1}$ and $\mathbf{H}_l$. Specifically, $\mathbf{H}_{l-1}$ is the first processed with a linear projection via weight $\mathbf{W}_l$ and bias $\mathbf{b}_l$; then, the pointwise nonlinear activation follows on:

$$\mathbf{H}_l^{\text{FC}} = Actv(\mathbf{H}_{l-1} \cdot \mathbf{W}_l + \mathbf{b}_l). \tag{3}$$

There are several nonlinear activation functions that have been proposed. Here, we adopt the rectified linear unit (ReLU) [40] for all the convolutional layers and fully connected layers due to its superior effectiveness and efficiency

$$Actv(x) = \max(0, x). \tag{4}$$

Generally, a task-specific classifier is used as the last layer L to do prediction. A logistic regression classifier can be utilized to do two-class classification. It is similar to a fully connected layer except that there is only one output node and the sigmoid activation function is used

$$H_L = Actv(\mathbf{H}_{L-1} \cdot \mathbf{W}_L + \mathbf{b}_L) \tag{5}$$

where

$$Actv(x) = \frac{1}{1 + e^{-x}}. \tag{6}$$

The value of H_L indicates how likely the input belongs to the positive class.

The model parameters Θ consist of weights $\{\mathbf{W}_1, \ldots, \mathbf{W}_L\}$ and biases $\{\mathbf{b}_1, \ldots, \mathbf{b}_L\}$. They can be jointly optimized by minimizing the cross entropy loss $L(\Theta)$ between the predicted probability H_L and the binary ground truth (GT) label Y, which utilizes the gradient descent algorithm via back propagation [41]

$$L(\Theta) = Y \log H_L + (1 - Y) \log(1 - H_L), \quad Y \in \{0, 1\}. \tag{7}$$

B. Eye Fixation Prediction Using Mr-CNN

We consider the problem of the eye fixation prediction as a pixelwise classification task. For a specific pixel, when we tend to infer that whether it is salient, we consider three types of information, i.e., how do the center region, the surrounding region, and the whole image look like? When we combine all these three types of information, we can measure the local and global contrasts. Then, we could infer the saliency value of the pixel. To implement this idea, we build a CNN architecture with multiple resolution inputs with different contexts, i.e., the Mr-CNN, to predict eye fixations as shown in Fig. 2. To be specific, for each pixel, we feed three image regions into the three streams of the Mr-CNN. The first CNN stream takes a local image region centered on the pixel with the finest resolution as the input to model the property of the center region of the pixel (see the red box in the image in Fig. 2). The second CNN stream takes another bigger local image region but with a coarser resolution as the input to model the surrounding context of the pixel (see the blue box in the image in Fig. 2). The third CNN stream takes the whole image with the coarsest resolution as the input to model the global context (see the yellow box encompassing the whole image in Fig. 2). Each CNN stream consists of three convolutional layers, each of which is followed by a max-pooling layer. Thus, the activation values on the third max-pooling layers in the three CNN streams represent the deep features of the center region, the surrounding region, and the whole image. Then, the three types of deep features come from different contexts are combined in the subsequent two fully connected layers, in which local and global contrasts can be inferred. In addition, top-down factors can be learned automatically in higher layers. At last, the saliency value can be inferred in the last logistic regression layer via integrating both bottom-up saliency and top-down factors.

As for the image region resolution, we first directly warp each image to the size of 500×500 without considering its aspect ratio. Then, we extract three image regions for each pixel on the rescaled image. The center region is extracted with size 79×79 and directly inputted to the Mr-CNN. The surrounding region is extracted as two times larger of the center region, i.e., size of 158×158. Next, we rescale it to size of 79×79 first, and then we feed it into the Mr-CNN so that the two CNN streams can share the same architecture. As for the whole image, if we rescale it to the same size as other two image regions, i.e., 79×79, we will lose too much information since the resolution is too coarse. If we keep its size as 500×500, the resolution will be too fine and the CNN structure will be too large. Thus we rescale it to two times larger of other two regions, i.e., size of 159×159, as the global context input (see Fig. 2).

TABLE I
ARCHITECTURE OF THE PROPOSED Mr-CNN

	Size of Feature Maps	Conv1	Pool1	Size of Feature Maps	Conv2	Pool2	Size of Feature Maps	Conv3	Pool3	Size of Feature Maps	Full4	Full5	Full6
Central Region Stream	3× 79×79	64×7×7 St. 2 pad 3 norm 0.1 ReLU	2×2 Max	64× 20×20	160×5×5 St. 1 pad 2 ReLU	2×2 Max	160× 10×10	384×3×3 St. 1 pad 1 ReLU	2×2 Max	384×5×5 dropout	1024 ReLU dropout	512 ReLU dropout	1 Sigmoid
Surrounding Region Stream	3× 79×79		2×2 Max	64× 20×20		2×2 Max	160× 10×10		2×2 Max	384×5×5 dropout			
Whole Image Stream	3× 159×159		2×2 Max	64× 40×40		2×2 Max	160× 20×20		4×4 Max	384×5×5 dropout			

Each CNN stream consists of 3 convolutional layers (Conv 1–3) and 3 pooling layers (Pool 1–3), then the three streams are merged and three fully connected layers (Full 4–6) follow on. Parameters in each convolutional layer across these three CNN streams are shared. The sizes of the feature maps and convolutional kernels in each convolutional layer are given by the form "channel × width × height". For each pooling layer, the size of each pooling window is given by the form "width × height", and the notation "Max" specifies that the pooling method is max-pooling. The number of neuron nodes of each fully connected layer is also given. In each convolutional layer and fully connected layer, the activation function is given by "ReLU" [40] or "Sigmoid". In convolutional layers, the convolution stride ("St.") and spatial padding size ("pad") are also given. In Conv1, "norm" means we use a weight constraint [42] of 0.1 to constrain the $\iota_2 - norm$ of each convolutional kernel. Dropout [42] is also used in the last convolutional feature maps, Full4 and Full5.

To make the learned features robust to scales, we share the weights of the same convolutional layers among these three CNN streams. We adopt 64 filters with size 7×7 and stride of 2 in the first convolutional layer, 160 filters with size 5×5 and stride of 1 in the second convolutional layer, and 384 filters with size 3×3 and stride of 1 in the third convolutional layer. The image borders are padded with 0 to perform the "same" convolution operations. All max-pooling layers are set with pooling size of 2×2 and stride of 2, except that the pooling size of the third max-pooling layer in the third CNN stream with the whole image input is set to 4×4; thus, we can obtain feature maps of the same size in the third max-pooling layers of all the three CNN streams. The sizes of the two fully connected layers and the output layer are set to 1024, 512, and 1, respectively. To mitigate overfitting, we randomly turned off the neuron activations of the last convolutional layer and the subsequent two fully connected layers with probability of 0.5 via dropout [42]. For the convolutional kernels of the first convolutional layer, we also adopted a weight constraint [42] of 0.1. Once the ι_2-norm of a kernel is greater than the constraint, it is renormalized by division. This also relieves overfitting. The architecture of the whole network is shown in Table I.

In the training stage, a number of fixation and nonfixation pixels are randomly sampled in each training image based on the saliency values in the GT saliency map as the positive and negative samples, respectively, to train the Mr-CNN. For testing, we slide the Mr-CNN in each testing image to generate the saliency map.

Furthermore, for the RGB-D image saliency detection, we directly adopt Mr-CNN on the raw depth and RGB information, just feeding in each image with its depth map simultaneously and changing the input image channel from three dimensions to four dimensions (we refer the Mr-CNN used here with RGB-D image inputs as the 3-D Mr-CNN). By such a little modification, the 3-D Mr-CNN can learn various saliency cues from depth and RGB information jointly and their interactions from the training data. In the testing stage, the 3-D Mr-CNN works in the same sliding window way as Mr-CNN does.

III. EXPERIMENTS

This section presents experimental results to validate the proposed model for the task of predicting eye fixations. At first, eye fixation benchmark data sets and evaluation metrics used in our experiments are reported. Subsequently, we present the implementation details of the proposed Mr-CNN. Next, the results of the proposed model and comparisons with other ten state-of-the-art approaches are reported. Then, the influences of different resolutions are analyzed and hierarchical features obtained from our Mr-CNN are visualized. Finally, we demonstrate the experimental results of the proposed 3-D Mr-CNN model on the RGB-D image saliency detection task.

A. Data Sets

We evaluated our model on seven eye fixation prediction benchmark data sets with varied properties. The first data set we used is **MIT** [3], which is one of the most widely used eye fixation data sets. It consists of 1003 images selected from Flicker and LabelMe data sets, including 779 landscape, 228 portrait, and several synthetic images. The images have diverse resolutions ranging from 405×1024 to 1024×1024 pixels. The GT eye fixation points were obtained via free-viewing of 15 human subjects. The second data set, **Toronto** [13], is also widely used by previous works. It has 120 511×681 color images, each of which is either indoor or outdoor scenes. 20 human subjects freely viewed these images to obtain eye fixations. The third data set we used is the **Cerf** data set [25], which has 181 images. These images have a relatively large resolution of 1024×768 pixels. The images were viewed by seven subjects and the focused

image regions are usually faces and some other small objects, such as cell phones, toys, and so on. The fourth data set is **NUSEF** [22] with 758 images. The image contents contain rich semantics and affective contents, such as expressive faces, interesting objects, human actions, and so on, making this data set very challenging. Each image in this data set is viewed by 25 subjects on average. We only used 431 images in our experiments due to the copyright issue. The fifth data set is the **DUT-O** data set [43], which consists of 5168 images with the largest height or width of 400 pixels. Images of this database have one or more salient objects and relatively complex backgrounds. Each image is viewed by 5 subjects; then, a postprocessing step is applied to remove outlier eye fixation points that do not lie on a meaningful object. The sixth data set, **OSIE** [44], contains 700 images with the resolution of 800×600 pixels. The images are natural indoor and outdoor scenes, and aesthetic photographs from Flickr and Google, usually containing multiple dominant objects. Each image is viewed by 15 subjects. The last data set is the **SBU** data set [45], which contains 1000 images selected from the Pascal VOC2008 data set. Images in this data set are usually different types of daily life pictures, including various sceneries, animals, portraits, objects, and so on. Each image is viewed by three subjects.

B. Evaluation Metrics

The area under the receiver operating characteristic (ROC) curve (AUC) [13] metric is widely used to evaluate saliency models. Given an image and its GT eye fixation points, fixated points and other ones are regarded as the positive and negative sets, respectively. When computing the AUC score, the obtained saliency map is normalized to [0], [1] first. Then, it is binarily classified into salient regions and nonsalient regions by using a threshold. Through varying the threshold from 0 to 1, ROC curves can be obtained by plotting true positive rate versus false positive rate. Finally, the AUC is calculated as the AUC score. Nevertheless, AUC could be largely influenced by center-bias [46] and border cut [47]. It may score a central Gaussian blob highly, resulting in unfair evaluation. To deal with this problem, shuffled AUC is proposed by [46] and [47]. Different from AUC, shuffled AUC adopts all fixation points over all images from the same data set (except for the positive set), namely, the shuffled fixation map, as the negative set. Shuffled AUC does not benefit from center-bias. It scores 0.5 for a central Gaussian blob while a perfect prediction obtains a score of 1. Due to shuffled AUC score is sensitive to the levels of blurring of saliency maps, following many recent works [5], [7], we use small Gaussian filters to smooth the generated saliency maps with various standard deviations (STDs) σ. Then, the curve of average shuffled AUC scores over a data set versus various σ is shown and the best score under the optimal σ is reported to evaluate a model.

C. Implementation Details

1) Data Processing: We randomly sampled 60% images from the DUT-O, MIT, OSIE, and SBU data sets as the training set to train the Mr-CNN since the four data sets contain large amount of images with abundant visual contents, including semantic objects and complex backgrounds. Then, we tested our trained model on the rest images and other data sets.

To boost the generalization performance of Mr-CNN, we augmented training images via mirror-image flipping thus to increase training samples twice. When extracting the central and the surrounding image regions, when the center pixel is quite close to image borders, we cannot obtain sufficient pixels to extract the needed image region. For this case, image borders were copied to pad image regions. We also precomputed the mean value and the STD for each of the RGB channel of all pixels over the training set. Then, each image region was mean-centered and normalized to unit variance along the RGB channels using the precomputed parameters before it was inputted to the Mr-CNN.

During training, for each training image, we randomly picked out 10 fixated pixels and 40 nonfixation pixels to train the Mr-CNN. To be specific, we directly considered fixation pixels as those whose saliency values are greater than 0.8. As for the nonfixation pixels, if we just randomly sample pixels whose saliency values are smaller than a small threshold, the sampled pixels would mostly lie on regions near image borders, which are usually trivial backgrounds. As we have incorporated global contrast into our model, we would like to have some hard negative training samples that are distinctive in local contexts but trivial in the global context to learn global contrast. Thus we proposed a shuffled-AUC-like nonfixation pixel sampling scheme. In detail, for each image, we sampled nonfixation pixels whose saliency values are smaller than a threshold (0.2 in our experiment) from the shuffled fixation points. Thus, the sampled nonfixation points are at the locations of the fixation points of other images, many of which had complex image contents. As a result, this sampling strategy can guide the Mr-CNN to learn to deal with hard negative samples well via learning global contrast.

When testing, to reduce computational costs, we just evenly sampled 50×50 pixels for each testing image. Then, for each sampled pixel, we fed forward its corresponding multiresolution image regions to the Mr-CNN and got the output classification probability as its saliency value. Finally, the obtained 50×50 saliency map was rescaled to the original size of the testing image. We also horizontally flipped each testing image to generate two saliency maps and take the average map as the final saliency map.

2) Training Strategies: We implemented the Mr-CNN based on the deepnet[1] library using a GTX Titan black GPU to accelerate. During training the Mr-CNN, we set the size of minibatch to 128, training step to 60 000, and weight decay to 0.0002. We initially set the learning rate to 0.001 and afterward decrease it along with the increase of training step. We set the momentum to linearly increase from 0.8 to 0.99 along with the increase of training step. Meanwhile, to avoid overfitting, we used the rest 40% MIT images to validate the trained Mr-CNN models every 3000 training steps and selected the trained model with the best validation performance to do testing.

[1] https://github.com/nitishsrivastava/deepnet

TABLE II

AVERAGE SHUFFLED AUC SCORES WITH OPTIMAL GAUSSIAN BLURRING. THE BEST SCORE ON EACH DATA SET IS BOTH IN BOLD FACE FONT AND UNDERLINED, THE SECOND BEST IS SHOWN IN BOLD FACE FONT

Dataset	AWS [6]	BMS [7]	CA [8]	eDN [31]	HFT [48]	ICL [14]	IS [49]	LG [5]	QDCT [50]	SDSR [51]	Mr-CNN
Toronto [13]	0.7182	**0.7214**	0.6954	0.6733	0.6918	0.6982	0.7097	0.6997	0.7173	0.7069	**0.7256**
Opt. σ	0.010	0.015	0.025	0.010	0.035	0.015	0.040	0.035	0.015	0.040	0.015
MIT [3]	0.6916	**0.6923**	0.6671	0.6831	0.6522	0.6629	0.6640	0.6744	0.6655	0.6595	**0.7278**
Opt. σ	0.010	0.025	0.025	0.020	0.015	0.020	0.040	0.035	0.020	0.060	0.020
Cerf [25]	0.7242	**0.7357**	0.7143	0.7138	0.6999	0.7250	0.7267	0.7027	0.7269	0.7235	**0.7853**
Opt. σ	0.010	0.020	0.025	0.020	0.030	0.020	0.045	0.035	0.025	0.045	0.020
NUSEF [22]	**0.6401**	0.6322	0.6166	0.6322	0.6061	0.6097	0.6207	0.6259	0.6171	0.6103	**0.6694**
Opt. σ	0.020	0.030	0.030	0.020	0.030	0.020	0.045	0.045	0.020	0.050	0.025
DUT-O [43]	**0.7336**	0.7319	0.7223	0.6570	0.6808	0.6995	0.7103	0.6793	0.7129	0.7105	**0.7595**
Opt. σ	0.035	0.030	0.030	0.010	0.020	0.030	0.050	0.030	0.030	0.050	0.030
OSIE [44]	0.7541	**0.7637**	0.7390	0.7216	0.6928	0.7397	0.7301	0.7605	0.7369	0.7193	**0.8121**
Opt. σ	0.010	0.010	0.025	0.010	0.010	0.015	0.040	0.030	0.020	0.050	0.010
SBU [45]	0.6577	**0.6799**	0.6597	0.6514	0.6210	0.6484	0.6528	0.6479	0.6590	0.6511	**0.7143**
Opt. σ	0.040	0.040	0.040	0.010	0.035	0.040	0.050	0.050	0.040	0.065	0.040
Average	0.7028	**0.7082**	0.6878	0.6761	0.6635	0.6833	0.6878	0.6843	0.6908	0.6830	**0.7420**

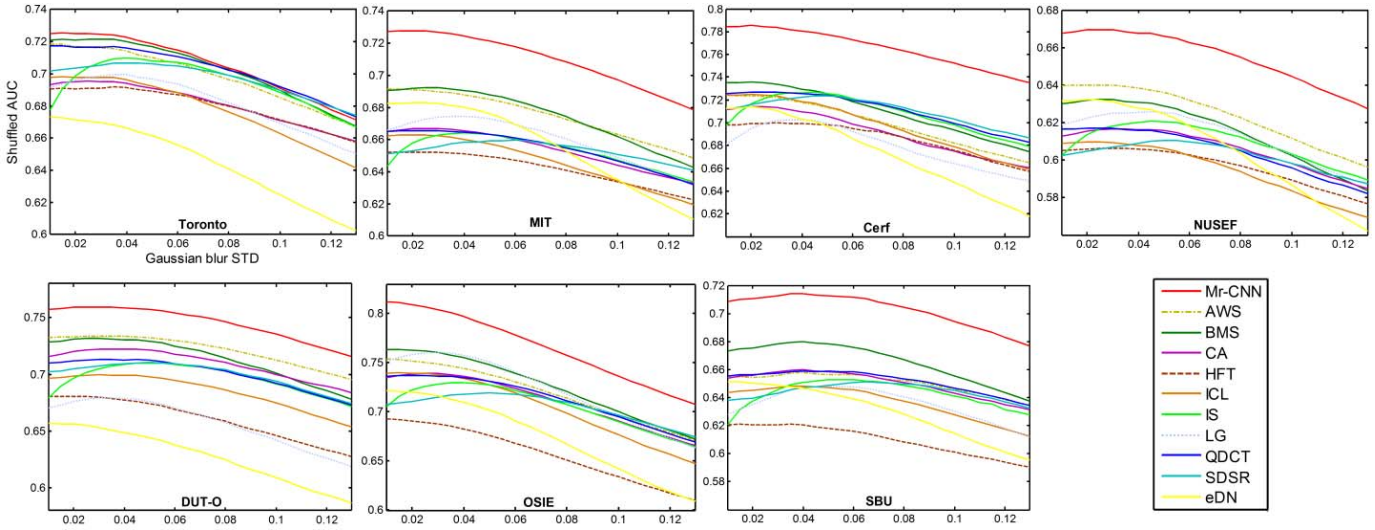

Fig. 3. Average shuffled AUC scores against the STDs of Gaussian blur. The proposed Mr-CNN model is compared with other ten state-of-the-art models on seven benchmark data sets. The x-axis indicates the Gaussian blur STD σ which is proportional to the largest dimension of the images. The y-axis indicates the average shuffled AUC score on each data set. Best viewed in color.

D. Comparison With State-of-the-Arts

To demonstrate the effectiveness of our proposed Mr-CNN for the task of eye fixation prediction, we compared it with other ten state-of-the-art models, including AWS [6], BMS [7], CA [8], eDN [31], HFT [48], ICL [14], IS [49], LG [5], QDCT [50], and SDSR [51]. As for eDN model, we used the saliency maps without center bias to calculate the shuffled AUC scores for fair comparison. At first, the shuffled AUC scores of our model and other ten models were computed for quantitative comparison. Gaussian kernels with various blur STD σ were used to smooth saliency maps first. Then, Fig. 3 shows average shuffled-AUC scores of each model on seven data sets by varying σ. Table II reports the best score of each model on each data set and its corresponding optimal Gaussian blur STD.

As can be seen in Fig. 3 and Table II, our Mr-CNN model can consistently obtain the best performance on all seven benchmark data sets. Specifically, it outperforms other ten models by a large margin on almost all the tested data sets, except for the Toronto data set, on which our model is slightly better than AWS, BMS, and QDCT. This may be from the fact that the Toronto data set mainly composes of relatively simple images and contains trivial objects, which can be handled effectively by traditional methods. Among compared models, AWS and BMS performed much better than the others. However, they are still significantly outperformed by the proposed Mr-CNN, especially on the data sets containing complex images and semantic meaningful objects, e.g., human, cars, and animals, demonstrating the superior capacity of deep models on the saliency detection task. We also notice that our model is much better than the eDN model, another deep neural network-based method, which indicates that the effectiveness of the designed local and global contrast-based multiresolution architecture.

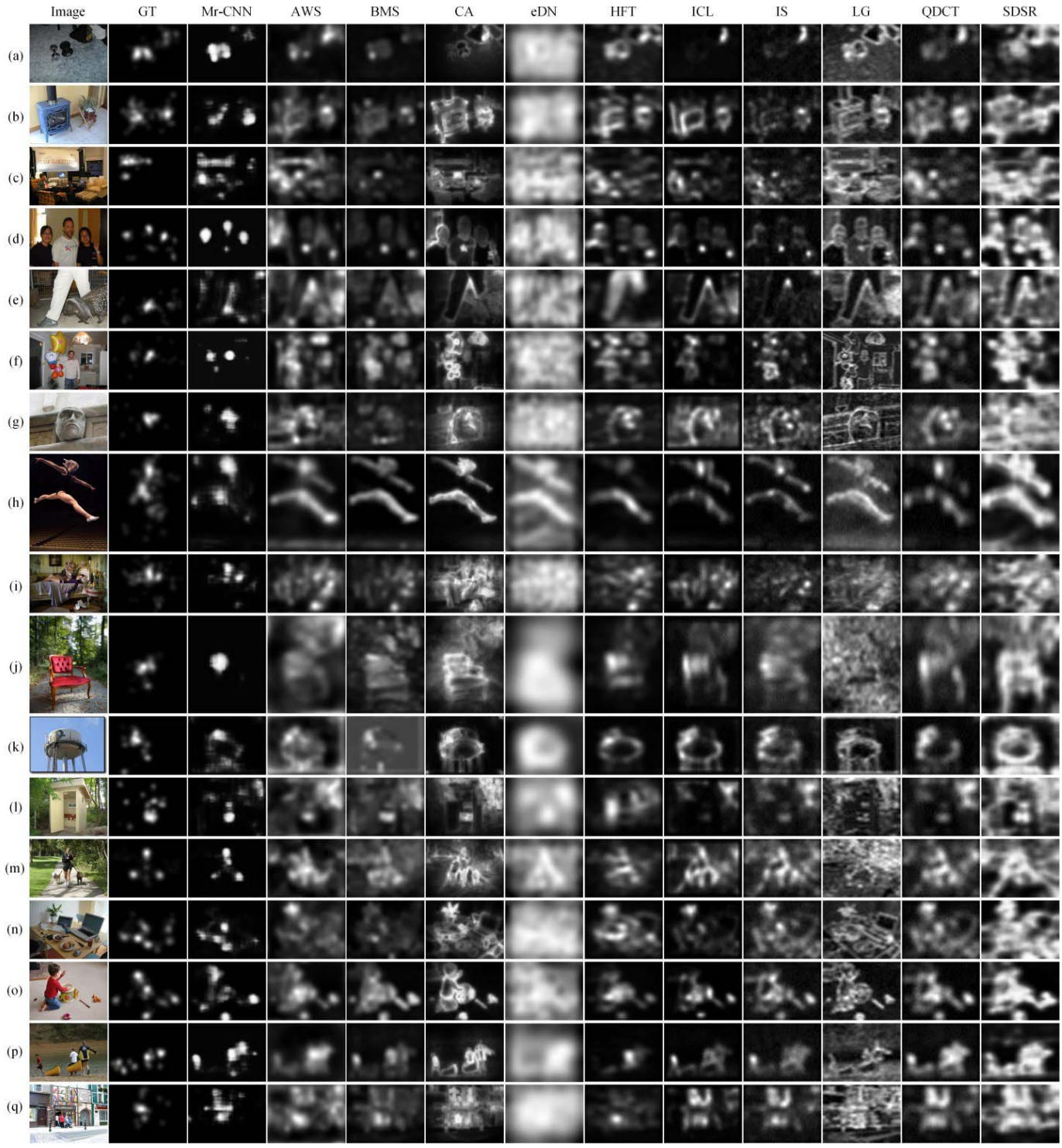

Fig. 4. Visual comparison of various models. We visualize a number of saliency maps of the proposed Mr-CNN model and other ten models. In the first column, we show (a) and (b) input images from the Toronto, (c)–(e) MIT, (f) and (g) Cerf, (h) and (i) NUSEF, (j)–(l) DUT-O, (m)–(o) OSIE, and (p) and (q) SBU data sets. In the second column, the corresponding GT fixation density maps that are yielded using Gaussian blur on the raw eye fixation point maps are shown. Best viewed in zoomed-in PDF file.

We also present the qualitative comparison of the proposed model with other ten approaches in Fig. 4. As can be seen, our Mr-CNN model can generate more accurate saliency prediction results than other methods. It can be less distracted by high-contrast edges and complex backgrounds, can detect bottom-up saliency patterns with diverse scales [see (a), (b), (j), (l), and (n)], and can also deal with local contrast and global contrast well [especially see (n) which contains many objects in one image]. What is more important, Mr-CNN can highlight various top-down factors, such as faces [see (c), (d), (f), (g), (i), and (m)], text [(c) and (q)], human heads [(h), (o), (p), and (q)], and animal heads [(e) and (m)]. It can also deal

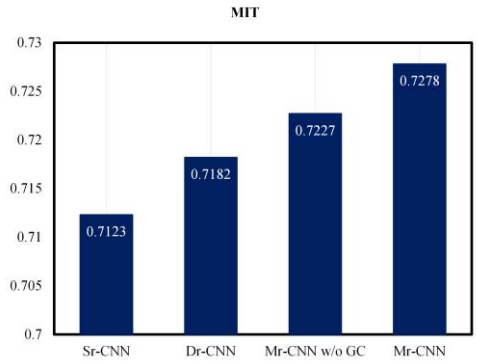

Fig. 5. Performance of models with different resolutions on the MIT data set. Sr-CNN: single-resolution-CNN with the central region; Dr-CNN: dual-resolution-CNN with the central and surrounding regions; Mr-CNN without GC: triple-resolution-CNN without the global context.

with the integration of contrasts and top-down factors well. To be specific, in (c) and (i), even though the images contain many attention-grabbing objects or complex backgrounds, Mr-CNN can still highlight semantic meaningful regions preferentially. While on the contrary, among traditional methods, CA and LG are apt to highlight object boundaries and high-frequency noises instead of real salient regions. SDSR and eDN are easily distracted by complex backgrounds. What is more important, it is very hard for these traditional models to perceive real semantic concepts, and deal with complex interactions among local contrast, global contrast, and top-down factors as the proposed Mr-CNN does.

E. Influence of Different Resolutions

Here, we analyze how the model performance is influenced by different resolutions. We trained another three networks, i.e., single-resolution-CNN with the central region (denoted as Sr-CNN), dual-resolution-CNN (denoted as Dr-CNN) with the central and surrounding regions, and triple-resolution-CNN without the global context (denoted as Mr-CNN without GC), for which we sampled 79×79 image regions from the whole image with size 159×159 for each pixel. We tested the performance of models with different resolutions on the MIT data set and the results are shown in Fig. 5. As we can see, the Sr-CNN performs the worst, but it still achieves the shuffled AUC score of 0.7123, which is much better than traditional methods, indicating the superiority of the features and top-down factors learned by deep networks. When surrounding regions are incorporated, the performance can be improved by local contrast inference (see Dr-CNN and Mr-CNN without GC). By comparing the performance between Mr-CNN without GC and Mr-CNN, we can see that, when the global context is included, the performance can be further improved by incorporating global contrast.

F. Feature Visualization

In order to understand the reason of the effectiveness of the proposed Mr-CNN, we visualize the features learned hierarchically in the convolutional layers. Since it is intractable to directly visualize the learned features of higher layers in a CNN, for each neuron, we uniformly show nine optimal image patches which activate it most strongly. We only show 64 neurons per layer due to the limitation of space, which forms an 8×8 grid (see Fig. 6). As can be seen from Fig. 6, the proposed Mr-CNN mainly learns a variety of edges and color blobs in layer 1, and corners and edge/color conjunctions in layer 2. As shown in the rightmost image in Fig. 6, the features obtained in layer 3 are quite informative. A number of low-level patterns, for example, complex corners [(Row 5, Col 1), (Row 5, Col 2), and (Row 6, Col 5)], edge conjunctions [(Row 1, Col 4), (Row 1, Col 5), (Row 1, Col 6), (Row 8, Col 8), and so on], complex textures [(Row 1, Col 1), (Row 1, Col 6), and so on] and other contrast-like patterns [(Row 7, Col 6), (Row 8, Col 1), and so on] can be found. These features are essentially relevant to saliency information. Some middle-level features are also learned, for example, the circle pattern [(Row 5, Col 1), (Row 7, Col 7), and (Row 7, Col 8)], the stripe pattern [(Row 5, Col 6) and (Row 6, Col 7)], and the grid pattern [(Row 7, Col 1) and (Row 7, Col 2)]. What is more important is that layer 3 also learns various high-level semantic concepts, for instance, human faces and heads (Row 2), text and signs (Row 3), and human body profiles [(Row 8, Col 2) and (Row 8, Col 3)], which are difficult to be handled in traditional methods. In conclusion, the visualization results indicate that the proposed Mr-CNN can infer bottom-up saliency factors and high-level top-down factors, which is an important reason for the superiority of our model.

G. Results on the RGB-D Image Saliency Detection Task

To evaluate the effectiveness of the proposed 3-D Mr-CNN model on the RGB-D image saliency detection task, we conducted experiments on the NUS3D data set [27], which is the only one public RGB-D eye fixation data set with a relatively large amount of images. This data set contains 600 RGB-D images. For each image, its GT eye fixation points are acquired from an average of 14 viewers. We randomly sampled 450 images for training the 3-D Mr-CNN and another 50 images for validation. Then, we tested our model on the rest 100 images. The data processing and the CNN parameters of the 3-D Mr-CNN were almost the same with the Mr-CNN. The differences are that we changed the input channels to 4, changed the training steps to 10 000, and changed the model evaluation and selection step interval to 500.

As the NUS3D data set published saliency maps instead of raw fixation points, we adopted the correlation coefficient (CC) [52] as the similarity measure between each saliency map and the corresponding GT saliency map to compare the proposed 3-D Mr-CNN and Lang's model [27]. We used the best CC score of 3-D fixation reported in [27] as their performance, i.e., the result achieved by combining depth prior with the Graph-Based Visual Saliency model (GBVS) [17] using the multiplication fusion scheme. We show the quantitative and qualitative results in Table III and Fig. 7, respectively. As shown in Table III, 3-D Mr-CNN performs better than Lang's model [27], which is mainly because the early fusion scheme and the deep architecture we adopted. As shown in Fig. 7, saliency

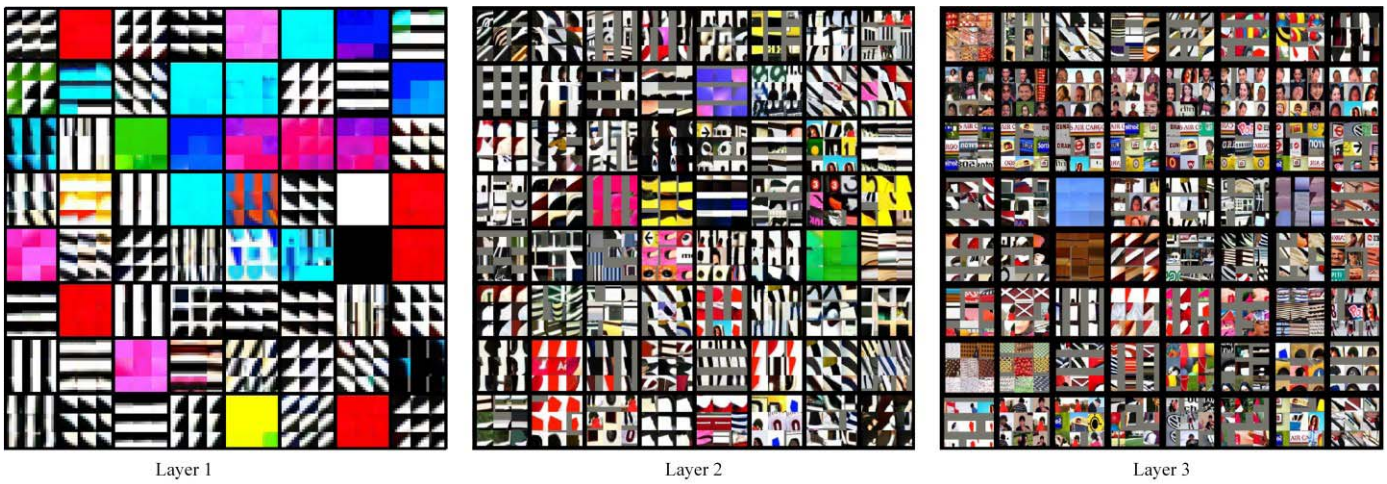

Fig. 6. Feature visualization. Best viewed in zoomed-in PDF file.

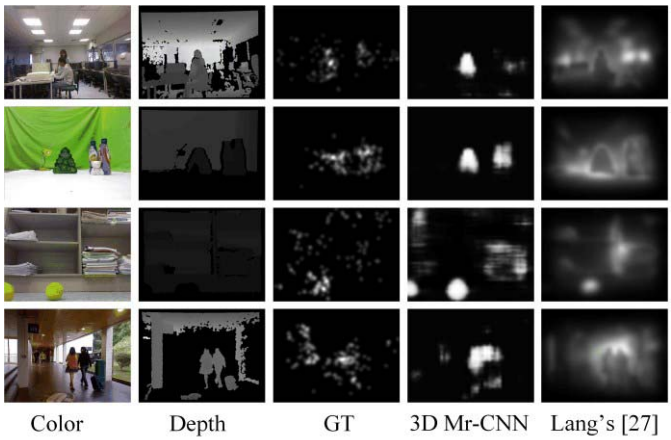

Fig. 7. Visual comparison. GT: ground truth saliency map.

TABLE III

COMPARISON OF CC ON NUS3D DATA SET. CC IS THE PEARSON CC BETWEEN GROUND TRUTH SALIENCY MAP G AND THE PREDICTED SALIENCY MAP S. $\text{CC} = \text{COV}(G, S)/(\sigma_G * \sigma_S)$

Methods	CC
Lang's [27]	0.4128
3D Mr-CNN	0.4454

maps produced by 3-D Mr-CNN match the GT well. Even though the color images contain complex backgrounds, 3-D Mr-CNN can still highlight salient regions accurately by combining depth information (see rows 1 and 4).

IV. CONCLUSION

This paper has proposed a novel deep CNN called Mr-CNN for eye fixation prediction, which can learn local contrast, global contrast, top-down factors, and their integration as well. The proposed Mr-CNN achieved better performance with significant improvements compared with other ten state-of-the-art saliency detection approaches on seven widely used saliency benchmark data sets. The superior performance of Mr-CNN implies that local contrast, global contrast, and high-level semantics are more likely to be processed jointly in human visual system rather than separately. The visualization of learned hierarchical features indicates that the proposed Mr-CNN learns both low-level bottom-up saliency cues and high-level semantic factors. More importantly, the architecture of our model can also facilitate the understanding of the intrinsic mechanism of human visual attention. We also applied our method to the RGB-D image saliency detection problem. By modeling depth and color saliency cues jointly and their interactions, the proposed 3-D Mr-CNN model achieved better performance than the traditional method.

In the future, we have two directions to work on. The first one is to develop more effective saliency models based on neural networks, e.g., using pretrained models and recurrent models. The second one is to leverage our saliency model to facilitate other tasks, e.g., salient object detection [56], [57], video saliency detection [58], multiresolution imaging [59], scene classification [60]–[62], and object detection [63].

REFERENCES

[1] E. Niebur and C. Koch, "Computational architectures for attention," in *The Attentive Brain*. Cambridge, MA, USA: MIT Press, 1998, pp. 163–186.
[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[3] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2106–2113.
[4] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 438–445.
[5] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 478–485.
[6] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.*, vol. 30, no. 1, pp. 51–64, Jan. 2012.
[7] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.
[8] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[9] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[10] F. Liu and M. Gleicher, "Video retargeting: Automating pan and scan," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 241–250.

[11] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1511–1518.

[12] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2004, pp. 37–44.

[13] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 155–162.

[14] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 681–688.

[15] B. Han, H. Zhu, and Y. Ding, "Bottom-up saliency based on weighted sparse coding residual," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1117–1120.

[16] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *J. Vis.*, vol. 8, no. 7, p. 13, 2008.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[18] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[19] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, p. 9, 2011.

[20] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, p. 18, 2008.

[21] L. Elazary and L. Itti, "Interesting objects are visually salient," *J. Vis.*, vol. 8, no. 3, p. 3, 2008.

[22] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 30–43.

[23] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vis. Res.*, vol. 46, no. 26, pp. 4333–4345, Dec. 2006.

[24] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, p. 4, 2007.

[25] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 241–248.

[26] N. Ouerhani and H. Hügli, "Computing visual attention from scene depth," in *Proc. IEEE 15th Int. Conf. Pattern Recognit.*, Sep. 2000, pp. 375–378.

[27] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 101–115.

[28] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," in *Proc. BMVC*, 2013, pp. 9–13.

[29] C. Shen, M. Song, and Q. Zhao, "Learning high-level concepts by training a deep network on eye fixations," in *Proc. NIPS Deep Learn. Unsupervised Feature Learn. Workshop*, 2012.

[30] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[31] Y. Lin, S. Kong, D. Wang, and Y. Zhuang, "Saliency detection within a deep convolutional architecture," in *Proc. AAAI Workshop*, 2014, pp. 31–37.

[32] M. Kümmerer, L. Theis, and M. Bethge. (2014). "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet." [Online]. Available: https://arxiv.org/abs/1411.1045

[33] 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.

[34] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu. (2015). "DeepFix: A fully convolutional neural network for predicting human eye fixations." [Online]. Available: https://arxiv.org/abs/1510.02927

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[36] 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 Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 262–270.

[37] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[41] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.

[42] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. (2012). "Improving neural networks by preventing co-adaptation of feature detectors." [Online]. Available: https://arxiv.org/abs/1207.0580

[43] 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.*, Jun. 2013, pp. 3166–3173.

[44] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, p. 28, 2014.

[45] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. Berg, "Studying relationships between human gaze, description, and computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 739–746.

[46] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, Mar. 2005.

[47] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

[48] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.

[49] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[50] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 116–129.

[51] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.

[52] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri, "Empirical validation of the saliency-based model of visual attention," *Electron. Lett. Comput. Vis. Image Anal.*, vol. 3, no. 1, pp. 13–24, 2004.

[53] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 362–370.

[54] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.

[55] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-stage learning to predict human eye fixations via SDAEs," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 487–498, Feb. 2016.

[56] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 678–686.

[57] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.

[58] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[59] X. Lu and X. Li, "Multiresolution imaging," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 149–160, Jan. 2014.

[60] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[61] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.
[62] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.
[63] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

Nian Liu received the B.E. and M.E. degrees from Northwestern Polytechnical University, Xi'an, China, in 2012 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the School of Automation.

His current research interests include computer vision and multimedia processing, especially on saliency detection and deep learning.

Junwei Han (M'12–SM'15) is a currently a Full Professor with Northwestern Polytechnical University, Xi'an, China. His current research interests include computer vision, multimedia processing, and brain imaging analysis.

Dr. Han is an Associate Editor of the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, *Neurocomputing*, *Multidimensional Systems and Signal Processing*, and *Machine Vision and Applications*.

Tianming Liu (SM'14) received the Ph.D. degree in computer science from Shanghai Jiaotong University, Shanghai, China, in 2002.

He was a Faculty Member with the Weill Medical College, Cornell University, Ithaca, NY, USA, and Harvard Medical School, Boston, MA, USA. He is currently an Associate Professor of Computer Science with the University of Georgia, Athens, GA, USA. His current research interets include computational brain imaging.

Dr. Liu was a recipient of the Microsoft Fellowship Award and the NIH NIBIBK01 Career Award.

Xuelong Li (M'02–SM'07–F'12) is currently a Full Professor with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China.