ANNUAL REVIEWS

Annual Review of Vision Science

Predicting Visual Fixations

Matthias Kümmerer and Matthias Bethge

Tübingen AI Center, University of Tübingen, Tübingen, Germany;
email: matthias.kuemmerer@bethgelab.org, matthias@bethgelab.org

Annu. Rev. Vis. Sci. 2023. 9:269–91

First published as a Review in Advance on July 7, 2023

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-120822-072528

Keywords

fixations, eye movements, saliency, model comparison, information theory, benchmarking, transfer learning, unifying framework, taxonomy

Abstract

As we navigate and behave in the world, we are constantly deciding, a few times per second, where to look next. The outcomes of these decisions in response to visual input are comparatively easy to measure as trajectories of eye movements, offering insight into many unconscious and conscious visual and cognitive processes. In this article, we review recent advances in predicting where we look. We focus on evaluating and comparing models: How can we consistently measure how well models predict eye movements, and how can we judge the contribution of different mechanisms? Probabilistic models facilitate a unified approach to fixation prediction that allows us to use explainable information explained to compare different models across different settings, such as static and video saliency, as well as scanpath prediction. We review how the large variety of saliency maps and scanpath models can be translated into this unifying framework, how much different factors contribute, and how we can select the most informative examples for model comparison. We conclude that the universal scale of information gain offers a powerful tool for the inspection of candidate mechanisms and experimental design that helps us understand the continual decision-making process that determines where we look.

1. INTRODUCTION

By moving our gaze, we can select which part of our visual environment to see in high detail. In fact, we are continuously making unconscious decisions—a few times per second—about where to look next. In this article, we review how accurately these decisions can be predicted.

Researchers, including early pioneers like Aristotle and Leonardo da Vinci, have long been interested in how we move our eyes (for an extensive overview over the history of eye movement research, see Wade 2010). Buswell (1935) was likely the first researcher to focus on where we look, more precisely, where we fixate on pictures. Buswell looked into many effects that still interest the field today, including what shapes the spatial fixation distribution and how it changes over presentation time, fixation durations, interobserver consistency, and the influence of instructions given to the subjects. The last factor was made famous by Yarbus (1967), who explored the influence of tasks on eye movements in great depth.

Understanding how gaze is directed is fundamental to studying visually guided behavior. Eye movements reflect a form of visual attention that is, in contrast to covert attention, comparatively easy to measure and is usually closely tied to overt attention in natural behavior (Henderson 2003). Therefore, eye movements can be used to better understand cognitive processes. Eye movements can also be measured during everyday behaviors and reflect navigational strategies in solving a task, for example, when making tea or sandwiches (Hayhoe 2000, Land et al. 1999). Finally, there are many applications where understanding eye movements can help, e.g., when optimizing warning signs or advertisements, when cropping images, in scene understanding, and in robotics.

In this review, we give an overview of the developments in predicting fixations during the past two decades. We use the framework of probabilistic modeling to unify and organize different settings such that we can quantify and compare prediction performances. We begin with a taxonomy of the most popular settings in Section 2. Section 3 describes the evaluation of these settings, introducing information gain as a universal currency for judging effect size in Section 3.2. In Section 4, we review the key ingredients that have led to significant improvement in prediction performance in recent years and how information gain can be used as a tool for the inspection of candidate mechanisms and for experimental design. Finally, in Section 5, we give an outlook on what we consider to be important unresolved issues for predicting fixations that can be addressed with the current technology.

2. PROBABILISTIC FIXATION MODELING

Where we look next is influenced by many interacting factors. Some of them are relatively easy to control for (e.g., where we looked before or explicitly given tasks), while others are much harder to measure (e.g., previous experiences, personal interests, implicit tasks, or more general brain states). In addition, there are noise sources in the brain and in the oculomotor system that introduce random variability. Together, these factors lead to complex uncertainties such that simple point estimates are not suitable for the predication of visual fixations. Instead, the field of view splits into different complex shaped regions for which the probability is higher, lower, or close to impossible. Therefore, the task of predicting where people will look based on the history of visual input and behavior requires a Bayesian approach that explicitly models the posterior distribution over all possible fixation locations: Given the history up to the current point in time, t, in the observer's world, $W_{\leq t}$, and in the behavior of the observer, $o_{<t}$ (including the previous fixation trajectory, $x_{<t}$), we need to model the full probability distribution, $p(x_t | W_{\leq t}, o_{<t})$, over all locations where the observer could look next.

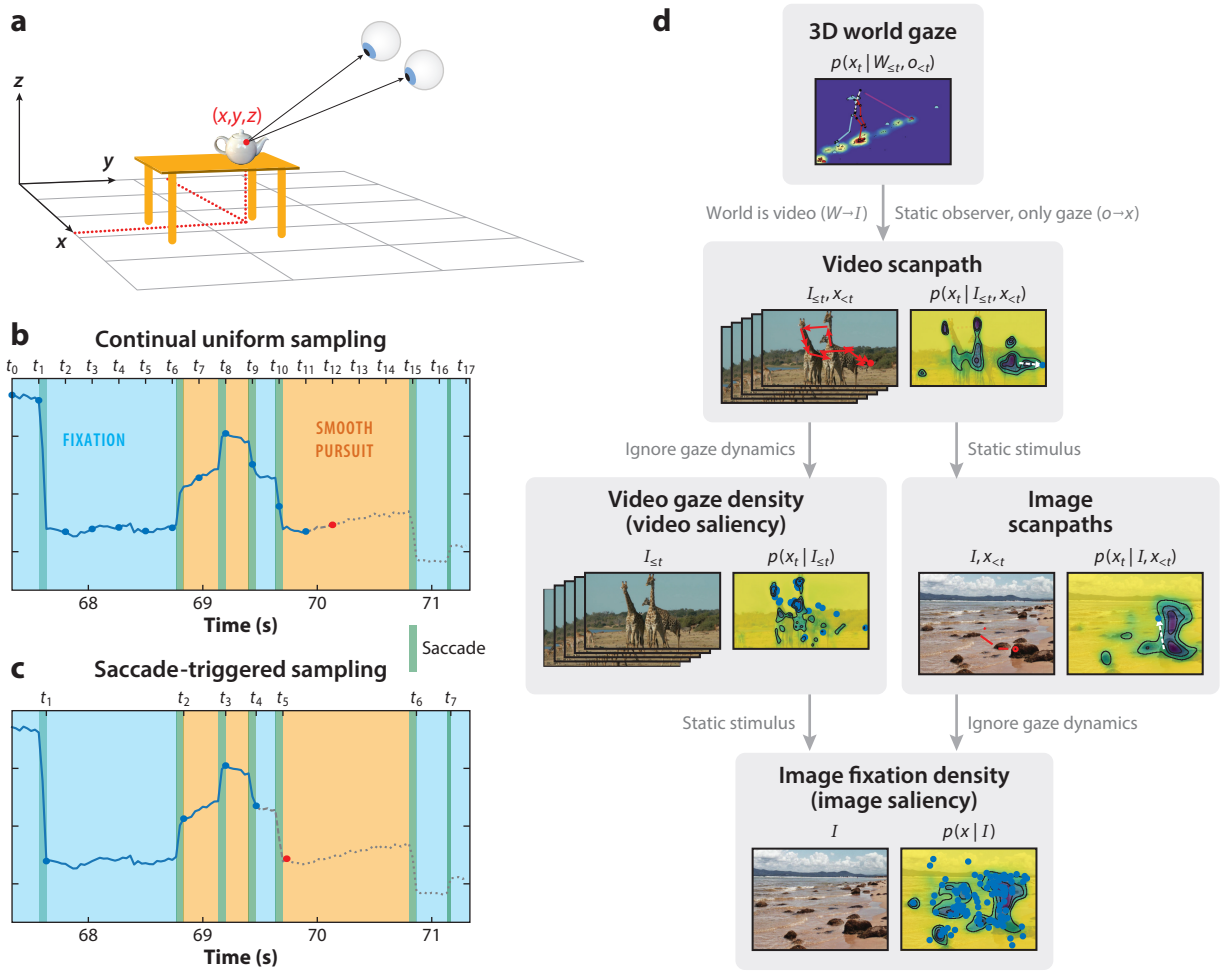

Figure 1

Predicting where we look. (*a*) Most generally, we look at points in the three-dimensional world. (*b*) Our gaze usually stays mostly constant for some time (fixations) or smoothly tracks a target (smooth pursuit), interleaved with fast repositioning eye movements (saccades). Gaze prediction in its most general form tries to predict the gaze trajectory by predicting where we will look after a short time span (x_{t+i}) given where we looked up to now ($x_{\leq t_i}$) or, more generally, the observer's behavior up to now ($o_{\leq t_i}$) and what happened in the world so far ($W_{\leq t_i}$). (*c*) Alternatively, we can try to predict when and where we decide to relocate our gaze. (*d*) When we successively simplify the assumptions about the world by ignoring dependencies, a taxonomy of fixation and gaze prediction tasks arises. Three-dimensional world gaze saliency illustration adapted with permission from Matthis et al. (2018).

2.1. A Taxonomy of the Problem

In the most general form, x is a three-dimensional trajectory of gaze locations (**Figure 1*a***) in the observer's environment. When studying the continual dynamics of gaze movement, one can uniformly sample the gaze trace and predict the next gaze position, $x_{t_{i+1}}$, from the previous data as $p(x_{t_{i+1}} \mid W_{\leq t_{i+1}}, x_{\leq t_i})$ (**Figure 1*b***).

However, as illustrated in **Figure 1*c***, the trajectory of gaze movements naturally decomposes into intervals of variable lengths for which, in most cases, the gaze position stays constant (fixations) or smoothly tracks a target (smooth pursuit, or vestibulo-ocular movements, when the

head instead of the target moves). These intervals are separated by very brief intervals of fast eye movements (saccades), which reflect unconscious decisions to redirect the gaze to a new target (of course, this picture is somewhat simplified; for an extensive review of eye movements, see Kowler 2011). In this review, we focus on predicting the new gaze positions, x_{t_k}, at times t_k given by the end points of these saccadic events. That is, we ignore the gaze changes during smooth pursuits, as they can be considered a fixation on a moving target. Mathematically, this can be described as a point process in space and time defined by the probability $p(x_{t+\tau}, t+\tau \,|\, W_{\leq t+\tau}, o_{\leq t})$ for both when $(t+\tau)$ and where $(x_{t+\tau})$ we fixate next (**Figure 1*c***). Usually, we only want to predict the locations of fixations. In this case, by conditioning on the time when the next saccade happens, we can further reduce the setting to purely spatial prediction, $p(x_{t+\tau} \,|\, t+\tau, W_{\leq t+\tau}, o_{<t+\tau})$.

Above, we frame the problem of spatial fixation prediction in the most general way; however, modeling all dependencies is difficult. Therefore, it is common to simplify the problem by focusing only on a reduced set of dependencies. This gives rise to a taxonomy of fixation prediction tasks (**Figure 1*d***). The general setting described above may be termed three-dimensional world gaze prediction. Commonly, the world, W, is replaced with a video, I, such that the observer does not interact or move in the world. This is the setting of video scanpath prediction, where the problem simplifies to predicting $p(x_t, \, |I_{\leq t}, \, x_{<t})$ over the the two-dimensional fixation position x_t on the screen.

From video scanpath prediction, there are two complementary ways to reduce the two factors $I_{\leq t}$ and $x_{<t}$ used for the prediction. One option, video gaze density prediction, ignores $x_{<t}$ (where the observer looked before) and uses only the video to predict the next fixation: $p(x_t | I_{\leq t})$. The other option, image scanpath prediction, keeps the gaze history but uses still images instead of videos as visual input. We then need to predict $p(x_t | I, x_{<t})$. Finally, both options, video gaze density prediction and image scanpath prediction, can be further simplified into the same, simplest case, where both the previous scanpath and the history of the visual input are ignored. In this case, known as image fixation density prediction, we only need to predict the average fixation density, $p(x|I)$, as a function of static images I.

This taxonomy categorizes the most prominent methods of gaze prediction, but not the only ones. For example, one might want to include additional potentially relevant factors, such as systematic differences between different subjects or different tasks. The probabilistic framework allows one to include any such factors as condition variables and to compare all of these models in a principled way.

The probabilistic framework can also be extended to predicting not only the next fixation, but also further behavioral variables such as present/nonpresent responses in a visual search task. Such a response can be considered an alternative decision to making a saccade; thus, instead of predicting a probability distribution for the next fixation location, we could predict a probability distribution for the next decision. Some probability mass would be assigned to making a fixation in the different image locations, and the remaining probability mass would be distributed over the response options.

2.2. Saliency Modeling

Where we look is heavily influenced by the task at hand (Yarbus 1967); thus, substantial research has been conducted on eye movements in specific tasks, such as making tea (Land et al. 1999), avoiding obstacles (Rothkopf et al. 2007), and walking (Matthis et al. 2018). However, the notion of saliency builds upon the idea that some visual inputs attract our attention and our gaze largely independent from the specific task at hand: They are salient. Saliency was originally proposed to be a very strong low-level property of visual scenes (Itti & Koch 2001a, Li 2002), and vision

scientists tried to identify the features that determine the saliency of different spatial regions in the visual input. The interest of computer vision research in saliency prediction focuses more on potential applications ranging from optimizing warning signs and advertisements to compression and smart image cropping.

Saliency research commonly uses so-called free-viewing conditions, i.e., conditions in which the observer does not have any specific task. This is motivated by the idea that the absence of a specific task should make the influence of saliency most visible. One common objection to these conditions is that not prescribing a task simply means that we do not control which task observers are doing (Tatler et al. 2005). However, we can argue that, by averaging across many different implicit tasks pursued by different observers, we obtain a residual net effect that defines salience, at least for practical purposes. With respect to our taxonomy, saliency research is most prominent in two areas of the taxonomy: image saliency prediction and video saliency prediction, which are fixation density prediction under free-viewing conditions on still images and videos, respectively.

The remainder of this review focuses on eye movement predictions in free-viewing conditions, for which there have been substantial efforts to provide public data sets and establish model comparability. The methods that we discuss, however, are not limited to this case and but can equally be used to predict fixations under task-dependent conditions.

3. EVALUATION OF FIXATION MODELS

"Since all models are wrong the scientist must be alert to what is importantly wrong" (Box 1976, p. 792). This quote[1] also applies to models of fixation prediction. Consider a case in which, presented with one or multiple models, we would like to know which model is better relative to the other(s). Ideally, we would also be able to establish a gold standard that allows us to judge quantitatively how close a model is to solving a given prediction problem on an absolute scale.

Being able to quantify prediction performance depends on having good scoring functions (defining what is "importantly wrong") and good data sets (specifying the range of cases for which we want the models to be valid). Oftentimes, different studies use different data sets and different metrics, making it difficult to compare their results. This problem can be alleviated by the use of common data sets and metrics for model evaluation. These so-called benchmarks facilitate comparability such that we can keep track of progress and the current state of the art (within the limited scope of the benchmark).

The field of computer vision is extremely benchmark driven. It puts so much emphasis on comparable model evaluation that concerns have been raised that the field values incremental progress on a benchmark more than scientific insights (Wagstaff 2012). In contrast, the field of vision science rarely uses benchmarking and often focuses more on studying specific mechanisms to explain observed effects. In this section, we take a synergistic view advocating the idea that quantitative model comparison can serve as a powerful tool for the identification of mechanisms. In contrast to significance testing for whether an effect exists, careful benchmarking tests also give feedback about effect size and thus relevance of a hypothesized mechanism. At the same time, it is also important not to get too attached to a single benchmark data set, which may lead to overfitting, especially when one is only looking at average performance without looking at the individual errors that different models make.

If we want to use benchmarking to probe the effect size of different candidate mechanisms, such as, for example, to what extent fixations are driven by low-level versus high-level features, then it is not enough to determine a ranking among different models; we also need to be able to

[1]A variant of the more famous "all models are wrong but some are useful" (Box 1979, p. 202).

say quantitatively how much more variability is explained by one model (e.g., one that includes high-level features) relative to another (e.g., one that excludes high-level features).

In this section, we discuss the different benchmarking methods most commonly used for static image saliency models, image scanpath models, and video gaze models and explain how the unifying framework of probabilistic modeling can facilitate direct comparisons among the three different settings. For example, one can directly compare the effect size of scanpath history or video input history to the effect size of predicting fixations based on a single image input only.

3.1. Image Saliency

The field of static image saliency research, i.e., the field of predicting the spatial distribution of free-viewing fixations on static images, has been extremely active since Itti et al. (1998) began it with their computational model. Many new models are published every year, and more than 100 models have been evaluated on the MIT/Tuebingen Saliency Benchmark and its predecessor, the MIT Saliency Benchmark [originally established by Judd et al. (2012) and continued by us since 2019]. These benchmarks typically score prediction performance on data sets consisting of fixation data from multiple subjects free viewing images for a few seconds, e.g., the MIT300 data set (300 natural scenes, 45 subjects, 3 s presentation time) (Judd et al. 2012) or the CAT2000 data set (2,000 images from different categories, 24 subjects, 5 s presentation time) (Borji & Itti 2015).

The history of the MIT/Tuebingen Saliency Benchmark reflects how the evaluation of saliency models has substantially evolved over time. Historically, saliency models have not been designed to directly predict the two-dimensional fixation density, $p(x|I)$, which we describe in the taxonomy of fixation prediction. Instead, the majority of saliency models predict so-called saliency maps. Saliency maps were originally envisioned to be computational modules of the attentional system, from which the brain selects locations of high saliency to attend next (see **Figure 2**; see the sidebar titled A Short History of Saliency: From Reaction Times to Eye Movements). However, it was never specified and commonly agreed on how saliency maps affect the actual fixations. Therefore, the notion of saliency maps became ambiguous, and it was unclear how to evaluate saliency models on ground truth fixation data. Over the years, a perplexing diversity of performance metrics has been proposed to quantify the predictive performance of saliency maps (see **Supplemental Appendix, section 1.1**).

Supplemental Material >

Multiple saliency metrics have long been used concurrently. The resulting evaluations appeared to be highly inconsistent, and the correlation between the scores obtained from different metrics

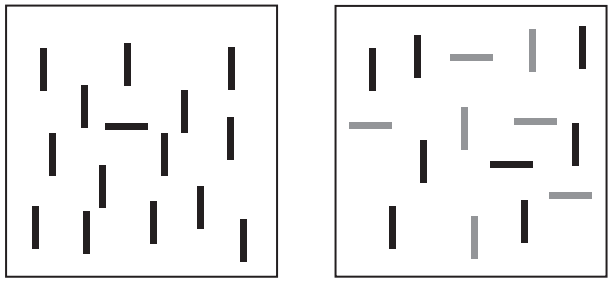

Figure 2

In visual search experiments, there are cases where the targets seem to pop out among the distractors (*left*; *horizontal bar* among *vertical bars*), while in other cases, the target does not immediately pop out and requires serial search (*right*; *horizontal black bar* among *horizontal* and *vertical black* and *gray bars*). These differences motivated the feature integration theory, which kickstarted the field of saliency models via the model of Itti et al. (1998).

A SHORT HISTORY OF SALIENCY: FROM REACTION TIMES TO EYE MOVEMENTS

The term "saliency" (or "saliency map") is used a lot in the context of eye movements but in different, ambiguous ways. The term saliency originated from attention research and made its way into eye movement research in multiple steps. In visual search experiments, some combinations of targets and distractors result in very short search durations, independent of the number of distractors: Targets seem to pop out from the stimulus (**Figure 2**, *left*). Other target–distractor combinations instead result in search durations depending linearly on the number of distractors, which have to be scanned serially (**Figure 2**, *right*). Treisman & Gelade (1980) proposed the feature integration theory to explain this effect with a two-stage attentional system. The first stage computes feature maps, which register elementary features highly parallel over the field of view and allow fast detection of targets defined by individual elementary features. Targets defined as conjunctions of multiple features require the second stage, a system of focused attention. This system can relate features to each other, but only in one location at a time, and thus requires serially shifting the focus of attention over candidate locations.

Koch & Ullman (1985) suggested a computational mechanism for the feature integration theory: For many elementary features, such as color and orientation, conspicuity maps are computed that single out locations that differ significantly from their environment with respect to the respective feature. The conspicuity maps are subsequently combined into one global conspicuity map, the so called saliency map, from which the most conspicuous location is selected with a winner-take-all mechanism. If the attended location does not turn out to be the actual search target, then the saliency map is locally suppressed, resulting in a shift of attention.

Itti et al. (1998) implemented the proposed mechanism of Koch & Ullman using color, intensity, and edges at different scales as elementary features and center-surround differences to detect feature popout. An important advantage of this model that made it particularly useful to a broad community was that it was image computable, i.e., it not only could process stimuli for which the elementary features were known by construction (e.g., Didday & Arbib 1975), but also was applicable to arbitrary images.

The mechanism of Koch & Ullman and the model of Itti et al. tried to explain covert shifts of attention; their goal was not to model eye movements. However, there is a close connection between covert and open attention (e.g., Henderson 2003). Itti & Koch (2001a,b) themselves suggested comparing model predictions to eye movements in follow-up work; this comparison was eventually performed by Parkhurst et al. (2002), who found that fixated image locations have above-average model saliency. Although Koch & Ullman originally envisioned the saliency map as a module of the covert attentional system, the evaluation by Parkhurst et al. of the model of Itti et al. established the terms saliency map and saliency model in the field of spatial fixation prediction. While most early models stayed close to the original concept of saliency as low-level feature popout (e.g., Bruce & Tsotsos 2009, Kienzle et al. 2009, Zhang et al. 2008), over the years, models have incorporated more and more high-level features (Judd et al. 2009) or even deep learning (Kümmerer et al. 2015a, Vig et al. 2014). It is important to be aware that, in vision science, the term saliency is commonly used with a meaning closer to its original one, inspired by the feature integration theory, whereas currently in computer vision, saliency is directly equated with fixation probability.

was as low as 0.31 (Pearson correlation) or 0.49 (Spearman rank correlation) (see also **Supplemental Figure 1**). Depending on the metric used, the same saliency model could yield state-of-the-art or below-baseline performance. This made it hard to assess the state of the field and has created some confusion in the literature. For example, while Einhäuser & König (2010, p. 389) stated that "Recent elaborations of such stimulus-driven models are now approaching the limits imposed by intersubject variability," a few years later, Borji et al. (2013b, p. 61) found "that a significant gap still exists between the best models and human interobserver agreement." At some point, it was even concluded that it is conceptually impossible to determine a best model independent of the different metrics; instead, it was believed that one must decide which metric is most

Supplemental Material >

important for one's work and accept bad scores in other metrics (Bylinskii et al. 2018, Riche et al. 2013a, Wilming et al. 2011).

The common denominator in all approaches to interpreting saliency maps is that higher saliency values should correspond to the expectation of more fixations. However, this still leaves large freedom in interpreting the actual scale of saliency values. For example, can we say what it means to double the amount of saliency? What does zero saliency mean? Framing saliency prediction as fixation density estimation, where the model has to predict a probability distribution $p(x|I)$ over image locations x depending on the observed image I (Baddeley & Tatler 2006, Barthelmé et al. 2013, Kümmerer et al. 2015b), resolves this ambiguity. For any possible region in the image, the probability distribution encodes how probable the model considers it that the fixation happens there. Twice the probability means that fixations are expected to occur twice as frequently in the considered region, and zero probability means that the model predicts that no fixation will ever occur there.

In the following sections, we explain how phrasing the task of spatial fixation prediction as a Bayesian fixation density prediction task gives rise to an evaluation framework that is both principled and intuitive and that can explain and resolve the disagreement among different saliency metrics.

3.1.1. Bayesian decision theory. Probabilistic modeling provides a way to understand and mostly solve the problem of inconsistent metrics. Following Bayesian decision theory, the evaluation of decisions is decomposed into two components: (*a*) a probability distribution over possible uncertain events that can be expected to happen and (*b*) a loss (utility) function that defines for each event and each decision how large the loss (utility) of the decision would be. Even if the probability distribution is perfectly known, optimal decisions still depend on the choice of the loss function. A well-known example of this dependency is given by the mean and the median of a distribution, which are optimal under the choice of the squared error loss and the absolute error, respectively, and which can differ substantially for, e.g., skewed distributions.

Similarly, even if the probability distribution over possible fixation locations is perfectly known, saliency maps can differ substantially if optimized for different metrics. This is demonstrated in **Figure 3**, where, for the same (posterior) distribution over all possible fixation locations in the image (fixation distribution) (**Figure 3*a***), substantially different saliency maps turn out to be optimal (**Figure 3*b***) with respect to different average utilities (the metric score) (**Figure 3*c***). While deriving the saliency map with the highest expected metric score for a given fixation distribution seems like a hard problem at first, it turns out that, for most commonly used saliency metrics, these saliency maps can be computed analytically or at least well approximated (for details, see Kümmerer et al. 2018).

The differences among the different optimal saliency maps for the same fixation distribution are quite striking. Some metrics require saliency maps to be smoother than the fixation density (CC and KL-Div), while others require them to have many zeros (SIM) or not to include the central fixation bias (sAUC). A saliency map that is optimal for one metric may perform very poorly on other metrics (**Figure 3*c***), and the performance variations among different metrics are often bigger than the differences among different models. This is also reflected by the fact that saliency maps often look more consistent if they are from different models, but for the same metric, than if they are from the same model, but for different metrics (**Supplemental Figure 2**).

Supplemental Material >

To resolve the large inconsistency among different metrics, it is important to tie the notion of a saliency model to the probability distribution of fixations, which is independent of the metrics needed for the evaluation. For any saliency model, we then need to generate different optimal saliency maps depending on the metric that we want to use for the evaluation. Even if one might

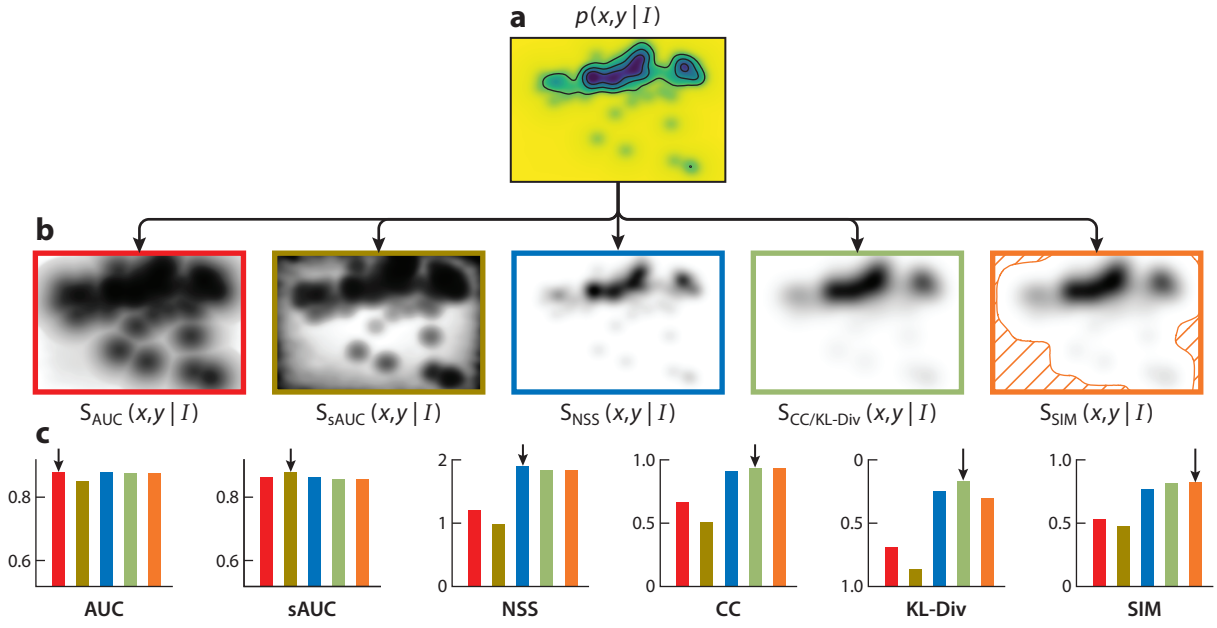

Figure 3

Bayesian decision theory provides a framework to derive from a fixation density (*a*) saliency maps optimal for different saliency metrics (*b*). The saliency maps differ dramatically due to the different properties of the metrics but always reflect the same underlying model. Areas with zero saliency are marked in red. The metric scores of the saliency maps under the different metrics (*c*) differ substantially per metric. The optimal saliency map (*bars marked with arrows*) always yields the highest metric score, while other saliency maps can be penalized substantially: No single saliency map can perform best in all metrics. Figure adapted with permission from Kümmerer et al. (2018).

still desire to use different metrics for different applications, one should never judge the mechanisms of different models based on comparisons of saliency maps that have been optimized for different metrics.

If models are evaluated in each metric using the correct metric-specific saliency maps, then it is possible to achieve highly consistent rankings among different models across virtually all commonly used metrics (Kümmerer et al. 2018). This important result resolves earlier concerns that objective model ranking was impossible (Riche et al. 2013a). It also resolves apparently contradictory results in the literature, such as the debate over whether objects or saliency attract fixations (Borji et al. 2013a, Einhäuser 2013, Einhäuser et al. 2008) and the diverging opinions on the progress of the state of the art (Borji et al. 2013b, Einhäuser & König 2010), as we detail in the **Supplemental Appendix, section 1.3**.

Supplemental Material >

3.1.2. The framework of Bayesian decision theory is backwards compatible. The framework of Bayesian decision theory can be adopted without the need to redo everything from scratch. If researchers design their new saliency model as a probabilistic model, then they can evaluate each metric on the correct saliency map. In this way, they can also compare their model to other saliency map models for which no density model is available using the saliency metric for which the other model was designed.

Existing (nonprobabilistic) saliency map models can only be optimal, and thus can only be evaluated fairly, with respect to the metric for which they have been designed. If we are interested to know how a (nonprobabilistic) saliency model performs with respect to other metrics, then it is

sufficient to optimize a postprocessing consisting of a pixelwise nonlinearity and a multiplicative center bias (Kümmerer et al. 2015b, 2018) for information gain (see Section 3.2). Subsequently, the resulting fixation density model can be used to derive optimal saliency maps for various metrics.

3.2. Judging Effect Size Using Information Theoretic Units as Universal Currency

In Section 3.1, we explain how the framework of Bayesian decision theory and the use of optimal saliency maps for different metrics yield consistent model rankings across different metrics. However, the choice of metric is still not arbitrary because the scales used by the different metrics can differ by arbitrary monotonic nonlinear mappings, which challenges the possibility of judging effect sizes. In this section, we go one step further and ask whether there is a particular choice of metric for which the scale is well calibrated such that it would allow us to judge quantitatively how much better one model explains the data in comparison to another model.

Following rate distortion theory (Cover & Thomas 2005), predictions always rely on an information channel, and the amount of information provided by the considered channel is a universal currency for the comparison of arbitrary information channels. Importantly, there is a precise notion for the amount of information transmitted by a channel that is independent of the choice of loss (utility) function (Shannon & Weaver 1949). Differences in the amount of information can be mapped directly to intuitively perceivable quantities, e.g., in the form of differing transmission times required to send images, videos, or text messages over the Internet.

In saliency modeling, the information channel is defined by the conditional distribution $p(x|I)$. The explainable information can be measured by the mutual information between fixation locations x and image I, which measures the difference in entropy, $h(x) - h(x|I)$, between the prior distribution, $p(x)$ (usually refered to as "center bias" because, on average, people tend to look more in the center of images; Tatler et al. 2005), and the posterior distribution, $p(x|I)$, after observation of an image I [averaged over all images according to $p(I)$]. The entropy of a distribution is measured in bits and can be directly related to the average coding cost of reporting an event drawn from this distribution, which we experience as waiting time in our daily life when digitally transmitting or storing data. The notion of bits relates to the degree of uncertainty (i.e., the minimum average number of yes/no questions that one must ask to remove uncertainty) and provides a universal scale for any type of model comparison that goes well beyond the saliency modeling problem (Shannon 1948).

To compare different models with respect to the amount of information that they can explain, it is sufficient to choose the model's log-likelihood as a metric. Differences in log-likelihood between two models can then be interpreted as follows: If model A's log-likelihood is one bit higher than model B's, then model A on average finds the data twice as likely (or is only half as surprised by the data; Itti & Baldi 2009). Alternatively, the log-likelihood difference in bits can also be illustrated by how much model A manages to shrink the area of a uniform distribution expressing the uncertainty of another model B (see **Figure 4**). Every bit of improvement in the log-likelihood of model A over model B reduces the uncertainty expressed by the area of the uniform distribution by a factor of two.[2]

When comparing to a baseline model such as the center bias, we call the log-likelihood difference information gain.[3] The information gain relative to the center bias quantifies how much

[2] This statement assumes that the support of the true fixation density is a subset of the support of the probability distribution of model A.

[3] Technically, information gain is the log-likelihood difference between a prior distribution, such as the image-independent center bias $p(x)$, and a posterior distribution, such as the image-dependent distribution

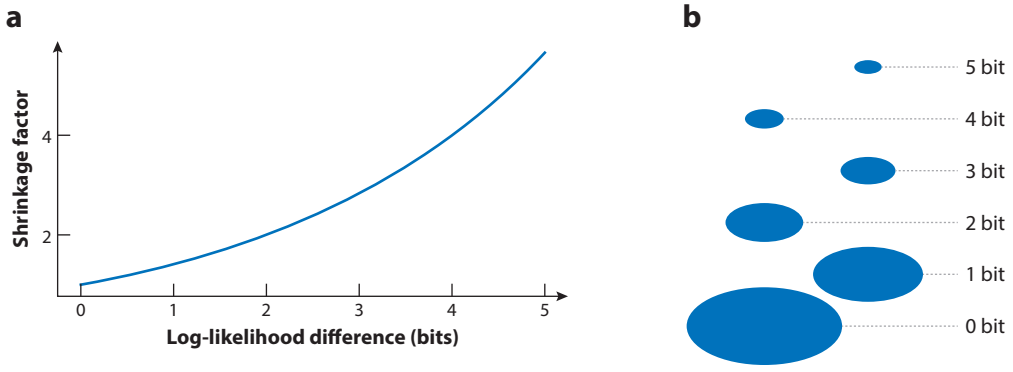

Figure 4

Log-likelihood differences. Intuitively speaking, log-likelihood differences measure how much a model manages to shrink the area of potential fixation locations. (*a*) One more bit means that the model can successfully constrain the scale of the area of fixation locations to $\frac{1}{2}\sqrt{2}$ of the previous size. (*b*) The area itself is halved with each bit.

additional information a model can gain from the given image about the spatial positions of fixations. Information gain defines a ratio scale where relative differences between models measured in bits can be compared just like relative differences in storage space or transmission times. This will become important below, when we discuss how much of the explainable information is captured by the current state of the art (see Section 4.2). A more detailed discussion of log-likelihood and information gain for fixation prediction is provided by Kümmerer et al. (2015b).

3.3. Scanpath Modeling on Static Images

To score the predictive performance of scanpath models, most studies resort to so-called scanpath-similarity metrics: For each ground-truth scanpath, they sample one or multiple new scanpaths from the model and compute a distance metric between human scanpaths and the best-matching model scanpath. The distance metrics are usually based on string-edit distances and compare fixations and saccades with respect to location, saccade amplitude, saccade direction, and other properties (for a review, see Anderson et al. 2015).

However, scanpath similarity metrics for model evaluation are inherently inconsistent, as wrong models can score systematically better than even the ground-truth model. The underlying problem is easy to grasp by an example: The smallest average distance to samples of a Gaussian distribution is not achieved by samples of the same Gaussian distribution, but instead by sampling from a delta-distribution (i.e., a Gaussian with vanishing variance) with the same mean. This also applies to scanpath distributions. Sampling multiple scanpaths per ground-truth scanpath and using only the best-matching one, as is commonly done, does not fully resolve the problem and introduces additional issues (for more details, see Kümmerer & Bethge 2021).

The problems can be resolved again by using probabilistic modeling. In our taxonomy, we define scanpath prediction as predicting the conditional probability distribution $p(x_t|x_{<t}, I)$ or, in terms of the sequence of fixations, $p(x_i|x_0, \ldots, x_{i-1}, I)$, i.e., as predicting the next fixation given where the observer looked before. The connection to predicting a full scanpath becomes clear when applying the chain rule: the conditional predictions for the next fixations give rise to a probability distribution over whole scanpaths, $p(x_1, x_2, x_3, \ldots, x_N \mid x_0, I) = \prod_{i=1}^{N} p(x_i \mid x_0, \ldots,$

$p(x|I)$. Although many models are not explicitly formulated as posteriors computed by conditioning a prior model on image information, we still use the term information gain in slight abuse of notation.

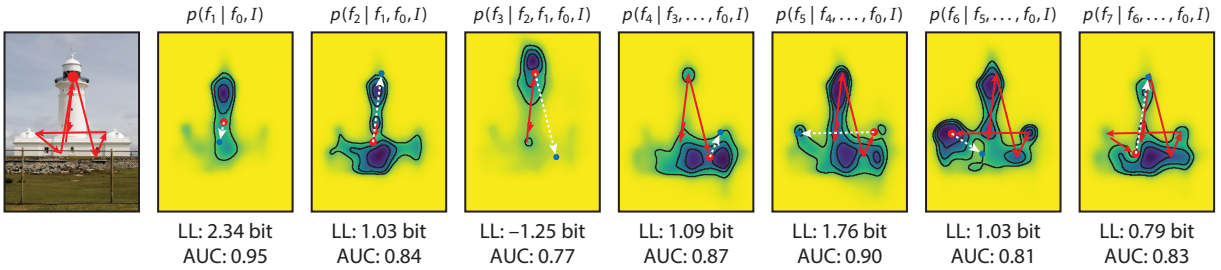

Figure 5

Scanpaths can be predicted and evaluated fixation by fixation. For each fixation, the model predicts a distribution over possible locations given the image and the previous scanpath history. The prediction quality can be scored for each fixation using, e.g., log-likelihood or AUC.

$x_{i-1}, I)$, where x_0 is the initial fixation at image onset (**Figure 5**). This framework has already been used for scanpath prediction in, e.g., the work of Schütt et al. (2017), Schwetlick et al. (2020), Malem-Shinitski et al. (2020), and Kümmerer et al. (2022).

Framing scanpath prediction as probabilistic next-fixation prediction allows us to employ the same metrics that we use for fixation density prediction: Using log-likelihood and information gain, we can score how well fixations are predicted on the same intuitive scale as in the fixation density case, making performance and effect size comparable. For example, starting from a simple spatial model, we can compare how much better predictions get by either improving the spatial fixation density prediction or adding scanpath dynamics via dependencies on previous fixations (Kümmerer et al. 2022). Using nonprobabilistic metrics like AUC and NSS allows us to compare to nonprobabilistic scanpath models, which can usually still be viewed in the framework of next-fixation prediction: For selecting fixation locations, these models often internally use a priority map, which corresponds to the conditional fixation distribution (for more details and a quantitative comparison of many scanpath models, see Kümmerer & Bethge 2021).

3.4. Video Saliency and Video Scanpath Prediction

Having discussed benchmarking in the case of image saliency above, discussing video saliency is rather straightforward because the prediction task is similar. In video saliency, models have to predict a fixation density (or gaze density) for each video frame, taking into account the preceding history of the video input. While there is no established video saliency benchmark with a holdout data set that would facilitate simple and fair comparison, there are some commonly used public data sets (Jiang et al. 2018, Mital et al. 2011, Wang et al. 2018). Most studies evaluate video saliency models frame by frame using the same saliency metrics as in image saliency, such as AUC, NSS, and CC. The evaluation is subject to the same problem of incongruent metrics as discussed for the case of static image saliency, and the problem can be solved in the same way by computing metric-specific saliency maps from the predicted densities or by directly going for information-theoretic model evaluation.

If the goal is to predict gaze at a certain time interval in the future, i.e., $p(x_{t+\delta} | I_{\leq t+\delta}, x_{\leq t})$ (Xu et al. 2018, Zhang et al. 2017), then the setting is identical to image scanpath prediction, with the only difference being that we condition on the previous video instead of the image. However, predicting when the next saccade happens can be more relevant than in the static image case. While predictions of both time and target of the next saccade can be evaluated in the same way, using log-likelihood, as when predicting only the next fixation, the resulting likelihoods are not comparable with purely spatial fixation prediction: $p(x_i, t_i | x_{i-1}, t_{i-1}, \ldots, x_0, t_0, I_{\leq t})$ predicts different data than, e.g., $p(x_i | x_{i-1}, t_{i-1}, \ldots, x_0, t_0, I_{\leq t})$. If comparability is relevant, and predicting saccade

timings is not crucial, then the easiest solution is to inform the model about the time of the next saccade, i.e., predict $p(x_i, |t_i, x_{i-1}, t_{i-1}, \ldots, x_0, t_0, I_{\leq t})$.

3.5. The MIT/Tuebingen Saliency Benchmark: The Next Generation of Fixation Prediction Benchmarking

We teamed up with the organizers of the previous MIT Saliency Benchmark to incorporate the probabilistic evaluation framework and further improve the functionality, which led to a relaunch in 2019 as the MIT/Tuebingen Saliency Benchmark (**https://saliency.tuebingen.ai**). The Benchmark now provides the option to submit fixation density predictions instead of saliency map predictions. In this case, we compute the appropriate saliency maps for all evaluated metrics and thus can evaluate the model fairly in all metrics. We also include previously published classic saliency map models in the same leaderboard, and researchers can still submit their models as saliency maps. However, as discussed above, a single saliency map cannot perform well on all metrics. This makes it worthwhile for researchers to switch to probabilistic modeling, and indeed, today, all three top-scoring models are probabilistic models.

In addition, we are planning to add a track for the evaluation of scanpath models on the same MIT300 and CAT2000 data sets to facilitate comparability to the benchmarking of spatial fixation prediction, and we are collecting a new data set that is more tailored toward scoring scanpath models. Finally, we are also planning to add tracks for tasks beyond free viewing, such as object search (Chen et al. 2021).

4. EVOLUTION AND KEY INGREDIENTS OF HIGH-PERFORMING SALIENCY MODELS

Since the first computational saliency model of Itti et al. (1998), a growing list of computational models for fixation prediction have been proposed. The oldest saliency models are usually inspired by the classic idea of feature popout. They use hand-crafted low-level features for computing saliency maps and either operate purely locally (Itti et al. 1998, Zhang et al. 2008) or, more often, add some global features or statistics (e.g., Bruce & Tsotsos 2009, Harel et al. 2007, Riche et al. 2013b, Torralba et al. 2006). Importantly, these models initially were designed without any explicit parameter fitting on fixation data sets, so that the exact shape of the saliency function had to be specified exclusively based on the inventor's reasoning. Alternatively, quite early on, researchers also applied machine learning to optimize relevant features from fixation data (e.g., Baddeley & Tatler 2006, Judd et al. 2009, Kienzle et al. 2009, Zhao & Koch 2011). Typically, data sets are used consisting of fixations from multiple subjects free viewing images for a few seconds (Borji & Itti 2015, Judd et al. 2009) (see also Section 3.1). More recently, due to the high cost and effort required to collect high-quality eye movement data and the high data demand of many modern machine learning methods, less precise webcam-based fixation data sets (Xu et al. 2015) or mouse-tracking-based pseudofixation data sets (Jiang et al. 2015) have also been used for training saliency models.

4.1. Key Ingredients of High-Performing Saliency Models

Until 2015, the best existing saliency models were able to explain approximately one-third of the explainable information in the spatial fixation structure (Kümmerer et al. 2015b), suggesting that there was still substantial room for improvement. We find that transfer learning, probabilistic modeling, and the use of good readout networks have been the key ingredients to achieving high benchmarking performance, as we roughly explain below.

4.1.1. Transfer learning. The first application of deep learning in saliency modeling was the eDN (Ensembles of Deep Learning) model (Vig et al. 2014), which used comparatively small

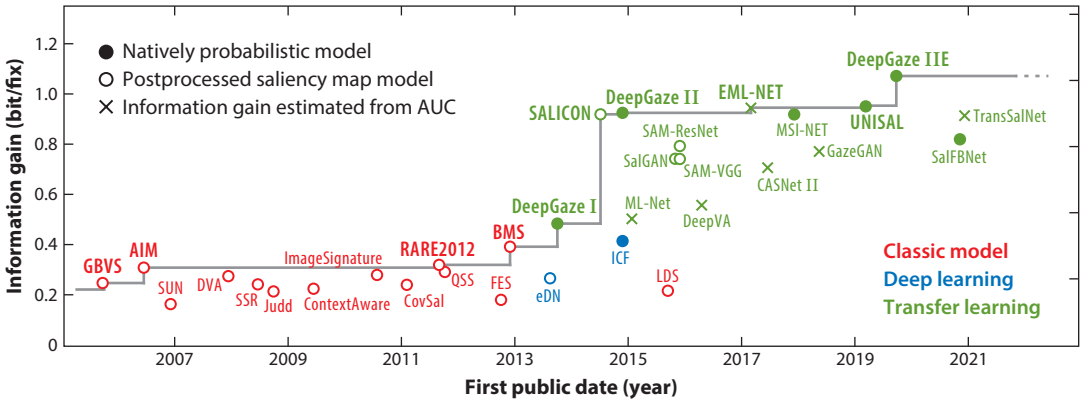

Figure 6

Progress in saliency prediction over the past 15 years in terms of explained information on the MIT300 data set of the MIT/Tuebingen Saliency Benchmark. Models that raised the state-of-the-art performances, as indicated by the solid line, are highlighted in bold. For details, readers are referred to the **Supplemental Appendix, section 1.5** and **Supplemental Figure 3**.

neural networks with three layers and achieved only a small improvement over previous models (see **Figure 6**): The data sets typically used for training models of fixation prediction (e.g., Borji & Itti 2015, Judd et al. 2009) are too small to train large neural networks from scratch.

This problem can be addressed by transfer learning: Features learned by deep neural networks on one task often transfer surprisingly well to other, loosely related tasks (Donahue et al. 2014). Kümmerer et al. (2015a) introduced transfer learning to fixation prediction with DeepGaze I by training a linear readout on top of the last convolutional layer of the object recognition model AlexNet (Krizhevsky et al. 2012). Since DeepGaze I, all high-performing saliency models have used transfer learning in various architectures, including multiscale architectures (Huang et al. 2015), generative adversarial networks (Pan et al. 2017), and multimodal training (Droste et al. 2020). In DeepGaze IIE, Linardos et al. (2021) used an ensembling approach to combine multiple models with different pretrained neural networks and have set the state of the art since 2020.

4.1.2. Probabilistic modeling. By formulating saliency models as probabilistic models of fixation density prediction, we get access to a principled and powerful loss function (log-likelihood). Training for log-likelihood encourages the model to extract and incorporate all available information from the ground-truth fixation data when approximating the underlying ground-truth fixation density. This allows the models to perform well in all sorts of possible downstream tasks, including evaluation on other metrics using Bayesian decision theory (see above).

4.1.3. Good readout networks. Although pretrained deep features transfer well enough to saliency prediction to train saliency models with the comparatively small data sets available, it is still important to regularize the model sufficiently to avoid overfitting. Using a small readout network that consists of only a few layers of, importantly, 1×1 convolutions seems to be a promising method for this regularization. The readout network is essentially a nonlinear function that is applied pixel by pixel to a vector of input features. Therefore, it is able to learn the best nonlinear transformation adjusting the scale of the input features and make use of interactions among those features. However, since the nonlinear function is applied to each pixel individually, readout networks cannot learn new spatial features. Given enough capacity, they can make use of all of the pixelwise mutual information between input features and predicted quantities such as fixation density and therefore provide a way to compare this mutual information between different sets of features.

4.2. Gold Standard and Explainable Information Explained

For assessing the state of a field, it is important not only to know the current state of the art, but also to develop (*a*) a relevant baseline and (*b*) a gold standard to quantify to what fraction the prediction problem has been solved (i.e., whether the best models are actually good). For the baseline, we use the best possible model of the prior distribution $p(x)$ (i.e., the center bias), which does not depend on the content of a particular image. For the gold standard, we use a nonparametric estimate of the fixation density for each individual image with leave-one-subject-out cross-validation (see the **Supplemental Appendix, section 1.4**). These two models essentially define 0% and 100% on the scale of how well the task is solved.

Supplemental Material >

To judge the degree to which the problem of fixation prediction is solved, we use the information gain scale (see Section 3.2), which measures information in bits such that differences at different locations on the scale are comparable and have the same meaning.

With adequate baseline models and a good scale, we can introduce new quantities that help us measure model performance and the state of the field in an even more intuitive way: The explainable information gain (IG) is the information gain of the gold standard model relative to the center bias model, which we write as $\mathrm{IG}(p_{\mathrm{gold}} \| p_{\mathrm{centerbias}})$ below. It quantifies how much of the information in the spatial fixation structure in the data set we can hope to explain. Using the explainable information gain, we can then define the explainable information gain explained (IGE) for a model $\hat{p}$: $\mathrm{IGE}(\hat{p}) = \frac{\mathrm{IG}(\hat{p} \| p_{\mathrm{centerbias}})}{\mathrm{IG}(p_{\mathrm{gold}} \| p_{\mathrm{centerbias}})}$. The IGE metric quantifies how much of the achievable information gain a certain model explains and therefore, on a scale from 0% to 100%, how well the model solves the task of predicting spatial fixations. Information gain and IGE can be viewed as information-theoretic versions of the variance explained and explainable variance explained values used in linear Gaussian regression.

The IGE of the best model can be interpreted as the percentage at which the field has solved the task of predicting fixations—at least on the data set at hand. For the MIT300 data set of the MIT/Tuebingen Saliency Benchmark, the best-performing model, DeepGaze IIE, explains more than 80% of the explainable information gain, compared to 30% for the best nontransfer model (ICF), demonstrating that transfer learning close to tripled the explained information (see **Figure 6**). This demonstrates the large progress the field has made over the past years but also shows that approximately 20% of the explainable information has still not been correctly predicted to date.

While most computational models have addressed the problem of spatial saliency, a growing number of models predict whole scanpaths of fixations on images (e.g., Adeli et al. 2017, Boccignone & Ferraro 2004, Engbert et al. 2015, Kümmerer et al. 2022, Le Meur & Liu 2015; for a review, see Kümmerer & Bethge 2021), video saliency (Bazzani et al. 2017, Tangemann et al. 2020), or even video scanpaths (Xu et al. 2018, Zhang et al. 2017). For all of these prediction tasks, deep learning has resulted in substantial prediction improvements.

In the case of scanpath prediction, estimating the explainable information gain is much harder than in the case of image saliency. Because the underlying scanpath distribution is much more high dimensional than in the case of spatial fixation prediction, it cannot easily be estimated by, e.g., a Gaussian kernel density estimate over the ground-truth data. However, if we want to know how much specific mechanisms contribute to scanpath prediction, then we can leverage high-performing black box models as useful lower bounds on the explainable information gain. In this way, we can obtain upper bounds on the fraction of explainable information explained for the considered mechanistic model. Kümmerer et al. (2022) used their scanpath model, DeepGaze III, which currently sets the state of the art in free-viewing scanpath prediction on images, to obtain such upper bounds. They found that, to date, all existing mechanistic scanpath models explain at most 28% of the explainable information in the dependencies between fixations in a scanpath.

4.3. Beyond Prediction, Toward Understanding

Since researchers began applying deep learning to fixation prediction, they have developed models that predict fixations well, e.g., explaining up to 80% of the explainable information in static image saliency. However, these models do not directly offer a concise description of the necessary mechanisms and do not seem to adhere to the principle of Occam's razor. Given that vision scientists want to understand how the visual system works, how can these models actually help improve our understanding? The probabilistic modeling framework, together with the universal scale of explainable information explained, offers a great opportunity to reliably evaluate how much the inclusion or exclusion of different factors or mechanisms actually changes the behavior of a model.

As a first example, Kümmerer et al. (2017) used this framework to disentangle quantitatively how much impact low-level versus high-level features have on the shape of fixation densities. Using a standardized architecture with highly constrained readout networks on top of the tested features, they could quantify how much information the tested features contain about fixation locations. By using fixed sets of intensity-contrast features, they constructed a low-level ICF model that performs better than all classic saliency models (**Figure 6**). By relating its performance to the gold standard, however, it becomes apparent that the ICF model explains only approximately one-third of the explainable information. By inspecting images where the ICF model leaves most explainable information unexplained, we can generate good hypotheses about what other features could be particularly important. Such features seem to be most prominently faces and text, providing strong evidence that high-level features have strong influence on fixation densities. As demonstrated by the approach of DeepGaze IIE (Linardos et al. 2021), the more recent improvements based on transfer learning are driven by the development of better ImageNet features that are closely linked to high-level semantic discrimination.

As a second example, scoring the relevance of different factors, Kümmerer et al. (2022) made use of the universal function approximation property of neural networks to quantify how strongly different earlier fixations and scene content influence fixation placement without having to know the precise mechanisms of their influence; this method also ruled out certain ways in which scene content and earlier fixations might interact in the fixation selection process.

A third instructive example, showing how the comparative evaluation of prediction performance can yield insights into the underlying mechanisms, is the comparison of video saliency models with static saliency models. Tangemann et al. (2020) compared models that explicitly included spatiotemporal video saliency mechanisms to a purely spatial saliency prediction of the DeepGaze MR model. They found that DeepGaze MR slightly outperformed the more complex spatiotemporal video saliency models and exhibited very similar behavior, as illustrated in **Figure 7*a***. In fact, all models can roughly reproduce the gold standard model most of the time, with the exception of some periods in which all models fail quite strikingly. The video segments during these periods indeed turned out to show clear salient temporal patterns such as suddenly appearing objects and movements.

The systematic inspection of test data when models fail and whether they agree or disagree is particularly useful for the design of more effective testing of mechanisms. For the example of video saliency, it is obvious that we need to focus on data sets where static saliency models fail (Tangemann et al. 2020). Similarly, Kümmerer & Bethge (2021) selected fixations for which a large range of scanpath models differed most strongly in their prediction and classified them with respect to the relevant mechanisms (see **Figure 7*b***). Using such controversial stimuli (Golan et al. 2020) is a highly informative strategy for model comparison and thus very useful for experimental design.

From a practical point of view, it would be desirable to be able to use saliency models directly for optimizing the saliency of a visual scene, instead of just for testing a given example. The straightforward way to change the saliency of an image is to use, e.g., gradient descent on an

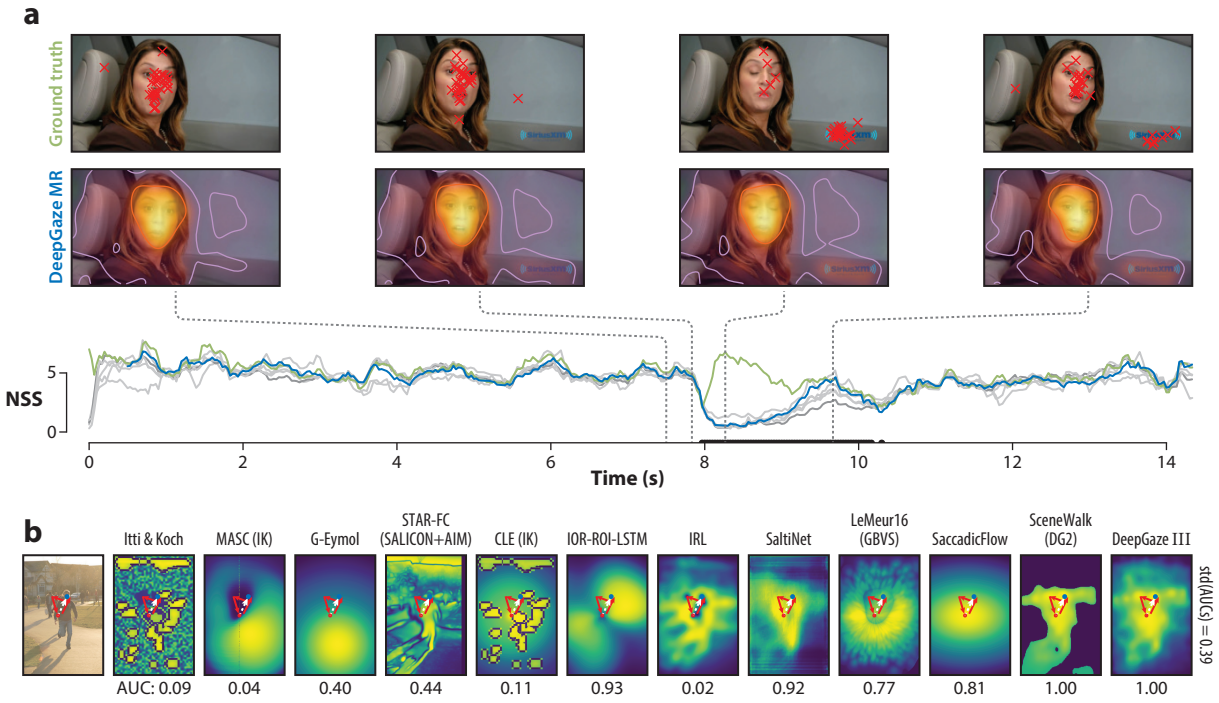

Figure 7

Counterfactual model testing. (*a*) Comparing a video saliency model, which by design cannot make use of temporal patterns (*blue* NSS line), to interobserver consistency (*green* NSS line) reveals video segments where gaze is driven by temporal patterns (*black horizontal lines in NSS plot*). All other video saliency models also fail to predict gaze in these segments, although they have access to temporal video features (other NSS lines). Ground truth data and DeepGaze MR predictions are shown for four example frames; their time points are marked with vertical dashed lines in the NSS plot. For more details, readers are referred to **Supplemental Figure 5**. Panel adapted with permission from Tangemann et al. (2020). (*b*) Selecting fixations in the MIT1003 data set for which a broad list of scanpath models differ strongly in their AUC score reveals important mechanisms; in this case, return saccades are predicted badly by models with slowly decaying inhibition of return. For more examples, readers are referred to **Supplemental Figure 4** and Kümmerer et al. (2018).

image until the predicted fixation density matches our desired outcome. This can work well for low-level saliency models (for a review of such approaches, see Mateescu & Bajic 2016). However, with state-of-the-art deep learning–based models, we run unavoidably into the problem of adversarial examples (Goodfellow et al. 2014): Tiny, imperceptible changes to the input are enough to result in arbitrary model outputs. For simple data sets like MNIST (Lecun et al. 1998), there are now adversarially robust deep neural networks (e.g., Schott et al. 2019) for which aspired output changes also lead to the desired perceptual changes, but in the case of large-scale natural image models, the problem is still very challenging (Croce et al. 2021).

Supplemental Material >

There are different ways to work around the problem of adversarial examples. Gatys et al. (2017) employed techniques from generative adversarial networks, texture synthesis, and style transfer to avoid adversarial examples when changing the fixation density of an image predicted by the DeepGaze II saliency model. Since then, multiple saliency manipulation methods using deep learning have been proposed, usually relying on adversarial approaches (e.g., Chen et al. 2019, Jiang et al. 2021); however, these methods are rarely confirmed by collecting fixation data on the generated images (for a notable exception, see Mejjati et al. 2020).

5. CONCLUSION AND OUTLOOK

In this article, we review recent advances in visual fixation prediction within a unifying probabilistic modeling framework for consistent evaluation of how well models predict eye movements and how to judge the contribution of different mechanisms. We show how the large variety of saliency maps and scanpath models can be translated into this unifying framework. Using explainable information explained, we evaluate how different factors contributed to the development of the state of the art across different settings such as static and video saliency and scanpath prediction. We also show how explainable information explained can be used to select the most informative examples for model comparison. Altogether, this opens a new door for vision science: While vision science has traditionally focused on improving the predictive power of interpretable models, we can now focus on improving the interpretability of well-predicting models.

Despite substantial progress over the past years, visual fixation prediction is far from being solved. Even on the current free-viewing benchmarks, with all their limitations, 20% of explainable information is still left unexplained. Part of this unexplained information manifests itself in systematic errors in all high-performing models. From the 10 classes of commonly missed gaze locations found by Bylinskii et al. (2016), only "animals" and "persons" now seem to be solved, whereas the majority still seems to be problematic for modern models (**Figure 8**).

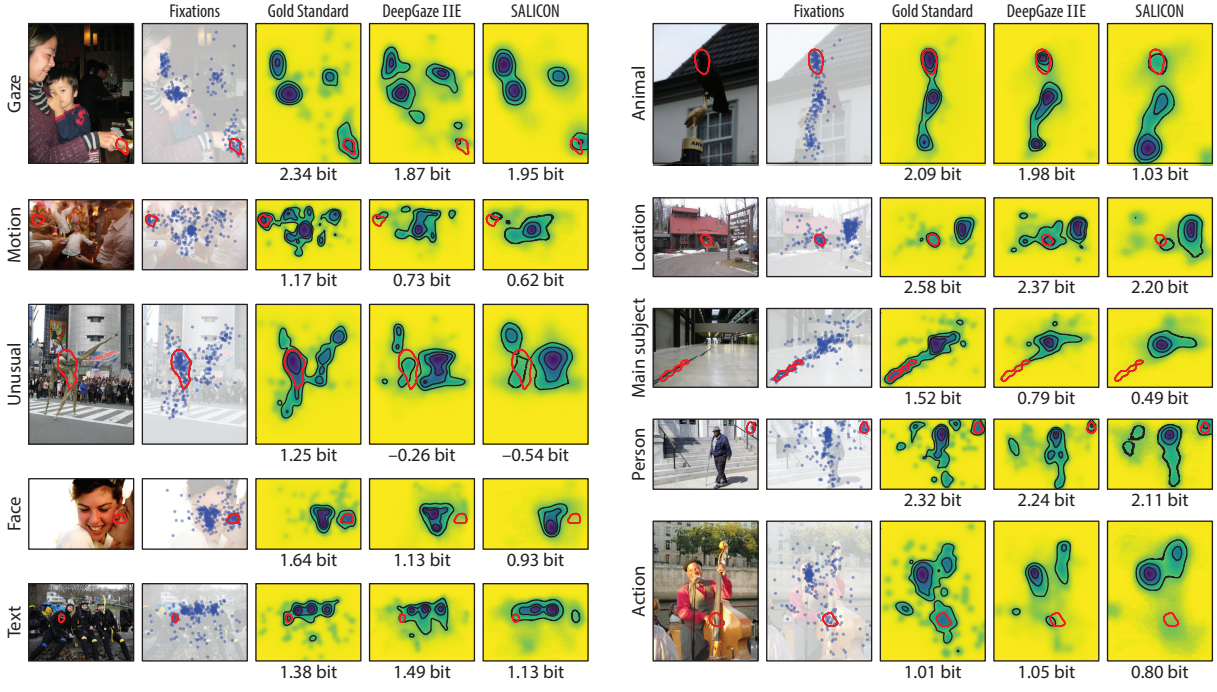

Figure 8

Images from the MIT300 benchmark data set, for which Bylinskii et al. (2016) found systematic errors (*red contours*) in the state-of-the-art models of 2016 (represented by the SALICON model). We revisit those problematic cases with the current state-of-the-art model, DeepGaze IIE. Some problematic cases look better now ("animals," "locations," "persons"), but most cases still seem mostly unsolved (note that, e.g., in "action," DeepGaze IIE sees the artist's hand, not the strings that the artist just plucked; a similar problem occurs for "gaze"). For model prediction, we display prediction performance as information gain in bits. For an extended version of the figure including more current models, readers are referred to **Supplemental Figure 6**.

Supplemental Material >

Fixing these errors likely requires models to have additional understanding of images, e.g., gaze targets of displayed people, areas of potential motion, action or interaction, and semantic relations (Pedziwiatr et al. 2021). Most of the 20% of information that is unexplained, however, is broadly distributed over virtually all images. Thus, the biggest challenge might be to further optimize the complex high-dimensional interplay between existing features, rather than to add missing ones.

In the future, the biggest opportunity for progress will come from large and very diverse data sets. While DeepGaze IIE has exhibited exceptional out-of-domain performance on several data sets, such as Bruce & Tsotsos (2009) and Li et al. (2014), it is clear that models trained on natural images can considerably struggle, e.g., on purely low-level images (Berga et al. 2019). New data sets that allow us to build and test models in settings that were previously unavailable (e.g., Chen et al. 2021) or where collecting eye movement itself was technically challenging (Kothari et al. 2020, Matthis et al. 2018) will open up exciting new opportunities.

SUMMARY POINTS

1. Inconsistent evaluation methods can severely impair the comparability among different fixation prediction models.
2. Formulating models as probability density models facilitates the use of information gain explained as a universal and intuitive measure, with which we can directly compare models and individual mechanisms across different prediction settings.
3. Prediction performance has made substantial progress, especially since the introduction of deep learning, but there are still qualitative deficiencies that are similarly evident in all current high-performing models.
4. Using differential comparisons (e.g., by including and excluding specific mechanisms of interest), benchmarking and deep learning can be powerful tools to support our understanding and identification of the most relevant mechanisms, e.g., how low-level and high-level image content interact in fixation selection or in which precise ways earlier fixations affect later fixations.

FUTURE ISSUES

1. The framework of systematic model comparisons presented in this review sets the stage for much more comprehensive testing of different mechanisms.
2. This framework, together with larger and more realistic data sets allowing us to probe much more extensively how well models predict fixation behavior in the real world, creates great opportunities for fast progress in the field, potentially including tasks changing over time and observers learning and adapting.
3. The two main future research directions are (*a*) decomposing current state-of-the-art models into more interpretable and mechanistic models and (*b*) building models that overcome the systematic errors of current state-of-the-art models, e.g., failure to predict the attraction of gaze by interactions, semantic inconsistencies, and uncommon objects and gaze targets, and that perform well over a wide range of different stimuli.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by German Federal Ministry of Education and Research (BMBF) grant FKZ 01IS18039A to Tübingen AI Center and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Germany's Excellence Strategy grants EXC 2064/1, 390727645, and SFB 1233 (Robust Vision: Inference, Principles and Neural Mechanisms), project number 276693517.

LITERATURE CITED

Adeli H, Vitu F, Zelinsky GJ. 2017. A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *J. Neurosci.* 37(6):1453–67

Anderson NC, Anderson F, Kingstone A, Bischof WF. 2015. A comparison of scanpath comparison methods. *Behav. Res. Methods* 47(4):1377–92

Baddeley RJ, Tatler BW. 2006. High frequency edges (but not contrast) predict where we fixate: a Bayesian system identification analysis. *Vis. Res.* 46(18):2824–33

Barthelmé S, Trukenbrod H, Engbert R, Wichmann F. 2013. Modeling fixation locations using spatial point processes. *J. Vis.* 13(12):1

Bazzani L, Larochelle H, Torresani L. 2017. Recurrent mixture density network for spatiotemporal visual attention. In *ICLR 2017: International Conference on Learning Representations*. N.p.: ICLR

Berga D, Vidal XRF, Otazu X, Pardo XM. 2019. SID4VAM: a benchmark dataset with synthetic images for visual attention modeling. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8788–97. Piscataway, NJ: IEEE

Boccignone G, Ferraro M. 2004. Modelling gaze shift as a constrained random walk. *Phys. A* 331(1):207–18

Borji A, Itti L. 2015. CAT2000: a large scale fixation dataset for boosting saliency research. arXiv:1505.03581 [cs.CV]

Borji A, Sihite DN, Itti L. 2013a. Objects do not predict fixations better than early saliency: a re-analysis of Einhäuser et al.'s data. *J. Vis.* 13(10):18

Borji A, Sihite DN, Itti L. 2013b. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Trans. Image Proc.* 22(1):55–69

Box GEP. 1976. Science and statistics. *J. Am. Stat. Assoc.* 71(356):791–99

Box GEP. 1979. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, ed. RL Launer, GN Wilkinson, pp. 201–36. Cambridge, MA: Academic Press

Bruce NDB, Tsotsos JK. 2009. Saliency, attention, and visual search: an information theoretic approach. *J. Vis.* 9(3):5

Buswell GT. 1935. *How People Look at Pictures*. Chicago: Univ. Chicago Press

Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F. 2018. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intel.* 41(3):740–57

Bylinskii Z, Recasens A, Borji A, Oliva A, Torralba A, Durand F. 2016. Where should saliency models look next? In *Computer Vision—ECCV 2016: Proceedings of the 14th European Conference, Amsterdam, The Netherlands, October 11–14*, pp. 809–24. Berlin: Springer

Chen Y, Yang Z, Ahn S, Samaras D, Hoai M, Zelinsky G. 2021. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Sci. Rep.* 11:8776

Chen YC, Chang KJ, Tsai YH, Wang YCF, Chiu WC. 2019. Guide your eyes: learning image manipulation under saliency guidance. In *Proceedings of the British Machine Vision Conference (BMVC)*, ed. K Sidorov, Y Hicks, pp. 24.1–12. Durham, UK: BMVA Press

Cover TM, Thomas JA. 2005. Rate distortion theory. In *Elements of Information Theory*, pp. 301–46. New York: Wiley

Croce F, Andriushchenko M, Sehwag V, Debenedetti E, Flammarion N, et al. 2021. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. N.p.: NeurIPS

Didday RL, Arbib MA. 1975. Eye movements and visual perception: a "two visual system" model. *Int. J. Man-Mach. Stud.* 7(4):547–69

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, et al. 2014. DeCAF: a deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning*, pp. 647–55. N.p.: ICML

Droste R, Jiao J, Noble JA. 2020. Unified image and video saliency modeling. In *Computer Vision—ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, August 23–28, 2020*, ed. A Vedaldi, H Bischof, T Brox, JM Frahm, pp. 419–35. Berlin: Springer

Einhäuser W. 2013. Objects and saliency: reply to Borji et al. *J. Vis.* 13(10):20

Einhäuser W, König P. 2010. Getting real—sensory processing of natural stimuli. *Curr. Opin. Neurobiol.* 20(3):389–95

Einhäuser W, Spain M, Perona P. 2008. Objects predict fixations better than early saliency. *J. Vis.* 8(14):18

Engbert R, Trukenbrod HA, Barthelmé S, Wichmann FA. 2015. Spatial statistics and attentional dynamics in scene viewing. *J. Vis.* 15(1):14

Gatys LA, Kümmerer M, Wallis TSA, Bethge M. 2017. Guiding human gaze with convolutional neural networks. arXiv:1712.06492 [cs.CV]

Golan T, Raju PC, Kriegeskorte N. 2020. Controversial stimuli: pitting neural networks against each other as models of human cognition. *PNAS* 117(47):29330–37

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 2672–80. Red Hook, NY: Curran Assoc.

Harel J, Koch C, Perona P. 2007. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, ed. B Schölkopf, JC Platt, T Hoffman, pp. 545–52. Cambridge, MA: MIT Press

Hayhoe M. 2000. Vision using routines: a functional account of vision. *Vis. Cogn.* 7(1–3):43–64

Henderson JM. 2003. Human gaze control during real-world scene perception. *Trends Cogn. Sci.* 7(11):498–504

Huang X, Shen C, Boix X, Zhao Q. 2015. SALICON: reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 262–70. Piscataway, NJ: IEEE

Itti L, Baldi P. 2009. Bayesian surprise attracts human attention. *Vis. Res.* 49(10):1295–306

Itti L, Koch C. 2001a. Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2(3):194–203

Itti L, Koch C. 2001. Feature combination strategies for saliency-based visual attention systems. *J. Electron. Imag.* 10(1):161–70

Itti L, Koch C, Niebur E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intel.* 20(11):1254–59

Jiang L, Xu M, Liu T, Qiao M, Wang Z. 2018. DeepVS: a deep learning based video saliency prediction approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 602–17. Berlin: Springer

Jiang L, Xu M, Wang X, Sigal L. 2021. Saliency-guided image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16509–18. Piscataway, NJ: IEEE

Jiang M, Huang S, Duan J, Zhao Q. 2015. SALICON: saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1072–80. Piscataway, NJ: IEEE

Judd T, Durand F, Torralba A. 2012. *A benchmark of computational models of saliency to predict human fixations*. Tech. Rep., MIT, Cambridge, MA

Judd T, Ehinger K, Durand F, Torralba A. 2009. Learning to predict where humans look. In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 2106–13. Piscataway, NJ: IEEE

Kienzle W, Franz MO, Schölkopf B, Wichmann FA. 2009. Center-surround patterns emerge as optimal predictors for human saccade targets. *J. Vis.* 9(5):7

Koch C, Ullman S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4:219–27

Kothari R, Yang Z, Kanan C, Bailey R, Pelz JB, Diaz GJ. 2020. Gaze-in-wild: a dataset for studying eye and head coordination in everyday activities. *Sci. Rep.* 10:2539

Kowler E. 2011. Eye movements: the past 25 years. *Vis. Res.* 51(13):1457–83

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, ed. F Pereira, CJC Burges, L Bottou, KQ Weinberger, pp. 1097–105. Red Hook, NJ: Curran Assoc.

Kümmerer M, Bethge M. 2021. State-of-the-art in human scanpath prediction. arXiv:2102.12239 [cs.CV]

Kümmerer M, Bethge M, Wallis TSA. 2022. DeepGaze III: modeling free-viewing human scanpaths with deep learning. *J. Vis.* 22(5):7

Kümmerer M, Theis L, Bethge M. 2015a. Deep Gaze I: boosting saliency prediction with feature maps trained on ImageNet. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR) 2015, San Diego, CA, USA, May 7-9*. N.p.: ICLR

Kümmerer M, Wallis TSA, Bethge M. 2015b. Information-theoretic model comparison unifies saliency metrics. *PNAS* 112(52):16054–59

Kümmerer M, Wallis TSA, Bethge M. 2018. Saliency benchmarking made easy: separating models, maps and metrics, In *Computer Vision—ECCV 2018: Proceedings of the 15th European Conference, Munich, Germany, September 8–14, 2018*, ed. V Ferrari, M Hebert, C Sminchisescu, Y Weiss, pp. 798–814. Berlin: Springer

Kümmerer M, Wallis TSA, Gatys LA, Bethge M. 2017. Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4789–98. Piscataway, NJ: IEEE

Land M, Mennie N, Rusted J. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* 28(11):1311–28

Le Meur O, Liu Z. 2015. Saccadic model of eye movements for free-viewing condition. *Vis. Res.* 116:152–64

Lecun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11):2278–324

Li Y, Hou X, Koch C, Rehg JM, Yuille AL. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–87. Piscataway, NJ: IEEE

Li Z. 2002. A saliency map in primary visual cortex. *Trends Cogn. Sci.* 6(1):9–16

Linardos A, Kümmerer M, Press O, Bethge M. 2021. DeepGaze IIE: calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12919–28. Piscataway, NJ: IEEE

Malem-Shinitski N, Opper M, Reich S, Schwetlick L, Seelig SA, Engbert R. 2020. A mathematical model of local and global attention in natural scene viewing. *PLOS Comput. Biol.* 16(12):e1007880

Mateescu VA, Bajic IV. 2016. Visual attention retargeting. *IEEE MultiMedia* 23(1):82–91

Matthis JS, Yates JL, Hayhoe MM. 2018. Gaze and the control of foot placement when walking in natural terrain. *Curr. Biol.* 28(8):1224–33.e5

Mejjati YA, Gomez CF, Kim KI, Shechtman E, Bylinskii Z. 2020. Look here! A parametric learning based approach to redirect visual attention. In *Computer Vision—ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, August 23–28, 2020*, ed. A Vedaldi, H Bischof, T Brox, JM Frahm, pp. 343–61. Berlin: Springer

Mital PK, Smith TJ, Hill RL, Henderson JM. 2011. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cogn. Comput.* 3(1):5–24

Pan J, Ferrer CC, McGuinness K, O'Connor NE, Torres J, et al. 2017. SalGAN: visual saliency prediction with generative adversarial networks. arXiv:1701.01081 [cs.CV]

Parkhurst D, Law K, Niebur E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vis. Res.* 42(1):107–23

Pedziwiatr MA, Kümmerer M, Wallis TSA, Bethge M, Teufel C. 2021. Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition* 206:104465

Riche N, Duvinage M, Mancas M, Gosselin B, Dutoit T. 2013a. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*, pp. 1153–60. Piscataway, NJ: IEEE

Riche N, Mancas M, Duvinage M, Mibulumukini M, Gosselin B, Dutoit T. 2013b. RARE2012: a multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Proc. Image Commun.* 28(6):642–58

Rothkopf CA, Ballard DH, Hayhoe MM. 2007. Task and context determine where you look. *J. Vis.* 7(14):16

Schott L, Rauber J, Bethge M, Brendel W. 2019. Towards the first adversarially robust neural network model on MNIST. In *ICLR 2019: International Conference on Learning Representations*. N.p.: ICLR

Schütt HH, Rothkegel LOM, Trukenbrod HA, Reich S, Wichmann FA, Engbert R. 2017. Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychol. Rev.* 124(4):505–24

Schwetlick L, Rothkegel LOM, Trukenbrod HA, Engbert R. 2020. Modeling the effects of perisaccadic attention on gaze statistics during scene viewing. *Commun. Biol.* 3(1):727

Shannon CE. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27(1):379–423, 623–56

Shannon CE, Weaver W. 1949. *A Mathematical Theory of Communication*. Urbana: Univ. Ill. Press. 1st ed.

Tangemann M, Kümmerer M, Wallis TSA, Bethge M. 2020. Measuring the importance of temporal features in video saliency. In *Computer Vision—ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, August 23-28, 2020*, pp. 667–84. Berlin: Springer

Tatler BW, Baddeley RJ, Gilchrist ID. 2005. Visual correlates of fixation selection: effects of scale and time. *Vis. Res.* 45(5):643–59

Torralba A, Oliva A, Castelhano MS, Henderson JM. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. Rev.* 113(4):766–86

Treisman AM, Gelade G. 1980. A feature-integration theory of attention. *Cogn. Psychol.* 12(1):97–136

Vig E, Dorr M, Cox D. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2798–805. Piscataway, NJ: IEEE

Wade NJ. 2010. Pioneers of eye movement research. *i-Perception* 1(2):33–68

Wagstaff K. 2012. Machine learning that matters. In *Proceedings of the 29th International Conference on Machine Learning*, ed. J Langfort, J Pineau, pp. 1851–56. Norristown, PA: Omnipress

Wang W, Shen J, Guo F, Cheng MM, Borji A. 2018. Revisiting video saliency: a large-scale benchmark and a new model. arXiv:1801.07424 [cs.CV]

Wilming N, Betz T, Kietzmann TC, König P. 2011. Measures and limits of models of fixation selection. *PLOS ONE* 6(9):e24038

Xu P, Ehinger KA, Zhang Y, Finkelstein A, Kulkarni SR, Xiao J. 2015. TurkerGaze: crowdsourcing saliency with webcam based eye tracking. arXiv:1504.06755 [cs.CV]

Xu Y, Dong Y, Wu J, Sun Z, Shi Z, et al. 2018. Gaze prediction in dynamic 360° immersive videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5333–42. Piscataway, NJ: IEEE

Yarbus AL. 1967. *Eye Movements and Vision*. New York: Plenum Press

Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW. 2008. SUN: a Bayesian framework for saliency using natural statistics. *J. Vis.* 8(7):32

Zhang M, Ma KT, Lim JH, Zhao Q, Feng J. 2017. Deep Future Gaze: gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4372–81. Piscataway, NJ: IEEE

Zhao Q, Koch C. 2011. Learning a saliency map using fixated locations in natural scenes. *J. Vis.* 11(3):9