# Global visual salience of competing stimuli

**Alex Hernández-García**
Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany
Max Planck School of Cognition, Osnabrück, Germany

**Ricardo Ramos Gameiro**
Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

**Alessandro Grillini**
Department of Ophthalmology, University Medical Center Groningen, Groningen, Netherlands

**Peter König**
Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany
Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

**Current computational models of visual salience accurately predict the distribution of fixations on isolated visual stimuli. It is not known, however, whether the global salience of a stimulus, that is, its effectiveness in the competition for attention with other stimuli, is a function of the local salience or an independent measure. Further, do task and familiarity with the competing images influence eye movements? Here, we investigated the direction of the first saccade to characterize and analyze the global visual salience of competing stimuli. Participants freely observed pairs of images while eye movements were recorded. The pairs balanced the combinations of new and already seen images, as well as task and task-free trials. Then, we trained a logistic regression model that accurately predicted the location—left or right image—of the first fixation for each stimulus pair, accounting too for the influence of task, familiarity, and lateral bias. The coefficients of the model provided a reliable measure of global salience, which we contrasted with two distinct local salience models, GBVS and Deep Gaze. The lack of correlation of the behavioral data with the former and the small correlation with the latter indicate that global salience cannot be explained by the feature-driven local salience of images. Further, the influence of task and familiarity was rather small, and we reproduced the previously reported left-sided bias. Summarized, we showed that natural stimuli have an intrinsic global salience related to the human initial gaze direction, independent of the local salience and little influenced by task and familiarity.**

## Introduction

The guidance of eye movements in visual behavior is a dominant necessity for navigation and interaction with the environment (Liversedge & Findlay, 2000; Geisler & Cormack, 2011; König et al., 2016), also reflecting the individual personality (Rauthmann et al., 2012). We constantly have to decide where to look next and which regions of interest to explore, in order to process and interpret relevant information of a scene (Ramos Gameiro et al., 2017). As a consequence, investigating eye movement behavior has become a major field in many research areas (Kowler, 2011; Kaspar, 2013; König et al., 2016).

In this regard, a number of studies have shown that visual behavior is controlled by three major mechanisms: bottom-up, top-down, and spatial biases (Desimone & Duncan, 1995; Egeth & Yantis, 1997; Kastner & Ungerleider, 2000; Corbetta & Shulman, 2002; Connor et al., 2004; Tatler & Vincent, 2009; Kollmorgen et al., 2010; Ossandón et al., 2014). Bottom-up factors describe features of the observed image, which attract eye fixations, involving primary contrasts, such as color, luminance, brightness, and saturation (Itti et al., 1998; Reinagel & Zador, 1999; Baddeley & Tatler, 2006). Hence, bottom-up factors are typically based on the sensory input. In contrast, top-down factors comprise internal states of the observer (Connor et al., 2004; Kaspar, 2013). That is, eye movement behavior is also guided by specific characteristics, such as personal motivation, specific

Citation: Hernández-García, A., Ramos Gameiro, R., Grillini, A., & König, P. (2020). Global visual salience of competing stimuli. *Journal of Vision*, *20*(7):27, 1–18, https://doi.org/10.1167/jov.20.7.27.

https://doi.org/10.1167/jov.20.7.27 Received July 17, 2019; published July 28, 2020 ISSN 1534-7362 

search tasks, and emotions (Wadlinger & Isaacowitz, 2006; Einhäuser et al., 2008; Henderson et al., 2009; Rauthmann et al., 2012; Kaspar & König, 2012). Finally, spatial properties of the image, such as the image size, and motor constraints of the visual system in the brain may affect eye movement behavior (Ramos Gameiro et al., 2017, 2018). As a result, spatial properties and motor constraints then lead to specific bias effects, such as the central bias in natural static images (Tatler, 2007). Thus, investigating visual behavior necessarily implies an examination of bottom-up and top-down factors as well as spatial biases.

Based on these three mechanisms—bottom-up, top-down, and spatial biases—guiding visual behavior, Koch and Ullman (1987) first revealed a method to highlight salient points in static image scenes. Whereas this model was purely conceptual, Niebur and Koch (1996) later developed an actual implementation of salience maps. This was the first prominent proposal of topographically organized features maps that guide visual attention. Salience maps describe these topographic representations of an image scene, revealing where people will most likely look at while observing the respective scene (Itti et al., 1998; Itti & Koch, 2001). That is, salience maps can be interpreted as a prediction of the distribution of eye movements on images. Usually, salience maps include only bottom-up image features, predicting eye fixations on image regions with primary contrasts in color changes, saturation, luminance, or brightness, among others (Itti et al., 1998; Itti and Koch, 2001). However, in their first implementation, Niebur and Koch (1996) also tried to include top-down factors to build up salience maps and thus predict where people will look at most likely in image scenes. Current state-of-the-art computational salience models are artificial neural networks pretrained on large data sets for visual object recognition and subsequently tuned to predict fixations, as is the case of Deep Gaze II (Kümmerer et al., 2016). Such models do not rely only on bottom-up features any more but also incorporate higher-level features learned on object recognition tasks. Still, despite the better performance on salience benchmarks, deep nets-based models seem to fail at predicting the salience driven by low-level features (Kümmerer et al., 2017).

Salience maps provide a highly accurate and robust method to predict human eye movement behavior on static images by relying on local features to determine which parts of an image are most salient (Niebur & Koch, 1996; Itti et al., 1998; Itti & Koch, 2001; Kowler, 2011). However, these methods do not provide any information about the salience of the image as whole, which may depend on both local properties and also the overall semantic and contextual information of the image. Such global salience is of great relevance when an observer is faced with two or more independent visual stimuli in one context. These combinations describe situations when several stimuli compete with each other with regard to their individual semantic content, despite being in the same overall context. Such cases appear frequently in real life, for instance, when two billboards hang next to each other in a mall or when several windows are open on a computer screen or a monitor in an intensive care unit, to name a few examples. Thus, by placing two or more independent image contexts side by side, as described in the previous examples, classical salience maps may well predict eye movement behavior within each of the individual images as a closed system, but they will most likely fail to predict visual behavior across the whole scene involving all images. Specifically, they will fail at answering the question: Which stimulus is most likely to attract the observers' visual attention?

In this study, our primary hypothesis is (H1) that it is possible to measure and calculate the global salience of natural images. That is, the likelihood of a visual stimulus to attract the first fixation of a human observer, when it is presented in competition alongside another stimulus, can be systematically modeled. In the experiment presented here, participants were confronted with stimuli containing two individual natural images—one on the left and one on the right side of the screen—at the same time. The set of images used to build our stimuli consisted of urban, indoor and nature scenes; closeups of human faces; and scenes with people in a social context. During the observation of the image pairs, we recorded the participants' eye movements. Specifically, to characterize the global salience, we were interested in the direction—left or right—of the initial saccade the participant made after the stimulus onset. For further analysis, we also collected all binary saccade decisions on all the image pairs presented to the participants. We used the behavioral data collected from the participants to train a logistic regression model that successfully predicts the location of the first fixation for a given pair of images. This allowed us to use the coefficients of the model to characterize the likelihood of each image to attract the first fixation, relative to the other images in the set. In general, images that were fixated more often were ranked higher than other images. Hence, we computed a unique "attraction score" for each image that we denote "global salience," which depends on the individual contextual information of the image as a whole.

We also analyzed the local salience properties of the individual images and compared them to the global salience. We hereby claimed that the global salience cannot be explained by the feature-driven salience maps. Formally, we hypothesize that (H2): Natural images have a specific global salience, independent of their local salience properties, that characterizes their likelihood to attract the first fixation of human observers, when presented alongside another competing stimulus. A larger global salience leads to a higher attraction of initial eye movements.

In order to properly calculate the global salience, we accounted for general effects of visual behavior in stimuli with two paired images. Previous studies have shown that humans tend to exhibit a left bias in scanning visual stimuli. Barton et al. (2006) showed that subjects looking at faces make longer fixations on the eye on their left side, even if the faces were inverted, and the effect was later confirmed and extended to dogs and monkeys (Barton et al., 2006; Guo et al., 2009). For an extensive review about spatial biases, see the work by Ossandón et al. (2014), where the authors presented evidence of a marked initial left bias in right-handers but not in left-handers, regardless of their habitual reading direction. In sum, there is a large body of evidence of lateral asymmetry in viewing behavior, although the specific sources are yet to be fully confirmed. With respect to our study, we hypothesize that (H3): Presenting images in horizontal pairs leads to a general spatial bias in favor of the image on the left side.

In addition to the general left bias, in half of the trials of the experimental sessions, one of the images had been already seen by the participant in a previous trial, while the other was new. The participants also had to indicate which of the images was new or old. Thus, we also addressed the questions of whether the familiarity with one of the images or the task has any effect in the visual behavior and thus in the global salience of the images. Do images that show the task-relevant scene attract more initial saccades? Likewise, are novel images more likely to attract the first fixation? This challenge sheds some light on central-peripheral interaction in visual processing. Guo (2007), for instance, showed that during face processing, humans indeed rely on top-down information in scanning images. However, Açk et al. (2010) proposed that young adults usually rely on bottom-up rather than top-down information during visual search. In this regard we thus hypothesize that (H4): Task relevance and familiarity of images will not lead to higher probability of being fixated first. In order to account for any spatial bias effects that could influence the global salience model, we added coefficients to the logistic regression algorithm that could potentially capture any lateral, familiarity, and bias effects. This not only makes the model more accurate but allows us to analyze the influence of these effects. Furthermore, the location of the images in the experiments was randomized across trials and participants.

Finally, in order to better understand the properties of the global salience of competing stimuli, we also analyzed the exploration time of each image. In this regard, we hypothesize the following (H5): Images with larger global salience will be explored longer than images with low global salience.

Code and data of this work are available at https://github.com/alexhernandezgarcia/global-salience.

# Methods: experimental setup

The present study was conducted in the neurobiopsychology lab at the Institute of Cognitive Science of the University of Osnabrück, Germany. The experimental methods were approved by the Ethical Committee of the University of Osnabrück, Germany, and performed in accordance with the guidelines of the German Psychological Society. All participants gave written consent to participate in this study.

## Participants

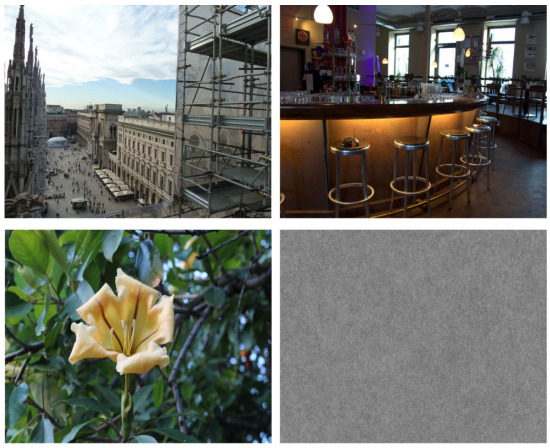
Forty-nine healthy participants (33 females, mean age = 22.39 years, $SD = 3.63$) with normal or corrected-to-normal vision took part in this study. All participants were instructed to freely observe the stimuli on the screen. In part of the measurements, they had to indicate after the trial the old or new image of a pair as further described below.

## Apparatuses

We presented the stimuli on a 32-in. widescreen Samsung monitor (Apple, Cupertino, CA) with a native resolution of 3,840 × 2,160 pixels. For eye movement recordings, we used a stationary Eye Link 1000 eye tracker (SR Research Ltd., Ottawa, Ontario, Canada) providing binocular recordings with one head camera and two eye cameras with a sampling rate of 500 Hz.

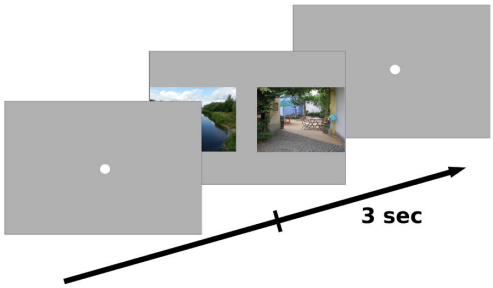
Participants were seated in a darkened room at a distance of 80 cm from the monitor, resulting in 80.4 pixels per visual degree in the center of the monitor. We did not fixate the participant's head with a headrest but verbally instructed the participants not to make head movements during the experiment. This facilitated comfortable conditions for the participants. However, the eye tracker constantly recorded four edge markers on the screen with the head camera, in order to correct for small head movements. This guaranteed stable gaze recordings based on eye movements, independent of residual involuntary head movements.

The eye tracker measured binocular eye movements. For calibration of the eye-tracking camera, each participant had to fixate on 13 black circles (size 0.5°) that appeared consecutively at different screen locations. The calibration was validated afterward by calculating the drift error for each point. The calibration was repeated until the system reached an average accuracy of $<0.5°$ for both eyes of the participant.



(a) Four images from the set

(b) The stimulus at one trial




(c) Procedure in blocks 1 and 4, where no task was given


(d) Procedure in blocks 2 and 3, where participants had to perform a simple task after the stimulus offset

Figure 1. Experimental setup.

## Stimuli

The images set consisted of 200 images, of which 197 were natural photographs and 3 were randomly generated pink noise images. Altogether, the stimulus set was divided into six categories, according to the image content: human faces, urban scenes, natural landscapes, indoor scenes, social activities, and pink noise. All photographs were obtained from either the internal image database of the neurobiopsychology laboratory at the University of Osnabrück, Germany, or the NimStim database. Each image was scaled to a resolution of 1,800 × 1,440 pixels. Some examples are shown in Figure 1a.

Each trial consisted of one stimulus with a resolution of 3,840 × 2,160 pixels, matching the full-size screen resolution of the display monitor (32-in. diagonal; 47.8° × 26.9°). Within each presented stimulus, two images were randomly paired, that is, one image was shown on the left screen side and the other image on the right screen side. Between both images, each stimulus contained a central gap of 240 pixels, as illustrated by Figure 1b. The background area of the stimuli was set to middle gray.

## Procedure

The experiment consisted of 200 trials divided into four blocks, at the beginning of which the eye-tracking system was recalibrated. The blocks were designed such that each had a different combination of task and image novelty:

- **Block 1** consisted of 25 trials formed by 50 distinct, novel images (new/new). This block was task-free, that is, participants were guided to freely observe the stimuli (Figure 1c).
- **Block 2** consisted of 75 trials, each formed by one new image and one of the previously seen images (new/old or old/new). In this block, the participants were guided to freely observe the stimuli and, additionally, they were asked to indicate the *new* image of the pair after the stimulus offset (Figure 1d).
- **Block 3** consisted of 75 trials, each formed by one new image and one of the previously seen images (new/old or old/new). In this block, the participants were asked to indicate the *old* image of the pair.

- **Block 4** consisted of 25 trials formed by 50 previously seen images (old/old). Like Block 1, this block was also task-free.

The decision in Blocks 2 and 3 was indicated by either pressing the left (task-relevant image is on the left side) or right (task relevant-image is on the right side) arrow button on a computer keyboard.

The image pairs were formed by randomly sampling from the set of 200 images, but some constraints were set in order to satisfy the characteristics of each block and keep a balance in the number of times each image was seen by the participant. The sampling process was as follows: In Block 1, 50 images were randomly sampled to form the 25 pairs. In Blocks 2 and 3, in order to construct the new/old and old/new pairs, the new image was randomly sampled from the set of remaining unseen images and the old image was randomly sampled of previously seen images, with two additional constraints: It must have been chosen only one time before and not in the previous five trials. Finally, in Block 4, a set of exactly 50 images that had been shown only once remained. These were used to randomly sample the remaining 25 trials. In all blocks, after sampling the two images, the left/right configuration was also randomly chosen with probability 0.5.

The sampling process was different for each participant, that is, they saw different sets of pairs from the 40,000 different pairs and in different order. This aimed at reducing the predictability of the process while satisfying the experimental constraints. Overall, we collected data from 9,800 pairs, some of which might have been repeated across participants. However, note that each participant saw each image exactly twice; therefore, the frequency of presentation of the images was balanced across the whole experiment. As we will see in the following section, the amount of data was enough to fit the computational model.

In all cases, the presentation time for each stimulus was 3 s and it was always preceded by a blank, gray screen with a white, central fixation dot. The stimulus was displayed only after the participant fixated the central dot.

The majority of our analyses focused on the first fixation. As a preprocessing stage, we discarded the fixations (a) due to anticipatory saccades; (b) shorter than 50 ms or longer than $\mu_{dur} + 2\sigma_{dur}$ ms, where $\mu_{dur} = 198$ ms and $\sigma_{dur} = 90$ ms are the mean and standard deviation of all fixation durations, respectively; and (c) located outside any of the two images. The discarded fixations were less than 4% of the total.

# Methods: computation of global salience

In order to characterize the global salience of competing stimuli, we trained a logistic regression model with the behavioral data from the eye-tracking experiments. Provided that the model can accurately predict the location of the first fixation—left or right—the coefficients for each image will represent the likelihood of the image to attract the first fixation, and this, in turn, can then be interpreted as the global image salience. The intuition is that images that are more often the target of the first fixation after the stimulus onset have a higher global salience and vice versa.

## Logistic regression for pairwise estimates

Typically, logistic regression is used in binary classification problems, as is this case where the initial fixation after stimulus onset can land either on the left ($y = -1$) or on the right ($y = 1$) image. The classifier simply estimates a probability $h_w(\mathbf{x})$ for the binary event on the linear hypothesis $\mathbf{w}^T\mathbf{x}$ by applying a logistic function:

$$h_w(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}} = \frac{e^{\mathbf{w}^T\mathbf{x}}}{1 + e^{\mathbf{w}^T\mathbf{x}}} \quad (1)$$

where $\mathbf{x}$ is a vector that represents the independent or explanatory variables (features) and $\mathbf{w}$ the coefficients to be learned. Thus, the likelihood of the binary outcome given the data is the following:

$$P(y|\mathbf{x}) = \begin{cases} h_w(\mathbf{x}) & \text{if } y = 1 \\ 1 - h_w(\mathbf{x}) & \text{if } y = -1 \end{cases} = \frac{e^{y\mathbf{w}^T\mathbf{x}}}{1 + e^{y\mathbf{w}^T\mathbf{x}}}$$

The coefficients are then optimized by minimizing the negative log-likelihood $-log(P(y|\mathbf{x}))$ through gradient descent. Typically, a regularization penalty is added on the coefficients, controlled by the parameter $C$ (inverse of the regularization strength). In our case, we applied $L_2$ regularization, and therefore the algorithm solves the following optimization problem, given a set of $N$ training data points (trials):

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} log(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}) \quad (2)$$

The optimization problem was solved through the LIBLINEAR algorithm (Fan et al., 2008), available in the scikit-learn Python toolbox.

In our particular case, for every trial $i$ (stimulus pair seen by a participant), each feature $x_{ij}$ corresponded

to one image $j$, and only two images were shown at each trial. Therefore, we were interested in modeling the probability that one image $u$ receives the first fixation when presented next to another image $v$; hence, $p(u > v)$. This simplifies the standard logistic regression model to a special case for pairwise probability estimates, known as the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 2005), where the probability $h_w$ is the following:

$$h_w(u, v) = p(u > v) = \frac{e^{w_u}}{e^{w_u} + e^{w_v}} = \frac{e^{w_u - w_v}}{1 + e^{w_u - w_v}} \quad (3)$$

where $w_u$ and $w_v$ are the coefficients of image $u$ and $v$. This is a special case of the function in Equation 1, where all the elements in the feature vector $\mathbf{x}$ are zero except for the two paired features $x_u$ and $x_v$, which are set to 1 and $-1$ respectively. Note that in the BTL model, the coefficients still refer to the whole set of features and therefore are described by an $M$-dimensional vector $\mathbf{w} = \{w_1, w_2, ..., w_M\}$, where in our case, $M = 200$, the total number of images in the set. After training the model, each learned coefficient $w_j$ will be related to the average likelihood of image $j$ of receiving the first fixation when presented next to other images from the set. As stated above, we interpret these coefficients $\mathbf{w}$ as a measure of the global image salience.

In order to estimate the coefficients $\mathbf{w}$, the logistic regression model was trained on the data set arranged into a design matrix $X$ of the following form:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_M^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_M^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_M^{(N)} \end{bmatrix} \quad (4)$$

where each row represents one measured data point, that is, one trial where one participant was presented a pair of images $u$ and $v$ (the total number of trials was in our case $N = 49$ participants $\times$ 200 trials per participant $= 9800$) and where the columns represent the values of the different features (images) that were tested ($M = 200$). According to Equation 3, if image $u$ is presented on the right and image $v$ is presented on the left at trial $i$, then $x_u^{(i)} = 1$, $x_v^{(i)} = -1$ and $x_j^{(i)} = 0, \forall\ j \neq u, v$. Finally, the outcome of each trial is given as a vector $\mathbf{y}$:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

such that $y^{(i)} = 1$ if the right image was fixated first, and $y^{(i)} = -1$ if the left image was fixated first at trial $i$.

## Task, familiarity, and lateral bias

Not only were we interested in modeling the likelihood of every image of receiving the first fixation, but also the contribution of other aspects of the experiment, namely, the effect of having to perform a small task when observing the pair of images and the familiarity with one of the two images. More specifically, we were interested in answering the following questions: Do light task demands, such as having to determine which image is new or old, influence the direction of the first saccade? Also, are unseen stimuli more likely to receive the initial saccade than previously observed stimuli when presented together or vice versa?

We addressed these questions by adding new features to the model that capture these characteristics of the experimental setup. These features were assigned coefficients that, after training, will indicate the magnitude of the contributions of the effects. In particular, we added the following features columns to every row $i$ of the design matrix:

- $t^{(i)}$: 1 if the target of the task (select new/old image) was on the right at trial $i$, $-1$ image if on the left, 0 if no task.
- $f^{(i)}$: 1 if at trial $i$, the image on the right had been already shown at a previous trial (familiar), while the image on the left was still unseen; $-1$ if the familiar image was on the left; 0 if both images were new or familiar.

Not only did these new features enable new elements for the analysis, but they also added more representational power to the model, which could potentially learn better coefficients to describe the global salience of each image. In this line, we added one more feature to the model to capture one important aspect of visual exploration: the lateral bias. Although a single intercept term in the argument of the logistic function ($\mathbf{w}^T\mathbf{x} + b$) would capture most of the lateral bias, since the outcome $\mathbf{y}$ describes exactly the lateral direction, left or right, of the first saccade, we instead added subject-specific features to model the fact that the trials were generated by different subjects with an individual lateral bias. This was done by adding $K = 49$ (number of participants) features $s_k^{(i)}$, with value 1 if the trial $i$ was performed by subject $k$ and 0 otherwise. Altogether, the final design matrix $X'$ extends the design matrix $X$ defined in Equation 4 as follows:

| | AUC | Tjur $R^2$ | Accuracy |
|---|---|---|---|
| Test | 0.8884 (0.0180) | 0.4287 (0.0460) | 81.36% (0.32) |
| Train | 0.8865 (0.0040) | 0.4240 (0.0214) | 81.99% (1.52) |
| Random baseline | 0.5 | 0.0 | 60.70% (2.32) |

Table 1. Test, train, and baseline performance of the logistic regression model. Values within brackets indicate the standard deviation across the folds. AUC = area under the curve.

$$X' = \left[\begin{array}{ccc|c|c|ccc} x_1^{(1)} & \dots & x_M^{(1)} & t^{(1)} & f^{(1)} & s_1^{(1)} & \dots & s_K^{(1)} \\ x_1^{(2)} & \dots & x_M^{(2)} & t^{(2)} & f^{(2)} & s_1^{(2)} & \dots & s_K^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & \dots & x_M^{(N)} & t^{(N)} & f^{(N)} & s_1^{(N)} & \dots & s_K^{(N)} \end{array}\right] \quad (5)$$

Note that the leftmost block of $X'$ is identical to $X$ (defined in Equation 4). While the shape of $X$ is 9,800 × 200, $X'$ is a 9,800 × 251 matrix, since 200 + 1 + 1 + 49 = 251.

### Validation and evaluation of the model

In order to ensure the successful training of the model, we carried out a fivefold cross-validation of the regularization parameter $C$ of the model, described in Equation 2. That is, we split our data set into five different folds of 39 subjects for training and 10 for validation (7,800 and 2,000 trials, respectively) and evaluated the performance with 10 different values of $C$, according to the following search space:

$$C = 10^p \quad \text{with} \quad p = -3 + \frac{2}{3}(n-1) \quad \text{and} \quad n = 1, ..., 10$$

The value that provided the best average performance across the folds was selected.

In order to reliably assess the model performance while taking the most out of the data set, we embedded the cross-validated model into a *leave-two-participants-out* cross-evaluation. That is, we constructed 25 different folds of data, each with the trials of 23 participants for training and of 2 participants for evaluation. We report here the average performance across the 25 test and train partitions together with the standard deviation (within brackets). In particular, in Table 1, we include the area under the curve (AUC), the Tjur[1] coefficient of discrimination $R^2$, and the accuracy. For the sake of an easier interpretation, we include the theoretical baseline values of the AUC and $R^2$, as well as the empirical baseline accuracy on our test partitions.

The results in Table 1 show that the logistic regression model successfully learned the behavioral patterns from the experimental data and hence accurately predicted the direction of the first saccade, with very low overfitting, since train and test performance were very similar and have low variance. As a conclusion, this implies that the learned coefficients can be meaningfully used for further analysis, as will be presented in Results section.

## Methods: salience maps of competing stimuli

In order to test whether the global salience is independent from the lower-level, salience properties of the stimuli (H2), we also computed salience maps both of each individual image and of each full stimulus shown at each trial, that is, the pair of images with gray background, as shown in Figure 1b. For the computation of the salience maps we used the Graph-Based Visual Salience algorithm (GBVS), by Harel et al. (2007), which is a computational salience model that makes use of well-defined low-level features.

Moreover, we also analyzed the connection between global salience and a less restricted salience model, Deep Gaze II (Kümmerer et al., 2016), whose features include higher-level cues, since it is a deep neural network model pretrained for large-scale, image object recognition tasks, with additional features optimized for salience prediction.

In order to compare the salience maps with the behavioral data from the observation of competing stimuli, as well as with our derived global salience, we performed the following tests:

### Predictivity of salience maps for the first fixation

In this case, our aim was to evaluate the performance of salience maps in predicting the landing location of the first fixation when two competing stimuli are presented. To do so, we computed the Kullback-Leibler divergence between the first fixation distribution $F_j(b)$ and the salience distribution $S_j(b)$ for every image $j$ in the set of 200 images:

$$D_{KL}(F_j||S_j) = \sum_{b=1}^{B} F_j(b) \log\left(\frac{F_j(b)}{S_j(b)+\epsilon} + \epsilon\right) \quad (6)$$

where $\epsilon$ is a small constant to ensure numerical stability and $b$ refers to $B$ bins of one $1 \times 1$ degrees of visual field angle.

The first fixation distribution, $F_j(b)$, is the probability density distribution of all the first fixations made by all observers on each image $j$. To compute $F_j(b)$, we divided every image into sections of one squared degree of visual field angle and counted the number of first fixations made by all participants on each bin to obtain a histogram. Then, the histogram was smoothed using a Gaussian kernel with a size of 1 degree of visual field angle and normalized such that it became a probability distribution. The salience distribution, $S_j(b)$, is the smoothed and normalized (likewise) salience map (computed with GBVS or Deep Gaze II) of each individual image $j$.

Hence, according to the definition in Equation 6, a low $D_{KL}(F_j \| S_j)$ would imply a good match between the location of the first fixations and the salience map of image $j$.

## Comparison between global and local salience

In order to compare the local salience maps and the global salience scores learned by the computational model presented in Methods: computation of global salience section, we analyzed the GBVS and Deep Gaze salience maps of both the individual images and the whole stimuli, in relation to the global salience scores.

### *Individual images*

First, we compared the Kullback-Leibler divergence between the first fixations distribution and the salience maps of the individual images, as computed in Equation 6, and the global salience scores, that is, the coefficients learned by the optimization defined in Equation 2. This aimed at analyzing whether, for instance, images whose local salience properties indeed drove the location of the first fixation have a higher global salience score and vice versa.

### *Trials*

Second, we looked at the properties of the salience map of the final stimulus seen by the participants at each trial, that is, the paired competing images with a gray background (see Figure 1b). As a metric of the contribution of each image to the salience map, for each trial $i$, we computed the relative salience mass $M$ of each image, left and right:

$$M_i^L = \int_{x \in X_L} S_i(x) \quad M_i^R = \int_{x \in X_R} S_i(x)$$

where $S_i(x)$ is the normalized salience map of the whole stimulus presented at trial $i$ and $X_L$ and $X_R$ are the stimulus locations corresponding to the left and right images, respectively. A significant positive correlation between $\Delta_M^{(i)} = M_i^L - M_i^R$ and the difference between the global salience scores of the images on the left and right, $\Delta_{GS}^{(i)} = w_L^{(i)} - w_R^{(i)}$, would indicate that the local salience properties can partly explain the direction of the first fixation.

# Results

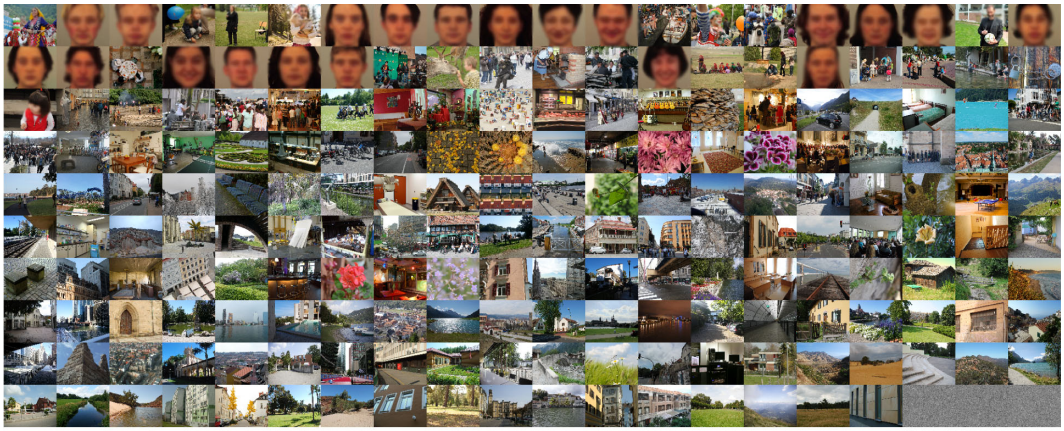
In this section, we present the main results of our analyses and discuss the validity of the hypotheses presented in the Introduction. Each of the subsections focuses on one of the five hypotheses, in the natural order. All the scatterplots that show the relationship between two variables include the value of the Pearson correlation, as well as the line fit by a linear regression model, with 95% confidence intervals estimated using bootstrap with 1,000 resamples.

## Global visual salience

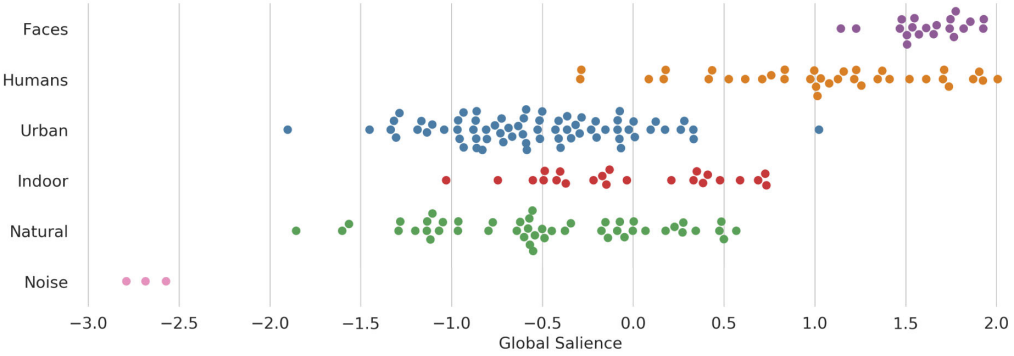
In our first hypothesis (H1), we stated that images can be ranked according to a specific global salience that leads to the attraction of initial eye fixations. In order to quantify the global salience of individual images, we have presented in Methods: computation of global salience section a computational model that successfully predicts the direction of the first fixation from the behavioral data, as validated by the results in Validation and evaluation of the model section, and thus we can analyze the coefficients of the model as indicators of the global salience of each image in the data set.

Importantly, the fact that the first fixation direction of the participants when exploring such competitive stimuli can be predicted by a computational model means that their behavior was not random but followed certain patterns. In order to shed some light on the nature of these patterns, in Figure 2a, we show the complete set of stimuli ranked according to the global salience score learned by our model and in Figure 2b the value of the global salience scores of each image, highlighting the differences between the image categories.

Figure 2 shows that there exists a clear, general tendency to first fixate on the images that contain either closeup faces or scenes with humans, even though the first fixations may occur, on average, as early as after 242 ms ($\sigma_{SD} = 66$ ms) from the stimulus onset. These two categories, faces and humans, were assigned the



(a) Experimental stimuli, ranked according to the learned global salience. The stimulus with the highest global salience score is on the top-left corner and the rest are sorted with the x-axis changing fastest (row-major order). Faces have been blurred to preserve the identity.




(b) Global salience score of each stimulus and image categories.

Figure 2. Global salience scores of the experimental stimuli.

highest global salience scores. Then, urban, indoor, and natural landscapes obtained significantly lower scores, with no big differences among the three categories. Finally, the three pink-noise images were assigned very low scores, which serves as a sanity check of our proposed method.

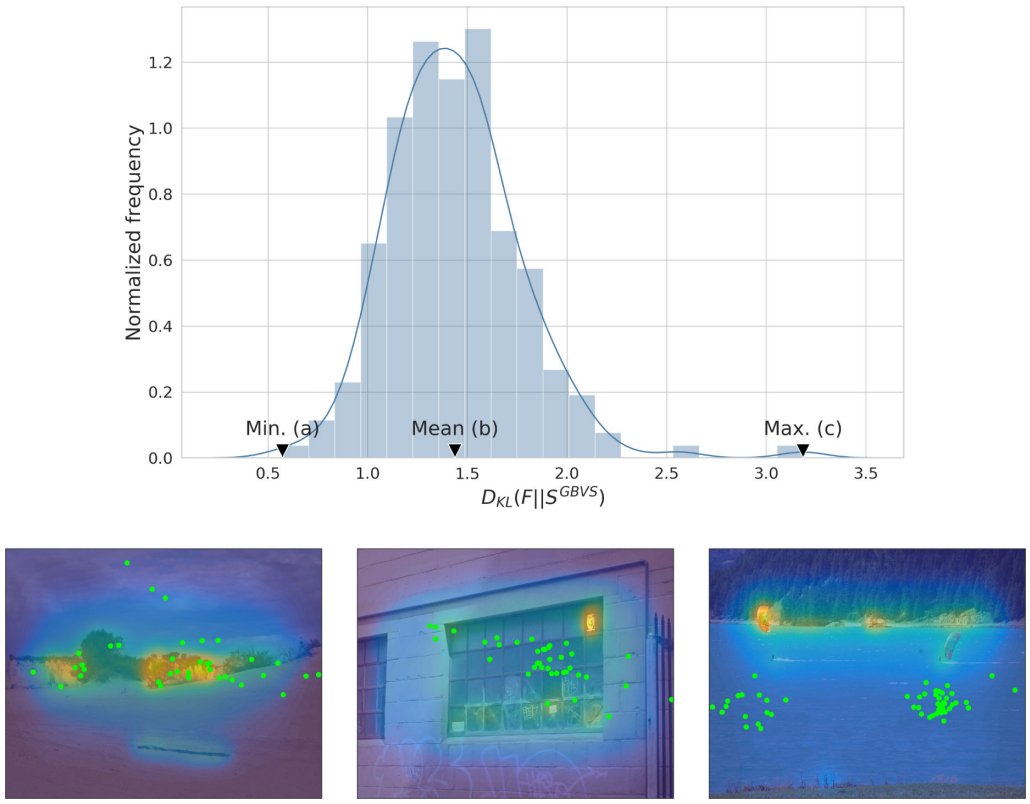
## Global versus local salience

A reasonable question in view of the results presented in Figure 2 is whether the global salience scores—and the ranking of the stimuli that arises from the scores—is a unique measure that assesses the initial visual behavior when facing competing stimuli or whether this behavior and thus our proposed global salience can be explained by the low-level properties of the respective stimuli.

In our second hypothesis (H2), we stated, instead, that the global salience is independent from the low-level local salience properties. So as to test this, we performed several tests, described in Methods: salience maps of competing stimuli section.

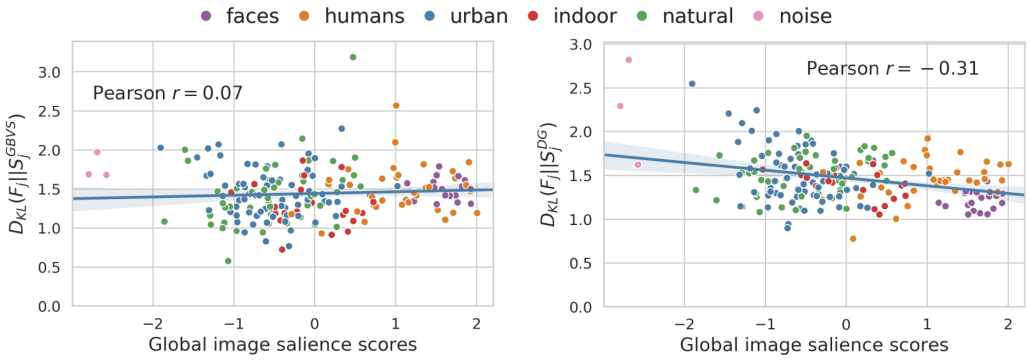
In Figure 3, we plot the distribution of the Kullback-Leibler divergence (KLD) between the first fixations maps and the GBVS local salience maps of the individual images (see Equation 6). The mean of the distribution is significantly nonzero (two-tail $t$ test $p < .001$, $\mu_{KLD} = 1.44$, $\sigma_{KLD} = 0.33$), which means that there is a significant loss of information when using a local salience map to predict the landing locations of the first fixations on a given image (Riche et al., 2013). In order to illustrate the mismatch, in Figure 3, we display three example images with the overlaid salience maps and the location of all the first fixations that landed on them. When the KLD value is minimum (a), the salience maps can approximate the fixations, although this happened rarely. Already with KLD values around the mean, the performance of a salience map in predicting the landing location of fixations is rather mediocre (b) and deteriorates further as the KLD increases (c).

Perhaps not surprisingly, in view of the poor match between the salience maps and the first fixation maps, Figure 4a shows that the Kullback-Leibler divergence between them does not correlate with the global salience scores. This means that the images that attract the first fixations toward salient regions (low KLD) do not tend to have high global salience scores or vice versa.




(a) $D_{KL}(F||S) = 0.57$ (b) $D_{KL}(F||S) = 1.44$ (c) $D_{KL}(F||S) = 3.19$

Figure 3. Top row: distribution of the Kullback-Leibler divergence between the first fixations map and the GBVS local salience maps. Bottom row: images with the minimum, closest to the mean and maximum KLD, with their overlaid salience map and the location of the first fixations.




(a) GBVS (b) Deep Gaze

Figure 4. Comparison between the global salience scores and the KLD between the first fixation distribution and the salience maps from the computational models.

Finally, we analyze in Figure 5 whether the direction of the first fixation when looking at competing stimuli, as modeled by our proposed global salience scores, can be explained by the difference in the low-level salience properties of the competing stimuli, as measured by the GBVS salience mass of each image (see Comparison between global and local salience section). Also in this case, we found no significant correlation.

The noisy images included in the stimulus set serve once more as a validation of the expected results. When one of the images (left or right) was pink noise, the difference in GBVS salience mass was either very high or very low, as is the difference in global salience scores. In this case, both metrics do correlate, but as shown by the central scatterplot of Figure 5, the feature-driven (GBVS) salience mass

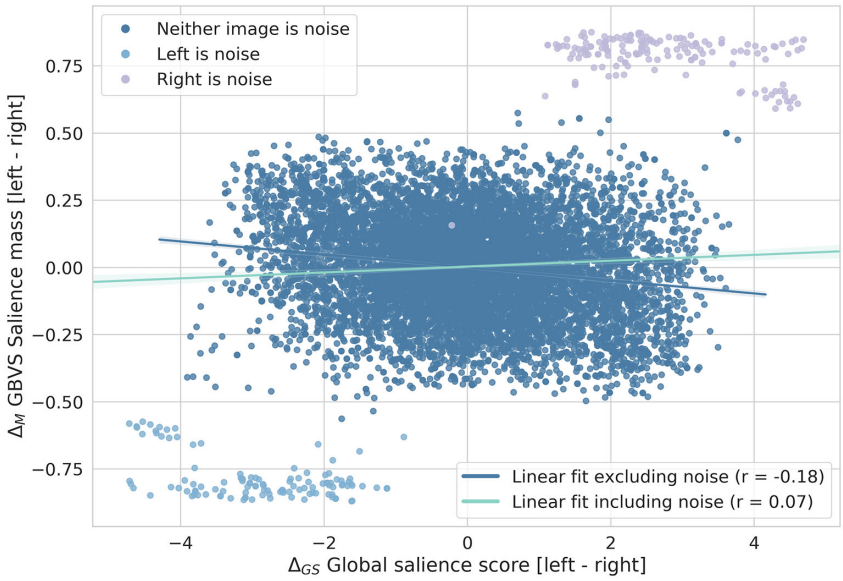



Figure 5. Correlation between the GBVS image salience mass and the global salience scores.

cannot explain the global salience scores learned by the model.

In order to better understand what drives the direction of the first fixation when faced with competing stimuli, we also compared our proposed global salience with properties of Deep Gaze II salience maps. As presented in Methods: salience maps of competing stimuli section, unlike GBVS, Deep Gaze does make use of higher-level information of the images to predict the salience maps, since it is a neural network pretrained on image object recognition tasks. This allows it to model salience driven by faces or objects (Kümmerer et al., 2017), and it becomes an interesting model to which to compare our global salience model, since we have seen in Global visual salience section that images containing faces and humans tend to get a higher global salience score.

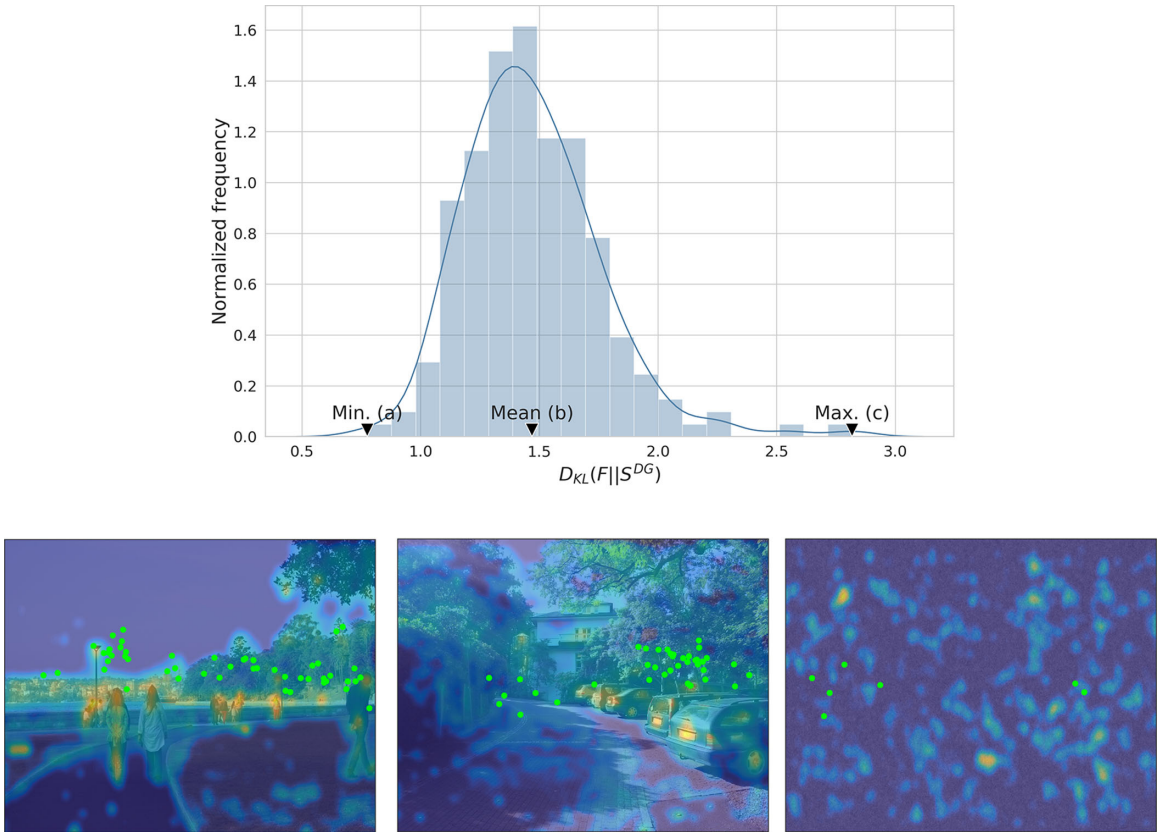
In general, we observe that unlike GBVS, measures derived from Deep Gaze salience maps exhibit a nonzero, yet moderate correlation with our proposed global salience. For instance, Figure 4b shows a slight negative correlation between global salience scores and the KLD between first fixation distributions and Deep Gaze salience maps. However, looking at the distribution of the Kullback-Leibler divergence in Figure 6, we see that the salience maps are also far from matching the location of the first fixations on the images. Finally, we also observed (see Figure 7) a nonzero correlation between the difference of global salience scores between the left and the right image, as well as the difference in salience mass computed with Deep Gaze.

Taken together, we can conclude that our proposed computational model provided a robust method to rank images according to a unique global image salience that is independent of the low-level local salience properties of the stimuli, and we observed a nonzero, yet moderate correlation with a computational salience model that incorporates higher-level cues.

## Lateral bias

Our third hypothesis (H3) stated that a general spatial bias leads to a higher likelihood to first fixate on the left rather than the right image. We thus calculated the number of first saccades that landed onto the left and the right image for each block separately (Figure 8). A 4 × 2 (block: 1, 2, 3, 4 × image side: left, right) repeated-measures analysis of variance (ANOVA) (Greenhouse-Geisser corrected) revealed a general spatial bias of the initial saccade toward the left image as indicated by a significant main effect according to the image side, $F(1, 48) = 30.833$; $p < .001$; $\eta_p^2 = .391$. No further effects were found (all $F \leq 2.594$; all $p \geq .074$, all $\eta_p^2 \leq .051$), showing that the left bias was present in all blocks to a similar extent. Thus, we can conclude that the participants generally targeted their initial saccades more on left-than-right sided images.

Nonetheless, the error bars in Figure 8 suggest a high variability of the lateral bias across subjects. In order to investigate this, we calculated the number of first saccades on the right image for each participant separately. Moreover, since our model included an individual bias term for each participant, as described in Methods: computation of global salience section, we can also look at the magnitude of the coefficients learned by the model. In Figure 9, we plot, for each participant, the percentage of first saccades toward the right image and their corresponding lateral bias term




(a) $D_{KL}(F||S) = 0.57$ (b) $D_{KL}(F||S) = 1.44$ (c) $D_{KL}(F||S) = 3.19$

Figure 6. Top row: distribution of the Kullback-Leibler divergence between the first fixations map and the Deep Gaze local salience maps. Bottom row: images with the minimum, closest to the mean and maximum KLD, with their overlaid salience map and the location of the first fixations.

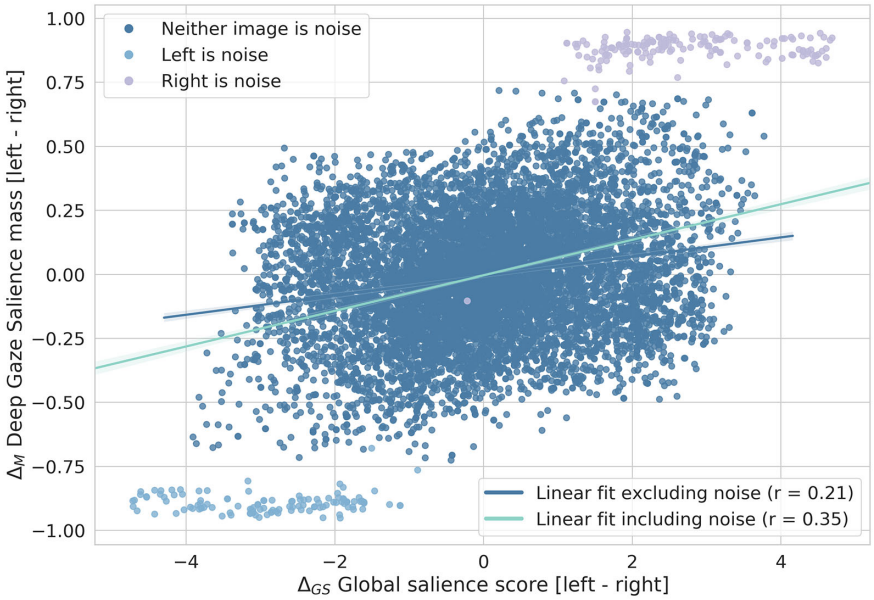



Figure 7. Correlation between the Deep Gaze image salience mass and the global salience scores.

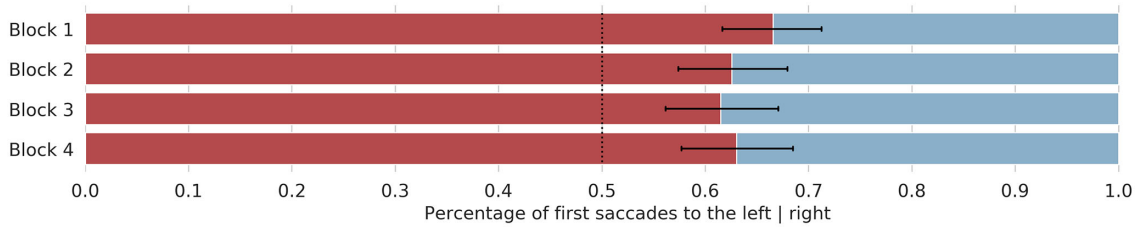



Figure 8. Percentage of first saccades that targeted on the left (red) and right (blue) images, at each block of the experimental session. Error bars depict the standard deviation of the mean. Note that considerably more first fixations landed on the left image, highlighting the lateral bias.

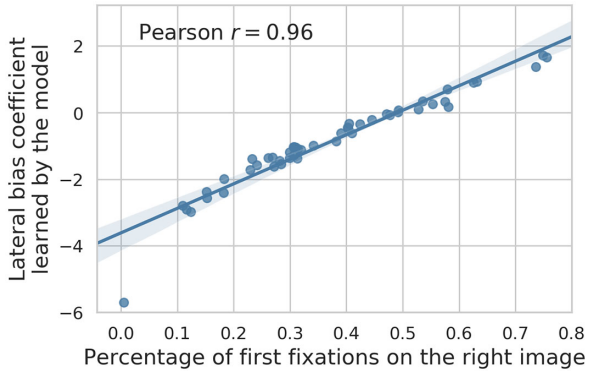



Figure 9. Lateral bias of each participant, as measured by the percentage of first fixations onto the right image and the lateral bias terms learned by our computational model. Both metrics are highly correlated and reveal the average left bias, but with high variability across participants.

learned by the computational model. Both metrics are highly correlated—further highlighting the validity of the model—and reveal a high variability in the lateral bias across participants. Overall, 63% of all first fixations landed on the left image.

## Task and familiarity

Next, we investigated the effect of the familiarity with one of the images and of the task of selecting the already seen or unseen image, which the participants had to perform in Blocks 2 and 3 of the experiment, respectively. In particular, we were interested in finding out whether there is a tendency to direct the initial saccade toward the task-relevant images or toward the new images, for instance. In our fourth hypothesis (H4), we stated that our task and familiarity should have little or no influence on the initial saccade. For that purpose, we first performed a 2 × 2 (task: select new, select old × fixated image: new image, old image) repeated-measures ANOVA analysis (Greenhouse-Geisser corrected). The results revealed no significant effects (all $F \leq 1.936$; all $p \geq .170$, all $\eta_p^2 \leq .039$) (Figure 10). Thus, the provided tasks did not bias the initial saccade decision to target one of the two presented images. Nevertheless, we found that participants correctly identified 91.43% of the new images in Block 2 and 91.16% of the old images in Block 3. Hence, the task performance was highly above chance (50%) and the participants were accurate in identifying the new and old images, respectively.

Also in this case, the same conclusion can be extracted from the coefficients learned by the model to capture the task and familiarity effects, which are −0.04 and −0.10, respectively, that is, very small and only slightly higher for the familiarity.

Taken together, spatial properties influenced the initial saccade in favor to fixate left-sided images first. Although task performance was very high, neither the task nor the familiarity with one of the images had an influence in the direction of the first fixation after stimulus onset. These results fully support our third and fourth hypotheses.

## Total exploration of images

In our fifth hypothesis (H5), we stated that images with higher global image salience lead to a longer exploration time than images with lower global salience. We thus calculated the relative dwell time on each image, left and right, for each trial. As an initial step, similar to the analysis of the initial saccade, we analyzed the potential effect of the spatial image location as well as the task and familiarity relevance on the exploration time.

With respect to the spatial image location, a 4 × 2 (block: 1, 2, 3, 4 × image side: left, right) repeated-measures ANOVA (Greenhouse-Geisser corrected) revealed a significant main effect according to the block, $F(2.368, 113.668) = 12.066$, $p < .001$, $\eta_p^2 = .201$, but no further effects (all $F \leq 2.232$; all $p \geq .109$, all $\eta_p^2 \leq .044$). Thus, the total time of exploration did not depend on the spatial location of the images, as also shown in Figure 11.

With respect to the task relevance (recall: Block 2—select new image; Block 3—select old image), we

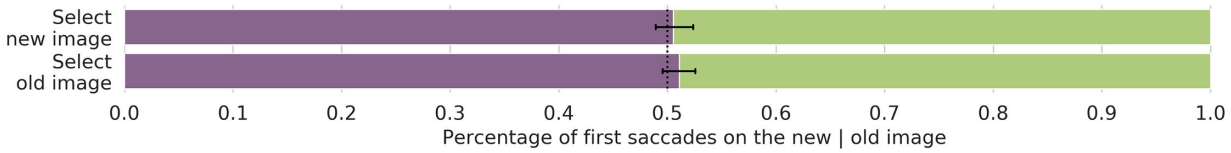



Figure 10. Percentage of first saccades that targeted on the new (purple) and old (green) images, at Blocks 2 and 3, where participants had the task of indicating the new and old image, respectively. Error bars depict the standard deviation of the mean. No significant bias can be appreciated in this case.

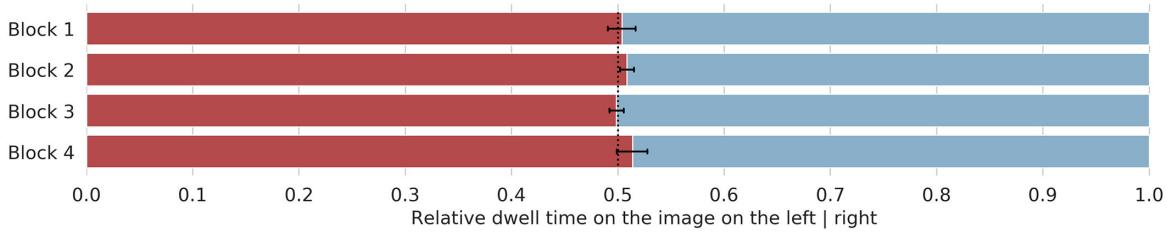



Figure 11. Exploration as measured by the relative dwell time on the left (red) and right (blue) images, at each block of the experimental session. Error bars depict the standard deviation of the mean.

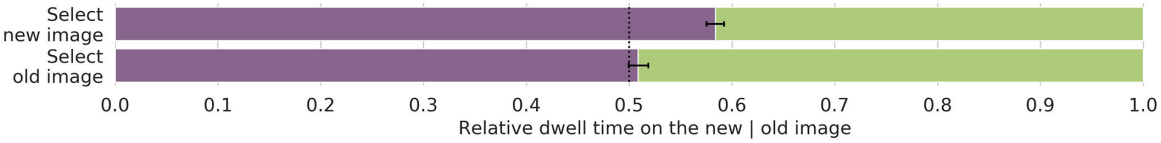



Figure 12. Exploration as measured by the relative dwell time on the new (purple) and old (green) images, at Blocks 2 and 3, where participants had the task of indicating the new and old image, respectively. Error bars depict the standard deviation of the mean.

calculated a 2 × 2 (task: select new, select old × fixated image: new image, old image) repeated-measures ANOVA (Greenhouse-Geisser corrected). The results revealed a significant main effect according to the task, $F(1, 48) = 4.298$, $p < .050$, $\eta_p^2 = .082$, and fixated image $F(1, 48) = 64.524$, $p < .001$, $\eta_p^2 = .573$, as well as an interaction between task and fixated image, $F(1, 48) = 36.728$, $p < .001$, $\eta_p^2 = .433$. As shown by Figure 12, our results showed that, in general, participants tended to spend more time exploring new instead of previously seen images. Furthermore, this effect was noticeably larger in Block 2, where the task was to select the new images, than in Block 3 (select old image).

Consequently, we found that the spatial location of images did not affect the total time of exploration. Instead, the task and familiarity had a considerable impact on the exploration time, revealing that new images were explored during a longer time than the counterpart.

For our main analysis regarding the interaction between exploration time and global image salience, we then contrasted the global salience score learned for each image with its respective dwell time averaged over all trials and subjects. The results revealed a significant

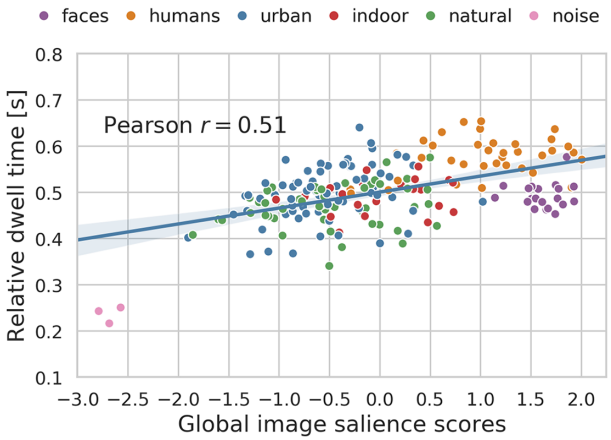



Figure 13. Dwell time versus global salience scores.

positive correlation, indicating that images with larger global image salience led to a more intense exploration (Figure 13). Thus, global image salience describes not only a measure of which image attracts initial eye movements, but is also connected to longer exploration time, suggesting that global salience may describe the relative engagement of images.

Taken together, our results suggest that the task and familiarity (but not the spatial location of images) influenced the exploration time with respect to higher dwell times on unseen images in combination with the task to select the new image. Note, however, that regarding the effects of task, our findings are restricted to the specific task assigned in our experiments, that is, selecting which image is new or old. The effect of task in visual attention is an active field in visual perception, and the results of multiple contributions should be taken together into consideration to draw robust conclusions. Finally, we also found that images with higher global salience correspondingly led to a larger time of exploration. These results fully support our fifth hypothesis.

## Discussion

We have presented a computational model trained on the saccadic behavior of participants freely looking at pairs of competing stimuli, which is able to learn a robust score for each image, related to its likelihood of attracting the first fixation. This fully supports our first hypothesis, and we refer to this property of natural images as the global visual salience.

The computational model consists of a logistic regression classifier, trained with the behavioral data of 49 participants who were presented 200 pairs of images. In order to reliably assess the performance of the model, we carried out a careful 25-fold cross-evaluation, with disjoint sets of participants for training, validating, and testing. Given a pair of images from the set of 200, the model predicted the direction of the first saccade with 82% accuracy and 0.88 area under the receiver operating characteristic curve.

Throughout the article, we have analyzed the general lateral bias toward the left image (H2), as well as other possible influences such as the familiarity with one of the images and the effect of a simple task (H3). Moreover, we have analyzed the relationship of our proposed global salience with the local salience properties of the individual images (H4). Finally, we have also studied the total exploration time of each image in the eye-tracking experiment and compared it to the global salience, which is based upon the first fixation (H5).

Regarding the lateral bias, we found that participants tended to look more frequently toward the image on the left. Such left bias is typical in visual behavior and has been found in many previous studies (Barton et al., 2006; Guo et al., 2009; Calen Walshe & Nuthmann, 2014; Ossandón et al., 2014). However, most of these studies presented only single images per stimulus. In this regard, it has been argued that cultural factors of the Western population who mostly take part in the research experiments may lead to a semantic processing of natural visual stimuli similar to the reading direction, that is, from left to right (Spalek & Hammad, 2005; Zaeinab et al., 2016).

In our study, about 63% of the first fixations landed on the left image. However, we also observed a high variability across participants, successfully captured by our computational model. In contrast, we showed that the given task in certain trials did not influence initial saccade behavior. Participants equally distributed the target location of saccades on the presented images, regardless of familiarity and task relevance. Consequently, the spatial location of an image affected saccade behavior, whereas the task as well as familiarity had no influence.

Importantly, we found that global salience, that is, the likelihood of an image attracting the first fixation when presented next another competing image, is independent of the low-level local salience properties of the respective images. The location of the first fixations made by the participants in the study did not correlate with the GBVS salience maps of the images, and the saccadic choice—left or right—was neither explained by the GBVS salience mass difference. Hence, our results provide some new insights in the understanding of visual perception of natural images, showing that the global salience of an image is rather affected by the semantics of the content. For instance, images involving socially relevant content such as humans or faces led to higher global salience than images containing purely indoor, urban, or natural scenes.

To gain further insight regarding this aspect, we computed the salience maps using Deep Gaze II (Kümmerer et al., 2016), a computational salience model that is not limited to low-level features but also makes use of high-level cues, obtained by pretraining the model with image object recognition tasks. We repeated the same analyses as with the GBVS model and we found that metrics derived from Deep Gaze salience maps did have a nonzero, yet moderate correlation with our proposed global salience. This, together with previous evidence about the importance of low- and high-level features in detecting fixations (Kümmerer et al., 2017), matches our finding that global salience cannot be explained by low-level properties of the images. However, the relatively low correlation further suggests that the initial preference for one of the images does not depend only on properties of the individual salience maps.

According to previous research, initial eye movements in young adults are based on bottom-up image features, whereas socially relevant content is fixated later in time (Açık et al., 2010). Interestingly, as described above, we found that this was not the case when two images have been shown at the same time. Considering the very short reaction time between stimulus onset and the observers, reaction to fixate one

of the two images, it seems surprising that participants had to prescan both images in their peripheral visual field before initializing the first saccade. Thus, in contrast to classical salience maps, we might argue that the global salience of an image highly relates to the semantic and socially relevant content.

In order to further investigate the effects of the global image salience, we also evaluated the total time of image exploration, that is, the dwell time. We hereby found that, different from the initial saccade, the spatial location of images did not affect the time participants explored the individual images of each image pair. However, the task and familiarity had an effect. We saw that in the task where participants had to select the new image, new images were explored longer than previously seen images. In contrast, the task asking to select the old image led to an almost equal exploration time on new and familiar images. Therefore, we conclude that participants in general tended to explore new images for a slightly longer time. Nevertheless and most important, we saw generally—and independent of the spatial location, task, and familiarity—that images with higher global salience were explored longer in time. Thus, images with larger global salience did not only attract initial eye movements after stimulus onset but also led to longer exploration times. These results support our assumption, that the global salience score of an image can also be interpreted as a measure of the general attraction of an image, in comparison to other images.

In this regard, note that although we considered the location of the first fixation as the target variable to model the global salience scores and carry out the subsequent analyses, the same computational model and procedures can be used to model alternative aspects of the behavioral responses. For instance, the model could be trained to fit the dwell time—which we have found to be positively correlated with the global salience based on the first fixations—the engagement (time until fixating away), or the number of saccades.

Despite the high performance of our computational model and its potential to assign reliable global salience scores to natural images, an important limitation is that the model and thus the scores are dependent on the image set that we used. Whereas local salience maps rely on image features, our proposed global salience model relies on the differences between the stimuli and the behavioral differences that they elicit on the participants. We observed significant differences between image categories, for example, humans versus indoor scenes, but this is only one initial step, and future work should investigate what other factors influence the global image salience. For example, it would be interesting to train a deep neural network with a possibly larger set of images and the global salience scores learned by our model as labels, similarly to how Deep Gaze was trained to predict fixation locations. This could shed more light on what features make an image more globally salient.

Another related, interesting avenue for future work is investigating the global salience in homogeneous data sets, that is, with images of similar content. Our work has shown that large differences exist between images with somehow different content, for instance, containing humans or not. However, we did not observe significantly different global salience between natural and urban scenes (see Figure 2b), although significant differences do exist between specific images. An interesting question is: *What* makes one image more likely to attract the first fixation, when presented alongside a semantically similar image? We think an answer to this question can be sought by combining a similar experimental setup to the one presented in this work, with additional data, and making use of advanced feature analysis, such as deep artificial neural networks, as mentioned above.

For instance, small changes in the context information of single images might already have a dramatic influence on reaction times in decision tasks (Kietzmann & König, 2015). In addition, the global salience was based on eye movement behavior of human data. Depending on the choice of participants, for example, different culture, age, personal interests, and emotions, our model could have revealed different results (Balcetis & Dunning, 2006; Dowiasch et al., 2015; Kaspar et al., 2015). Again, further studies might use the model on a wider range of participants, in order to validate the specific global salience and thus attraction of images.

In contrast, differences in the global salience between participant groups could be a great advantage in certain research fields. In medical applications, for instance, researchers could identify specific diseases, such autistic spectrum disorder (ASD). In such an example, our method could generate a model of the global visual salience of both control people and individuals with certain conditions, and then be used for diagnosis. Another use of our model would be marketing research, where the attraction of different images could be compared adequately based on intuitive visual behavior. Thus, depending on the research question, the global image salience might provide a new insight in prediction and analysis of visual behavior.

## Conclusion

Previous research has investigated the local salience properties of single images, which has helped understand visual behavior. However, assigning a single and unique global salience score to an image as a whole has been neglected. Here, we thus trained a logistic regression model to learn unique, global salience scores

for each tested image. We hereby showed that images can indeed be ranked according to their global salience, providing a new method to predict eye movement behavior across images with distinct semantic content. These results could be used in a variety of research, such as medicine or marketing.

*Keywords: image salience, visual behavior, overt attention, spatial exploration*

## Acknowledgments

The authors thank Deutsche Forschungsgemeinschaft (DFG) and Open Access Publishing Fund of Osnabrück University.

Supported by BMBF and Max Planck Society.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant agreement No 641805.

Commercial relationships: none.
Corresponding author: Alex Hernández-García.
Email: ahernandez@uos.de.
Address: Wachsbleiche 27, 49090 Osnabrück (Germany).

## Footnote

[1]While there is no consensus about the best metric for the evaluation of logistic regression, the coefficient of discrimination $R^2$ proposed by Tjur (2009) has been widely adopted recently, as it is more intuitive than other definitions of coefficients of determination and still asymptotically related to them.

## References

Açk, A., Sarwary, A., Schultze-Kraft, R., Onat, S., & König, P. (2010). Developmental changes in natural viewing behavior: Bottom-up and top-down differences between children, young adults and older adults. *Frontiers in Psychology, 1*.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46*, 2824–2833.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91*, 612–625.

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception, 35*, 1089–1105.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika, 39*, 324–345.

Calen Walshe, R., & Nuthmann, A. (2014). Asymmetrical control of fixation durations in scene viewing. *Vision Research, 100*, 38–46.

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology, 14*, R850–R852.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*, 201–215.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Dowiasch, S., Marx, S., Einhäuser, W., & Bremmer, F. (2015). Effects of aging on eye movements in the real world. *Frontiers in Human Neuroscience, 9*.

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology, 48*, 269–297.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14):18, 1–26.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research, 9*, 1871–1874.

Geisler, W. S., & Cormack, L. K. (2011). Models of overt attention. In S. P. Liversedge, I Gilchrist, & S. Everling (Ed.), *The Oxford handbook of eye movements* (pp. 439–454). Oxford, UK: Oxford University Press.

Guo, K. (2007). Initial fixation placement in face images is driven by topdown guidance. *Experimental Brain Research, 181*, 673–677.

Guo, K., Meints, K., Hall, C., Hall, S., & Mills, D. (2009). Left gaze bias in humans, rhesus monkeys and domestic dogs. *Animal Cognition, 12*, 409–418.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16*, 850–856.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in neural information processing systems*, pp. 545–552.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Kaspar, K. (2013). What guides visual overt attention under natural conditions? Past and future research. *ISRN Neuroscience, 2013*, Article 868491, 1–8.

Kaspar, K., & König, P. (2012). Emotions and personality traits as high-level factors in visual attention: a review. *Frontiers in Human Neuroscience, 6*, Article 321.

Kaspar, K., Ramos Gameiro, R., & König, P. (2015). Feeling good, searching the bad: Positive priming increases attention and memory for negative stimuli on webpages. *Computers in Human Behavior, 53*, 332–343.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience, 23*, 315–341.

Kietzmann, T. C., & König, P. (2015). Effects of contextual information and stimulus ambiguity on overt visual sampling behavior. *Vision Research, 110*, 76–86.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In L. Vaina (Ed.), *Matters of intelligence* (pp. 115–141). Dordrecht, Netherlands: Springer.

Kollmorgen, S., Nortmann, N., Schröder, S., & König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology, 6*, e1000791.

König, P., Wilming, N., Kietzmann, T. C., Ossandón, J. P., Onat, S., Ehinger, B. V., Ramos Gameiro, R., . . . Kaspar, K. (2016). Eye movements as a window to cognitive processes. *Journal of Eye Movement Research, 9*, Article 3, 1–16.

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research, 51*, 1457–1483.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.

Kümmerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4789–4798). Venice, Italy: IEEE.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences, 4*, 6–14.

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. New York, NY: Dover.

Niebur, E., & Koch, C. (1996). Control of selective visual attention: Modeling the "where" pathway. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Ed.), *Advances in neural information processing systems* (pp. 802–808). Denver, CO, USA.

Ossandón, J. P., Onat, S., & König, P. (2014). Spatial biases in viewing behavior. *Journal of Vision, 14*(2):20, 1–26.

Ramos Gameiro, R., Jünemann, K., Herbik, A., Wolff, A., König, P., & Hoffmann, M. B. (2018). Natural visual behavior in individuals with peripheral visual-field loss. *Journal of Vision, 18*(12), https://doi.org/10.1167/18.12.10.

Ramos Gameiro, R., Kaspar, K., König, S. U., Nordholt, S., & König, P. (2017). Exploration and exploitation in natural viewing behavior. *Scientific Reports, 7*, 2311.

Rauthmann, J. F., Seubert, C. T., Sachse, P., & Furtner, M. R. (2012). Eyes as windows to the soul: Gazing behavior is related to personality. *Journal of Research in Personality, 46*, 147–156.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the center of gaze. *Computation in Neural Systems, 10*, 341–350.

Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013). Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision* (pp. 1153–1160). Sidney, Australia: IEEE.

Spalek, T. M., & Hammad, S. (2005). The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science, 16*, 15–18.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, https://doi.org/10.1167/7.14.4.

Tatler, B.W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition, 17*, 1029–1054.

Tjur, T. (2009). Coefficients of determination in logistic regression modelsa new proposal: The coefficient of discrimination. *The American Statistician, 63*, 366–372.

Wadlinger, H. A., & Isaacowitz, D. M. (2006). Positive mood broadens visual attention to positive stimuli. *Motivation and Emotion, 30*, 87–99.

Zaeinab, A., Ossandón, J. P., & König, P. (2016). The dynamic effect of reading direction habit on spatial asymmetry of image perception. *Journal of Vision, 16*(11), https://doi.org/10.1167/16.11.8.