A Novel Method to Study Bottom-up Visual Saliency and Its Neural Mechanism

Cheng Chen, Xilin Zhang, Yizhou Wang, and Fang Fang

Abstract—In this study, we propose a novel method to measure bottom-up saliency maps of natural images. In order to eliminate the influence of top-down signals, backward masking is used to make stimuli (natural images) subjectively invisible to subjects, however, the bottom-up saliency can still orient the subjects attention. To measure this orientation/attention effect, we adopt the cuing effect paradigm by deploying discrimination tasks at each location of an image, and measure the discrimination performance variation across the image as the attentional effect of the bottom-up saliency. Such attentional effects are combined to construct a final bottom-up saliency map. Based on the proposed method, we introduce a new bottom-up saliency map database of natural images to benchmark computational models. We compare several state-of-the-art saliency models on the database. Moreover, the proposed paradigm is applied to investigate the neural basis of the bottom-up visual saliency map by analyzing psychophysical and fMRI experimental results. Our findings suggest that the bottom-up saliency maps of natural images are constructed in V1. It provides a strong scientific evidence to resolve the long standing dispute in neuroscience about where the bottom-up saliency map is constructed in human brain.

Index Terms—visual attention, bottom-up visual saliency measurement, saliency benchmark database, bottom-up saliency neural mechanism.

I. INTRODUCTION

Visual attention is essential for many cognitive processes, e.g. scene recognition. Information selection in these processes can be executed by top-down signals voluntarily, or triggered by salient stimuli in a bottom-up way. The former is driven by specific goals, while the latter is guided by a bottom-up saliency map. A bottom-up saliency map is defined as a topographical map to describe and predict the distribution of attention attraction on a visual input [22]. It is able to efficiently direct our focus of attention. Hence, understanding bottom-up saliency not only has a scientific value, but also can help us to build intelligent yet computation efficient models to facilitate real world applications such as object detection in clustered environment and unmanned vehicles ([31], [33]).

Many bottom-up saliency models have been proposed to simulate human intelligence in visual attention. Compared to the numerous proposed bottom-up visual saliency computational models, methods about how to accurately measure/quantize bottom-up saliency on natural images are under-studied. Currently, in order to evaluate computational saliency models, researchers use eye tracking data to generate groundtruth, e.g. [5], [35]. In traditional methods, participants are asked to free view an image presented on a monitor for a short period of time. After that, fixations of each participant are extracted by certain computational algorithms [5] from the original eye tracking data. And then, by fusing all the fixations across participants, researchers build a saliency benchmark dataset.

However, using eye tracking data to build a bottom-up saliency benchmark dataset has several disadvantages. First, eye fixations are mostly used to study overt attention. They are very easily to be influenced by top-down signals and center bias, and they are considered to be lack of precision and have large individual variances. Second, eye fixations are estimated by computational algorithms from raw eye tracking dataset, hence, the saliency values derived from the estimated fixations could be easily influenced by different algorithms. Therefore, it is not ideal/appropriate to use eye fixation data as the groundtruth to evaluate bottom-up saliency models. To solve this problem, it is imperative to propose a rigorous method to measure the bottom-up saliency map free from top-down signals.

In this paper, we propose a new method to measure bottom-up saliency maps of natural images. In the proposed paradigm, top-down control is blocked using backward masking method. In this way, stimuli (natural images) are made subjectively invisible to subjects, however, bottom-up saliency can still orient the subjects’ attention. To measure this orientation/attention effect, we adopt the cuing effect paradigm by deploying discrimination tasks at each location of an image, and measure the discrimination performance variation across the image as the attentional effect of the bottom-up saliency. Such attentional effects are integrated to construct a final bottom-up saliency map. Based on the proposed method, we introduce a new bottom-up saliency map dataset of natural images to benchmark computational models. Moreover, we evaluate several state-of-the-art saliency models using the dataset.

To justify the proposed paradigm, it is applied to investigate the neural basis of the bottom-up visual saliency by analyzing psychophysical and fMRI experimental results, especially focusing on identifying the brain areas that construct the bottom-up saliency maps of natural images. Currently, there are controversies about where the bottom-up saliency maps are constructed in human brain. Many brain regions that are believed to realize bottom-up saliency maps receive both bottom-up and top-down signals, and the judgement could be made based on the mixed signals rather than the pure bottom-up one. By using the proposed paradigm, top-down signals can be effectively eliminated, which makes it possible to identify the brain areas that realize the bottom-up saliency maps.

We design novel psychophysical and physiological experiments to locate the brain areas that construct bottom-up visual saliency. In the psychophysical experiment, by using the proposed method, we measure the attentional effect induced by
bottom-up saliency. In the physiological experiment, we record the blood-oxygen-level-dependent (BOLD) signals responding to the subjectively invisible natural images using functional magnetic resonance imaging (fMRI). We find that when top-down signals are eliminated, (i) the attentional effect increases with the degree of saliency, (ii) the BOLD signals increase with the degree of saliency only in V1, but not in other cortical areas such as V4, IPS or FEF, and more importantly, (iii) the attentional effect of bottom-up saliency significantly correlates with the BOLD signal in V1 across subjects. These findings suggest that bottom-up saliency is constructed in V1, which challenges the argument that bottom-up saliency can only be constructed in higher cortical areas such as IPS or FEF [3, 15].

In summary, the proposed has the following contributions. (i) We develop a novel method to measure bottom-up saliency maps of natural images, and construct a dataset to benchmark bottom-up saliency models more precisely. (ii) The research result is a step toward resolving a fundamental scientific dispute about where the bottom-up saliency is constructed in brain. (iii) The finding also suggests a clear reference to implement biologically-plausible computational models of bottom-up visual saliency, for example, by simulating the function of V1 neurons.

A. Related Work

There is a huge body of research focusing on studying competing computational bottom-up saliency models. However, few of them investigate a proper method to measure the bottom-up saliency maps of natural images in computer vision field. Bruce and his colleagues suggest to evaluate the computed saliency maps by using eye fixations collected in an image viewing task [5]. In their method, subjects are asked to observe an image presented on a CRT monitor without any particular instructions. For each image, a fixation density map is produced by convolving a 2D Gaussian function with each fixation collected in this image. The generated fixation density map could be further used to compare with a computed saliency map, so that the performance of a computational saliency model can be evaluated. Many other studies adopt the eye fixation dataset proposed by Bruce [5] to evaluate their own models, e.g., [10], [20], [36]. However, using eye fixation data to infer the distribution of bottom-up attention is far from satisfaction. There are two major drawbacks. First, eye fixations are very sparse. The saliency values of the locations not visited by eye fixations are not measurable; they can only be estimated via smoothing or set to 0 as in [5], which can be inappropriate. Second, as aforementioned, eye fixations can be easily influenced by top-signals [8] or the prior knowledge/experience of individuals. Therefore, the attentional effect resulted from the bottom-up saliency is hard to differentiate. Therefore, to evaluate/benchmark bottom-up saliency models, it is imperative to propose a new method to measure saliency maps of natural images derived from pure bottom-up signals and it should be able to measure saliency values at each location of an image.

To block the top-down signals, several methods can be used to achieve this goal, such as continuous flash suppression [13] and backward masking [4]. In the paradigm of backward masking, a stimulus is presented very briefly and followed by a high contrast mask, so that subjects cannot be aware of the stimulus [37]. The attentional attraction can be measured by using a modified version of the cueing effect paradigm proposed by Posner et al. [30]. In this paradigm, a target appears in one of two locations in the peripheral visual field randomly, and subjects need to perform a discrimination task with this target. Before the target presentation, a cue will indicate the location of the following target. Trials with a correct cue belong to the valid cue condition, while trials with an incorrect cue belong to the invalid cue condition. The attentional attraction could be measured by comparing the performance (response time or accuracy) between the two conditions [11].

Such a paradigm can also be used to solve the controversy about the neural basis of bottom-up saliency maps. Several studies had found that different brain regions could realize the saliency map. Subcortical structures such as superior colliculus [14] and pulvinar [34] were found to be able to construct the saliency map. Mazer and Gallant found that extrastriate ventral area V4 could realize a retinotopic saliency map to guide eye movement, providing evidence that brain activities in the ventral pathway could be correlated with the saliency map [28]. Their conclusion was also supported by Asplund and his colleagues, who found that the ventral network can account for stimulus-driven attention [1]. Geng and Mangun found that the anterior intraparietal sulcus (aIPS) was sensitive to the bottom-up influence driven by stimulus saliency [15]. Besides, frontal eye fields (FEF) was also found to play an important role in decoding the winner-take-all (WTA) stage of saliency processing [3]. Most of these findings were consistent with a dominant view, which argues that the final saliency map results from pooling different visual features channels after each visual feature channel construct its salient map independently. Thus, higher brain regions could be the possible candidates that realize the saliency map. However, these higher brain regions receive both bottom-up and top-down signals [16], it is hard to justify that the pure bottom-up saliency is constructed there.

On the other hand, Li proposed the V1 theory, which argues that neural activities in V1 could create the bottom-up saliency map via intracortical interactions that are manifested in contextual influences [25, 26]. Several studies also support the V1 theory. By measuring reaction times searching for a singleton that differs from its surrounding in more than one feature, Koene and Zhaoping found their results were consistent with the properties of some V1 neurons [23], which provides evidence for the V1 hypothesis of the bottom-up saliency map. Furthermore, Zhang and his colleagues measured the attentional effect of the bottom-up saliency map of simple texture stimuli, and found the degree of attention attraction correlated with the amplitude of the earliest component of the ERP as well as the V1 BOLD signal across subjects [37], which also supported the V1 hypothesis.

An important reason that the controversy exists is that most of these brain regions receive both bottom-up and top-down signals, which makes it difficult to determine whether
the saliency map they realized truly reflected the bottom-up attention attraction or not. In order to investigate the neural basis of the bottom-up saliency map, it is important to probe bottom-up signals free from top-down signals. Based on this requirement, Zhang and his colleagues used simple oriented bars as their stimuli, and found their results supported the V1 hypothesis [37]. However, as we know, neurons in V1 are highly tuned to these oriented bars. Such difference in stimulus preference between V1 and other cortical areas makes their conclusion not so convincing [2]. Moreover, our world contains much more complex visual features, thus the V1 hypothesis should be verified on natural scenes.

The rest of this paper is organized as follows. In Section 2, we introduce the material and method to measure the bottom-up saliency maps. The comparison results among different existing saliency models are presented in Section 3. In Section 4, we introduce our psychophysical and fMRI experiments to investigate the neural basis of the bottom-up saliency map of natural images. The results of these experiment are also presented in Section 5. Finally, we conclude and discuss the paper in Section 6.

II. MEASURING THE BOTTOM-UP SALIENCY MAP OF NATURAL IMAGES

In this section, we will introduce the protocol of our method to measure the bottom-up saliency map of natural images, followed by some descriptions about the visual stimuli and participants.

A. Experimental protocol

We perform psychophysical experiments measure bottom-up saliency maps from human. In the main experiment, visual stimuli are presented on a Gamma-corrected Iiyama HM204DT 22 inches monitor, with a spatial resolution of 1024 × 768 and a refresh rate of 60Hz. The viewing distance is 83cm. Participants’ head positions are stabilized using a chin rest and a head rest. A white fixation cross is always presented at the center of the monitor, and participants are asked to fixate the cross throughout the experiment.

To measure the bottom-up saliency map of a natural image, a low-luminance image (namely the image condition) or a blank (namely the baseline condition) is presented at the center of the screen for 50 ms, followed by a 100 ms full screen mask, and then 50 ms fixation interval. The bottom-up saliency map of the image serves as a cue to attract spatial attention, and the mask could ensure the image subjectively invisible to participants [4]. Then a Gabor orientates at ± 1.5° away from the vertical is presented at a random location within the area of the image. Participants are asked to press one of the two keys to indicate the orientation of the Gabor (left or right tilted). The duration of each trial is 2s. Fig. 1 shows the protocol of a trial. Each condition consists of 10 blocks of 96 trials. The location and the orientation of the Gabors are counterbalanced across trials.

At each location of an image, we measure the performance variation of the orientation discrimination task between the image and the baseline conditions. Such difference indicates the degree of attentional attraction due to the bottom-up saliency. Compared with the baseline condition, locations in the image condition with higher bottom-up saliency will attract more attention to improve the orientation discrimination task performance, while other locations will attract less attention so that the performance at these locations have less improvement.

Moreover, the mask following the image could ensure that top-down signals are eliminated from the experiment [4]. To ensure this precondition, subjects are asked to complete a two-alternative forced choice (2AFC) experiment in a criterion-free way before the main experiment. The protocol of this experiment is similar to the main experiment. Each trial begins with either a low-luminance image or a blank, followed by a full screen mask. Participants are asked to make a forced choice response to judge whether there is an image presented before the mask or not. We adjust the image luminance so as to make the subjects’ performance at chance level, this ensures that the masked images are indeed subjectively invisible, consequently, eliminate the top-down signals in the experiment.

B. Visual stimuli

There are three kinds of visual stimuli used in our experiment: natural images, masks and Gabors.

Natural images: 40 representative natural images are selected from the Internet and several public datasets. They cover a variety of image categories including ocean, mountain, open-country, street, tower, wild animal, etc. All these images are scaled to the same size (480 × 640 pixels, 14.54° × 19.38° of visual angle) with a mean low luminance (2.9 cd/m²). Such low luminance is aimed to ensure that these images could be easily masked so that top-down signals could be eliminated.

Mask: Mask stimuli are high contrast checkerboards with randomly arranged checkers. The size of each checker is about

---

Fig. 1. The psychophysical procedure to measure the bottom-up saliency map of natural images.
0.25° × 0.25°. The luminance of black checkers is 1.8 cd/m², and the luminance of white checkers is 79 cd/m².

**Gabor**: Gabor is a kind of visual pattern that human visual system is sensitive to. In our experiment, the Gabor has a spatial frequency of 5.5 cpd (cycle per degree) and its diameter is 2.5° with full contrast. The Gabor orientates at ± 1.5° away from the vertical.

C. Participants

Twenty human subjects (10 female and 10 male, 18-26 years old) participate in the experiment. All participants are naïve to the purpose of the study. All of them are right-handed, report normal or corrected-to-normal vision, and have no known neurological or visual disorders. They give written, informed consent in accordance with the procedures and protocols approved by the human subjects review committee of Peking University.

III. A DATASET OF THE BOTTOM-UP SALIENCY MAP OF NATURAL IMAGES

As mentioned above, using eye tracking data as the groundtruth of bottom-up saliency maps is not appropriate. The proposed method provides a viable means - it is able to measure the bottom-up saliency value at any position of an image without being affected by top-down signals. Based on the proposed method, we propose a more precise bottom-up saliency map dataset of natural images, so that bottom-up saliency models can be benchmarked.

In our experiment, each image is partitioned into an imaginary 6 × 8 grid. Although such resolution might be low, compared with traditional eye tracking methods, the density of the sample points is larger and our measurement covers the whole image. Moreover, the resolution can be increased by adding more sample points. We measure the saliency value at the center of each cell and assign the value to the rest pixels within the cell. We normalize the saliency values to [0, 1] on each image. Finally, the obtained map is convolved with a 2D Gaussian function to generate a smooth bottom-up saliency map. For each image, we collect data from 20 subjects. We apply this measurement with 40 natural images to build a bottom-up saliency map dataset.

We compare the Receiver Operator Characteristic (ROC) curves and the ROC areas of several state-of-the-art bottom-up computational saliency models: AIM [5], SWD [10], CAS [18], HOU [20], ITTI [21], and SER [36] on our dataset. The reason we select these saliency models is that they have public code. In the ROC analysis, the values of the measured saliency maps are normalized to [0, 1], and we use 0.5 as the threshold to binarize the saliency maps to salient regions and non-salient ones as the ground truth. To evaluate the computed saliency maps (also normalized), we adjust the threshold from 0 to 1, and identify the false positives and true positives at each threshold, then generate the ROC curves. The ROC curves are shown in Fig. 2, and the ROC areas are shown in Table I. The larger the ROC area, the better.

It can be seen that several bottom-up saliency models, such as ITTI [21], SWD [10], and SER [36] achieves better performance on our dataset. ITTI [21] is a traditional bottom-up saliency model based on the center-surround mechanism. SWD [10] is another bottom-up saliency model based on the spatially weighted dissimilarity. SER [36] simulates the signals transmission among the inter-connected V1 neurons, which is another bottom-up saliency model. Moreover, it is noticeable that the context-aware saliency model (CAS) proposed by Goferman et al. [18], which is based on top-down signals, has a poor performance on this dataset. More importantly, the settings of our experiment is convenient, e.g. our method does not require eye-tracking equipment, so that researchers could easily measure the bottom-up saliency map of natural images and evaluate their saliency models.

IV. NEURAL BASIS OF THE BOTTOM-UP SALIENCY MAP OF NATURAL IMAGES

Our paradigm could block the top-down signals, and thus also provide an efficient means to investigate the neural basis of the bottom-up saliency map of natural images. In our study, psychophysical and fMRI experiments using a similar paradigm are performed. In this section we will introduce the participants, the visual stimuli used in psychophysical and fMRI experiments, as well as the experimental procedures and data analysis processes for these two experiments.
A. Participants

Sixteen human subjects (9 female and 7 male, 19-26 years old) participate in the experiment. All of them participate in the psychophysical experiment. Thirteen of them participate in the fMRI experiment. Two subjects in the fMRI experiment are excluded from data analysis because of excessive head motion during fMRI scanning. All subjects are naïve to the purpose of the study except for one subject (one of the authors). All of them are right-handed, report normal or corrected-to-normal vision, and have no known neurological or visual disorders. They give written, informed consent in accordance with the procedures and protocols approved by the human subjects review committee of Peking University.

B. Visual stimuli

To study the neural basis of bottom-up visual saliency, the visual stimuli used in the psychophysical and fMRI experiments must be carefully processed. First we collect a large number of natural images from the Internet and several public datasets. All these images are scaled to the same size (11.63° × 31.03° of visual angle) with a mean low luminance (2.9 cd/m²) (Fig. 3A and Fig. 3B). Then we calculate the computational saliency map of each image by using the bottom-up visual saliency model proposed by Itti and his colleagues [21]. Then, we select 50 images for the current experiment. Each of these images has only one round salient region centers at about 7.2° eccentricity in the lower left quadrant (left salient images). The diameter of the salient region is about 4° (Fig. 3C and Fig. 3D). Moreover, to measure the bottom-up saliency map quantitatively, we classify all images into two groups: the high salient group and the low salient group, based on the proposed saliency index in the following formulation:

\[
Index(n) = \frac{S_I(n) - S_O(n)}{S_O(n)}, \quad (1)
\]

where \( n \) denotes the index of an image. \( S_I \) denotes the averaged saliency value of the round salient region in Fig. 3C and Fig. 3D, and \( S_O \) denoted the averaged saliency value of the residual region (the rest of the image). The higher value of the \( Index \), the higher saliency of the round region relative to the rest of the image. We selected half of images with an upper 50% \( Index \) as the high salient image set, and the rest as the low salient image set. An example of the high salient and the low salient images can be found in Fig. 3A and Fig. 3B. The averaged saliency maps of the 25 high salient images is shown in Fig. 3C, and the low salient counterpart is shown Fig. 3D.

In order to balance the location of the salient region, each of the left salient images is flipped horizontally to generate a new image, which has a salient region in the lower right quadrant (right salient images). Notice that the contents of the left and the right salient images are the same, the only difference between them is the location of the salient region. Thus, the stimuli used in the psychophysical and the fMRI experiments consist of two groups: the high salient group and the low salient group. Each group contains 50 images, half of which are the left salient images and the other half are the right salient ones. In both experiments, these stimuli are presented in the lower visual field on a dark screen (1.8 cd/m²).

Mask stimuli are high contrast checkerboards with randomly arranged checkers (Fig. 3E). The size of each checker is about 0.25° × 0.25°. The luminance of black checkers is 1.8 cd/m², while the luminance of white checkers is 79 cd/m².

C. Experimental procedure

Psychophysical experiment: Each trial starts with a fixation at the center of the screen. A natural image is presented on the lower half of the screen for 50 ms, followed by a 100 ms mask at the same position, and another 50 ms fixation interval. Then a Gabor centers at about 7.2° eccentricity from
the fixation is presented randomly at either the lower left quadrant or the lower right quadrant with equal probability for 50 ms. The location of the Gabor is either at the salient region of the preceding image or its contralateral counterpart, thus indicating the valid cue condition or the invalid cue condition. The Gabor has a spatial frequency of 5.5 cpd (cycle per degree) and its diameter is 2.5° with full contrast. The Gabor orientates at ± 1.5° away from the vertical. Subjects are asked to press one of the two keys to indicate the orientation of the Gabor (left or right tilted). The duration of each trial is 2s.

Fig. 3 shows the procedure of our experiment. The experiment consists of 10 runs. Each run contains 100 trials with two conditions: the high salient condition and the low salient condition. The images for the two conditions are randomly selected from the high and the low salient image groups correspondingly. The attentional effect of bottom-up saliency maps for each condition is quantified as the difference between the orientation discrimination task performance (discrimination accuracy) of the valid cue condition and the invalid cue conditions.

Moreover, we also ask subjects to complete an additional two-alternative forced choice (2AFC) experiment to check whether these natural images are subjectively invisible to subjects in our paradigm. In the additional experiment, each trial begin with either a low-luminance image or a blank, followed by a mask. Subjects are asked to make a forced choice response to judge whether there is an image presented before the mask or not (Fig. 3F). The performance at chance level in this experiment could provide an objective confirmation that the masked images are indeed invisible.

fMRI experiment: The fMRI experiment is conducted to investigate the neural basis of the bottom-up saliency map. In the fMRI experiment, an event-related design is adopted. The experiment consists of eight functional scans of 125 continuous trials. Each scan begins with 6s fixation and lasts 268s. There are five types of trials, including four types of salient image trials: two degrees of saliency (high and low) × two locations of salient region (left and right), and fixation trials. In all types of trials, a white fixation cross is always presented at the center of the screen. In a salient image trial, an image is presented on the lower half of the screen for 50 ms, followed by a 100 ms mask at the same position and 1850 ms fixation. Subjects are asked to indicate the location of the salient region, which is left to the fixation in half of salient image trials and right in the other half randomly. In a fixation trial, only the fixation point is presented for 2000ms. In a scan, there are 25 trials for each type. The order of the trials is counterbalanced across eight scans using M-sequences [6]. These are pseudo random sequences that have the advantage of being perfectly counterbalanced n trial back, so that each type of trials is preceded and followed equally by all types of trials, including itself.

Retinotopic visual areas (V1, V2, V3 and V4) are defined by a standard phase-encoded method developed by Sereno et al. [32] and Engel et al. [12], in which subjects view rotating wedge and expanding ring stimuli that create traveling waves of neural activity in visual cortex. A block-design scan is used to define the regions of interest (ROIs) in LGN, V1-
AC-PC space. The first 6s of BOLD signals are discarded to minimize transient magnetic saturation effects.

A general linear model (GLM) procedure is used for ROI analysis. The ROIs in LGN, V1-V4, LOC, IPS and FEF are defined by a localizer scan and retinotopic mapping scans (p < 10⁻⁸, uncorrected). Talairach coordinates of rLGN, ILGN, rFEF and IFEF are (21 ± 3, −26 ± 3, −1 ± 3), (−23 ± 3, −27 ± 2, −1 ± 2), (37 ± 7, −10 ± 5, 47 ± 4) and (−37 ± 6, −12 ± 4, 46 ± 3) respectively, consistent with previous studies (e.g. [27, 29]).

The event-related BOLD signals are calculated separately for each subject, following the method used by Kourtzi and Kanwisher [24]. For each event-related scan, the time course of the MR signal intensity is first extracted by averaging the data from all the voxels within the predefined ROI. The average event-related time course is then calculated for each type of trial, by selectively averaging from stimulus onset and using the average signal intensity during the fixation trials as a baseline to calculate percent signal change. Specifically, in each scan we average the signal intensity across the trials for each type of trial at each of 12 corresponding time points starting from the stimulus onset. These event-related time courses of the signal intensities are then converted to time courses of percent signal change for each type of trials by subtracting the corresponding value for the fixation trials and then dividing by that value. Because M-sequences have the advantage that each type of trials was precede and followed equally often by all types of trials, the overlapping BOLD responses due to the short interstimulus interval are removed by this averaging procedure [6]. The resulting time course for each type of trials is then averaged across scans and subjects.

F. Computational Saliency Model

We adopt an influential computational saliency model proposed by Itti et al. [21] to measure the bottom-up saliency map of each image. The model is based on the center-surround mechanism, and combined information from three channels: color, intensity and orientation. By using this model, we could predicate the degree of saliency of each image based on the formulation we proposed.

V. The Results of the Neural Basis of the Bottom-Up Saliency Maps

A. Psychophysical experiment

16 subjects participate in this experiment. In the additional 2AFC experiment, subjects report that they are unaware of the natural images. The percentages of correct detection (mean ± SEM) are 48.6±1.5% and 50.9±1.4% for the high salient and the low salient groups, respectively. The results are statistically indistinguishable from chance level, which indicates that the natural images in both two groups are subjectively invisible to subjects.

In the main experiment, considering that the salient region of a natural image could serve as a cue to attract attention, the attentional effect of the bottom-up saliency map of invisible natural images is quantified as the difference between the accuracy of the Gabor orientation discrimination performance in the valid cue condition, and that in the invalid cue condition. We find that the discrimination accuracy is higher in the valid cue condition than that in the invalid cue condition, for both high salient images (Valid: 81.31 ± 0.98%; Invalid: 72.88 ± 0.98%) and low salient images (Valid: 77.86 ± 0.93%; Invalid: 76.54 ± 0.88%). Thus, the attentional effect of the bottom-up saliency maps for the high salient and low salient groups are 8.43 ± 0.33% and 1.32 ± 0.47%, respectively (left panel in Fig. 5). The results indicate that the bottom-up saliency maps exhibit a positive cueing effect even when the image is subjectively invisible, which suggests that subjects attention is attracted to the salient region of the invisible images, so that they perform better in the valid cue condition than in the invalid cue condition. Moreover, the attentional effect in both two groups is significant (paired t-test high salient: t₁₅ = 18.126, p < 0.001; low salient: t₁₅ = 2.782, p = 0.014; significant level α = 0.05). The attentional effect of the high salient images is also significantly greater than that of the low salient images (t₁₅ = 9.665, p < 0.001). We also calculated the saliency index of the high salient and the low salient images based on the proposed formulation. The saliency index predicted the degree of the attention attraction of a bottom-up saliency map (right plane in Fig. 6). Psychophysical data were consistent with the prediction from the computational model.

B. fMRI Experiment

11 subjects participate in this experiment. The percentages of correct detection are 50.1 ± 0.7% and 50.3 ± 0.7% for the high salient and the low salient groups, respectively. The behavioral data confirms that the low-luminance natural images are indeed subjectively invisible to subjects. Contralateral and ipsilateral regions of interest (ROIs) in LGN, V1-V4 and IPS are defined as the cortical areas that respond to retinal inputs in the salient region and its contralateral counterpart. LOC and FEF in two hemispheres could be activated equally well by stimuli presented in the left and the right visual fields in the localizer scan. Thus, instead of presenting data...
Fig. 6. FMRI results in LGN, V1-V4 and IPS. (A). Event-related BOLD signals averaged across subjects in the contralateral and ipsilateral ROIs in LGN, V1-V4 and IPS. They are evoked by the bottom-up saliency map of natural images in the high salient and the low salient group. Error bars denote 1 SEM calculated across subjects at each time point. (B) Peak amplitude differences between the event-related BOLD signal in the contralateral ROI and that in the ipsilateral ROI in LGN, V1-V4 and IPS for the high salient and the low salient group. Error bars denote 1 SEM calculated across subjects.
in ipsilateral and contralateral regions in LOC and FEF, we
directly analyze event-related BOLD signals according to the
degree of saliency (the high salient and the low salient groups)
in these two cortical areas.

It is found that in V1-V4 and IPS, the natural images in both
two groups evoke larger BOLD signals in the contralateral
ROIs compared with the ipsilateral ROIs (Fig. 4A). It means
that the salient region could evoke stronger neural activity than
its contralateral counterpart. BOLD signal difference (namely
BSD) is quantified as the peak value difference of the BOLD
signals in the contralateral ROI and that in the ipsilateral ROI
(Fig. 4B). The BSDs of the high salient group and the low
salient group are compared by paired t-test. We find that the
BSD of the high salient group is significantly higher than
that of the low salient group ($t_{10} = 4.989$, $p = 0.001$) in
V1. However, in LGN, V2-V4 and IPS, we do NOT observe
significant BSD difference between the high salient group and
the low salient group (LGN: $t_{10} = −0.690$, $p = 0.506$; V2:
$t_{10} = 0.194$, $p = 0.850$; V3: $t_{10} = 0.194$, $p = 0.850$; V4:
$t_{10} = −0.159$, $p = 0.877$; IPS: $t_{10} = 0.540$, $p = 0.601$).
Moreover, we also measure the BOLD signal peak value
difference between the high salient and the low salient groups
in LOC and FEF (Fig. 7). We find NO significant difference
in these two areas (LOC: $t_{10} = −0.141$, $p = 0.891$; FEF:
$t_{10} = −0.690$, $p = 0.506$). As the attentional effect of
the bottom-up saliency map of the high salient group is also
significantly higher than that of the low salient group, these
findings reveal that neural activity in V1 is parallel to the
attentional effect.

C. Correlation Analysis

In order to further evaluate the role of neural activities of
early visual cortical areas in realizing bottom-up saliency
maps, we calculate the correlation coefficients between our
psychophysical and fMRI measures across individual sub-
jects. The attentional effect is significantly correlated with
the BOLD signal difference in V1 for the high salient group
($r = 0.633$, $p < 0.05$), but not for the low salient group
($r = 0.372$, $p = 0.260$) (Fig. 8A). However, no significant
correlation is found between the attentional effect and the
BOLD signal difference in the other cortical areas (Fig. 8B).
The results indicate a close relationship between the attentional
effect and V1 neural activities.

VI. CONCLUSION AND DISCUSSION

In summary, we develop a novel method to measure the
bottom-up saliency map of natural images. Our paradigm
is able to measure the bottom-up saliency at every location of
an image without affected by top-down signals, which
remedies several drawbacks of the traditional eye tracking
based methods. Based on the proposed method, a new database
of the bottom-up saliency maps of natural images is built.
which provides a more precise measurement to benchmark computational saliency models. Several state-of-art saliency models are compared on the proposed dataset. Moreover, a similar paradigm is applied to investigate the neural basis of the bottom-up saliency map of natural images. We measure the attentional effect and brain activities of the bottom-up saliency map on subjectively invisible natural images. We find that even if subjects are unaware of natural images, the attentional effect and the BOLD signal in V1 still increase with the degree of saliency. In addition, the attentional effect significantly correlates with the BOLD signals in V1, but not other cortical areas. These findings provide a strong scientific evidence to resolve the long standing dispute in neuroscience about where the bottom-up saliency map is constructed in human brain.

One important assumption of our study is that top-down signals are effectively eliminated in our paradigm. In the area of cognitive neuroscience, it has been proved and widely accepted that subjective awareness is determined by top-down signaling \[^9\]. Thus, rendering a stimulus invisible means eliminating top-down signals to the stimulus. No matter in psychophysical or fMRI experiments, we find that those natural images are invisible to participants, which confirms the assumption that top-signals are indeed eliminated in our paradigm. This issue is quiet important because several studies indicated that temporarily sluggish fMRI signals reflect neural activities resulting from both bottom-up and top-down processes, even in early visual cortex \[^9\]. Thus, blocking top-down signals make sure that we could observe a relative pure signal of the bottom-up saliency map of natural scenes in different brain regions.

Another important concern about our method is the resolution of the measured bottom-up saliency map. It can be laborious to measure the saliency value at each location/pixel of an image. However, we can propose a coarse-to-fine strategy to increase the sampling resolution non-uniformly. For example, we can re-sample more points around the salient regions based on the previous pass of sampling.

The most interesting observation in our experiment is that we find V1 plays a distinct role in realizing the bottom-up saliency map of natural scenes. Our findings challenge the dominant view that the bottom-up saliency map is created in higher brain regions such as IPS and FEF \[^22\], \[^15\], \[^3\]. Neural activities induced by the bottom-up saliency map are not observed in LOC, IPS, or FEF, which indicates that the observed neural activities in V1 are not attribute to the signals feedback from these areas. On the other hand, our results are consistent with previous findings that the neural response of V1 neurons is higher when their preferred stimuli pop out from the background \[^27\]. We suggest that the underlying neural mechanism of our observation may be attributed to the lateral connections \[^17\] between V1 neurons. More importantly, our observation is consistent with the V1 saliency theory \[^25\], \[^26\]), which states that V1 creates the bottom-up saliency map.

Our experiments extend previous evidences which support this theory from both psychophysical and physiological aspects. Instead of using textures consisted of simple oriented bars, we use natural scenes as our stimuli. Natural scenes contain richer naturalistic low-level features, including luminance, contrast, spatial frequency, curve etc. These features are basic units that our visual system needs to deal with. Therefore, this study is not only a critical complement to the previous study, but also provides an important evidence for the V1 saliency map argument.

**REFERENCES**

1. C. Asplund, J. Todd, A. Snyder, and R. Marois. A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nature Neuroscience*, 13(4):507–514, 2010.
2. T. Bezt, N. Wilming, C. Bogler, and J. Haynes. Dissociation between saliency signals and activity in early visual cortex. *Journal of Vision*, 13(14):1–12, 2013.
3. C. Bogler, S. Bode, and J. Haynes. Decoding successive computational stages of saliency processing. *Current Biology*, 21(19):1667–1671, 2011.
4. B. G. Breitmeyer and H. Ogmen. Recent models and findings in visual backward masking: A comparison, review, and update. *Perception and Psychophysics*, 62(8):1572–1595, 2000.
5. N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*, 18:155–162, 2006.
6. G. Buracas and G. Boynton. Efficient design of event-related fMRI experiments using m-sequences. *Neuroimage*, 16:801–813, 2002.
7. W. Chen, X. Zhu, K. Thulborn, and K. Ugurbil. Retinotopic mapping of lateral geniculate nucleus in humans using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 96(5):2430–2434, 1999.
8. M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Review Neuroscience*, 3(3):201–215, 2002.
9. A. D. Cul, S. Bueillet, and S. Dehaene. Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology*, 5(10):2408–2423, 2007.
10. L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. *IEEE Computer Vision and Pattern Recognition*, pages 473–480, 2011.
11. M. Eckstein, S. Schimoizaki, and C. Abbey. The footprints of visual attention in the posner cueing paradigm revealed by classification images. *Journal of Vision*, 2(1):25–45, 2002.
12. S. Engel, G. Glover, and B. Wandell. Retinotopic organization in human visual cortex and the spatial precision of functional mr. *Cerebral Cortex*, 7(2):181–192, 1997.
13. F. Fang and S. He. Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, 8(10):1380–1385, 2005.
14. J. Fecteau and D. Munoz. Saliency, relevance, and firing: a priority map for target selection. *Trends in Cognitive Science*, 10(9):25–45, 2006.
15. J. Geng and G. Mangun. Anterior intraparietal sulcus is sensitive to bottom-up attention driven by stimulus salience. *Journal of Cognitive Neuroscience*, 21(8):1584–1601, 2009.
16. C. Gilbert and W. Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14:350–363, 2013.
17. C. Gilbert and T. Wiesel. Clustered intrinsic connections in cat visual cortex. *Journal of Neuroscience*, 3(5):1116–1133, 1983.
18. S. Goerman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transaction in Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012.
19. S. Harrison and F. Tong. Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*, 50(22):2295–2307, 2010.
20. X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems*, 21:681–688, 2008.
21. L. Itti, C. Koch, and E. Niebur. A model of saliency based visual attention for rapid scene analysis. *IEEE Transaction in Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
22. C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.
23. A. Koene and L. Zhaoqing. Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in v1. *Journal of Vision*, 7(7):1–14, 2007.
24. Z. Kourtzi and N. Kanwisher. Cortical regions involved in perceiving object shape. *Journal of Neuroscience*, 20(9):3310–3318, 2000.
25. Z. Li. Contextual influences in v1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences of the United States of America*, 96(18):10530–10535, 1999.
26. Z. Li. A saliency map in primary visual cortex. *Trends in Cognitive Science*, 6(1):9–16, 2002.
[27] D. Marcus and D. V. Essen. Scene segmentation and attention in primate cortical areas v1 and v2. *Journal of Neurophysiology*, 88(5):2648–2658, 2002.

[28] J. Mazer and J. Gallant. Goal-related activity in v4 during free viewing visual search: Evidence for a ventral stream visual salience map. *Neuron*, 40(6):1241–1250, 2003.

[29] T. Paus. Location and function of the human frontal eye field: a selective review. *Neuropsychologia*, 34(6):475–483, 1996.

[30] M. Posner, C. Snyder, and B. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology*, 109(2):160–174, 1980.

[31] U. Rutishause, D. Wulff, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? *IEEE Computer Vision and Pattern Recognition*, 2:37–44, 2004.

[32] M. Sereno, A. Dale, J. Reppas, K. Kwong, J. Belliveau, T. Brady, B. Rosen, and R. Tootell. Borders of multiple visual areas in human revealed by functional magnetic resonance imaging. *Science*, 268(5212):889–893, 1995.

[33] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. *IEEE Computer Vision and Pattern Recognition*, 2:994–1000, 2005.

[34] S. Shipp. The brain circuitry of attention. *Trends in Cognitive Science*, 8(5):223–230, 2004.

[35] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao. Simulating human saccadic scanpaths on natural images. *IEEE Computer Vision and Pattern Recognition*, pages 441–448, 2011.

[36] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. *IEEE Computer Vision and Pattern Recognition*, pages 2368–2375, 2010.

[37] X. Zhang, L. Zhao ping, T. Zhou, and F. Fang. Neural activities in v1 create a bottom-up saliency map. *Neuron*, 73(1):183–192, 2012.