Objects guide human gaze behavior in dynamic real-world scenes

Nicolas Roth[1, 2,*], **Martin Rolfs**[1, 3, 4], **Olaf Hellwich**[1, 5], **and Klaus Obermayer**[1, 2, 4]

[1]Cluster of Excellence Science of Intelligence, Technische Universität Berlin, Marchstraße 23, 10587 Berlin, Germany
[2]Institut für Softwaretechnik und Theoretische Informatik, Technische Universität Berlin, Marchstraße 23, 10587 Berlin, Germany
[5]Institut für Technische Informatik und Mikroelektronik, Technische Universität Berlin, Marchstraße 23, 10587 Berlin, Germany
[3]Institut für Psychologie, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany
[4]Bernstein Center for Computational Neuroscience Berlin, Philippstraße 13, 10115 Berlin, Germany
[*]Correspondence to roth@tu-berlin.de

ABSTRACT

The complexity of natural scenes makes it challenging to experimentally study the mechanisms behind human gaze behavior when viewing dynamic environments. Historically, eye movements were believed to be driven primarily by bottom-up saliency, but increasing evidence suggests that objects also play a significant role in guiding attention. We present a new computational framework to investigate the importance of objects for attentional guidance. This framework is designed to simulate realistic scanpaths for dynamic real-world scenes, including saccade timing and smooth pursuit behavior. Individual model components are based on psychophysically uncovered mechanisms of visual attention and saccadic decision-making. All mechanisms are implemented in a modular fashion with a small number of well-interpretable parameters. To systematically analyze the importance of objects in guiding gaze behavior, we implemented four different models within this framework: two purely location-based models, where one is based on low-level saliency and one on high-level saliency, and two object-based models, with one incorporating low-level saliency for each object and the other one not using any saliency information. We optimized each model's parameters to reproduce the saccade amplitude and fixation duration distributions of human scanpaths using evolutionary algorithms. We compared model performance with respect to spatial and temporal fixation behavior, including the proportion of fixations exploring the background, as well as detecting, inspecting, and revisiting objects. A model with object-based attention and inhibition, which uses saliency information to prioritize between objects for saccadic selection, leads to scanpath statistics with the highest similarity to the human data. This demonstrates that scanpath models benefit from object-based attention and selection, suggesting that object-level attentional units play an important role in guiding attentional processing.

Author summary

There has long been an interest in understanding how we decide when and where to move our eyes, and psychophysical experiments have uncovered many underlying mechanisms. Under controlled laboratory conditions, objects in the scene play an important role in guiding our attention. Due to the visual complexity of the world around us, however, it is hard to assess experimentally how objects influence eye movements when observing dynamic real-world scenes. Computational models have proved to be a powerful tool for investigating visual attention, but existing models are either only applicable to images or restricted to predicting where humans look on average. Here, we present a computational framework for simulating where and when humans decide to move their eyes when observing dynamic real-world scenes. Using our framework, we can assess the influence of objects on the model predictions. We find that including object-based attention in the modeling increases the resemblance of simulated eye movements to human gaze behavior, showing that objects play indeed an important role in guiding our gaze when exploring the world around us. We hope that the availability of this framework encourages more research on attention in dynamic real-world scenes.

Introduction

Humans actively move their eyes about 2-3 times per second when visually exploring their environment (Buswell, 1935; Yarbus, 1967). This moves the part of the retina with the highest visual acuity – and allocates significant computational capacities in the brain – to parts of the scene that are of the highest interest at a given time. Hence, the decision of where and when to move our eyes determines which parts of the world around us we inspect in more detail and, conversely, which parts we might miss. Moving our eyes is therefore always a trade-off between different potential target positions that fundamentally shapes the way we perceive the world. The sequence of eye movements for a given visual scene is called a scanpath and reflects the allocation of overt attention (in contrast to covert attention, i.e., the ability to attend to parts in the visual field without moving the eyes). Most knowledge about overt (and covert) visual attention stems from highly controlled psychophysical experiments, which are carefully designed to isolate and examine a single attentional mechanism. Thus, these experiments only provide limited information on the extent to which these mechanisms actually guide our decisions on where to move our eyes in complex dynamic scenes like real-world videos. In this work, we present a modular, mechanistic computational framework to bridge the gap between attentional mechanisms unveiled under standardized conditions and their effect on *where* and *when* we move our eyes when observing ecologically valid scenes.

Computational models of visual attention in free-viewing conditions have proven to be a powerful tool for understanding human exploration behavior (Itti and Koch, 2001; Borji and Itti, 2012). Existing models of human eye movements, however, typically suffer from at least one of two major simplifications: the restriction to static scenes, and the static nature of the predicted behavior as an average spatial distribution of fixation locations rather than modeling eye movement patterns of individual observers. Scanpath prediction in static scenes was pioneered by the seminal work of Itti, Koch, and Niebur (1998), who implemented the previously postulated concept of a saliency map (Koch and Ullman, 1985) algorithmically and suggested a strategy to sequentially select locations in the saliency map based on a "winner-take-all" and a subsequent "inhibition of return" mechanism. A more detailed model of attentional dynamics and saccadic selection in static scene viewing was proposed with the *SceneWalk* model family (Engbert et al., 2015; Schütt et al., 2017; Schwetlick, Rothkegel, et al., 2020; Schwetlick, Backhaus, and Engbert, 2022). These models predict the likelihood of moving the gaze to a certain position based on a foveated access to saliency information and a leaky memory process. Using a similar mathematical approach, Malem-Shinitski et al. (2020) model the decision for choosing an upcoming saccade target as a selection process governed by either local or global attention. The currently best-performing scanpath prediction model for static scenes, DeepGaze III (Kümmerer, Bethge, and Wallis, 2022), uses a deep neural network for extracting high-level image features and combines them with the fixation history to predict the most likely fixation target.

The predicted outputs of these scanpath models for static scenes are ordered lists of (x, y) fixation locations for a given image. The saccade timing, i.e., when to switch from one predicted fixation location to the next, is typically not of interest. Similarly, the standard evaluation metrics for scanpaths only consider predicted fixation locations but do not take the duration of fixations into account (Kümmerer and Bethge, 2021). In dynamic scenes, the image content at a given (x, y) location also depends on the time, and the scanpath must be specified as a list of (x, y, t) locations and requiring prediction of saccade timing. Although the temporal aspect of scanpaths is much less investigated, there is preliminary evidence that variability in fixation duration depends on both cognitive factors and visual features within the scene (Henderson, Weeks Jr, and Hollingworth, 1999; Underwood and Foulsham, 2006). In the framework proposed here, a saccade is performed as soon as a potential target in the saccadic decision-making process has accumulated sufficient evidence. Similarly, Nuthmann, Smith, et al. (2010) model the decision-making mechanisms that account for saccade timing and programming as a random-walk process in the CRISP model which, however, does not make any prediction about the (x, y) location of the saccade target.

A few models exist that simultaneously account for when and where the eyes move when exploring static scenes. Kucharsky et al. (2021) argue for the need for models that combine the spatial and temporal aspects of gaze behavior and present the WALD-EM model. They introduced a spatiotemporal likelihood function to statistically model fixation positions and fixation durations, which are assumed to follow a WALD (inverse Gaussian) distribution, simultaneously and were successful in

reproducing many aspects of scanpath statistics. The most recent version of the SceneWalk model (Schwetlick, Backhaus, and Engbert, 2022) also incorporates fixation durations when generating scanpaths. The distribution of fixation durations is assumed to follow a Gamma distribution, which is parameterized such that the mean fixation duration at a location parametrically depends on the local image saliency at this position. As in the WALD-EM model (Kucharsky et al., 2021), fixation durations are then included in the likelihood function, allowing for Bayesian inference of the model parameters. This approach could, in theory, be extended to scanpaths simulation in dynamic scenes, but updating the saliency map for every frame leads to a number of practical implications that make it computationally infeasible. In the LATEST model (Tatler, Brockmole, and Carpenter, 2017), each pixel-location of an image accumulates evidence in favor of moving the gaze from the current position to this location depending on scene features (like salience and semantic interest) and oculomotor factors (like saccade amplitude or change in direction). Being a statistical model, each respective influence on saccade timing is estimated from human scanpaths using generalized linear mixed models, without focusing on explaining the biological mechanisms that influence these factors. The location that is quickest to reach sufficient evidence consequently becomes the saccade target, resulting in a simultaneous prediction of where and when to move the gaze position. Predicting the saccade timing and target location simultaneously is especially important in dynamic scenes, where the scene content at a given pixel location can change over time. The evidence accumulation process in LATEST, a linear approach to threshold with ergodic rate (Noorani and Carpenter, 2016), however, does not generalize to situations when scene features change during the accumulation process.

Apart from the fact that the designs of existing models (in their current form) are not able to process dynamic scenes, scene changes also lead to qualitatively different gaze behaviors. The presence of motion leads to a decrease in between-subject variability (Goldstein, Woods, and Peli, 2007; Dorr et al., 2010; Smith and Mital, 2013), and motion is the most important independent predictor for where people tend to look (Carmi and Itti, 2006; Itti, 2005; Mital et al., 2011). Dynamic scenes also give rise to a type of eye movement that models have yet to account for, smooth pursuit, where the eyes remain fixated on a moving object. Modeling scanpaths in dynamic scenes hence requires a new class of models that account for the location and timing of saccadic decisions, smooth pursuit eye movements, and scene changes as a result of object or camera movement.

Existing models for visual attention in *dynamic scenes* do not attempt to model the actual visual exploration behavior. Instead, they predict frame-wise saliency maps, quantifying the average spatial distribution of human gaze locations. Dynamic saliency models can be implemented by extending static saliency models to include motion features such as flicker and oriented motion energy. These motion conspicuity maps can be combined with maps of other low-level features to create a single saliency map for each frame (Niebur and Koch, 1995; Itti, 2005; Le Meur, Le Callet, and Barba, 2007). Molin, Etienne-Cummings, and Niebur (2015) suggest that the best way of integrating motion into a static, low-level visual saliency model is by spatiotemporal filtering of the image intensity, inspired by the temporal receptive field of simple cells in V1. With the availability of larger datasets in recent years (Marszalek, Laptev, and Schmid, 2009; Wang, Shen, et al., 2018), video saliency detection has also become a popular task in computer vision. Deep neural network (DNN) architectures, which include the temporal information in videos either through temporal recurrence (Linardos et al., 2019; Droste, Jiao, and Noble, 2020) or by using 3D convolutional networks (Min and Corso, 2019; Jain et al., 2021; Wang, Liu, et al., 2021), clearly outperform mechanistic models from computational neuroscience and psychology in predicting where humans tend to look. This boost in performance can be explained by the capabilities of these networks not just to encode information on low-level features like color or edges. Instead, the DNNs, which are trained on human eye tracking data, typically use pre-trained features from object detection or action recognition tasks and learn to extract high-level features characterizing objects and scene information during the learning process. Despite the remarkable performance of deep learning approaches in predicting where humans tend to look on average, the insights such models provide about the visual system are limited. Interpreting DNNs is notoriously hard and crucial information about how individuals decide to explore the scene is lost when averaging the viewing behavior of multiple observers.

A related task to video saliency detection is egocentric gaze prediction (Li, Fathi, and Rehg, 2013; Huang, Cai, et al., 2018; Tavakoli et al., 2019), which aims at predicting gaze position recorded with a mobile eye tracker from the first-person

perspective. The field of view of the video data changes depending on the body and head movement of the participant, which is typically not predicted in this task. Despite being ecologically relevant, egocentric gaze prediction is therefore limited in the sense that for each video stimulus, there is only one ground truth scanpath, and that gaze is only one component of the exploration behavior, with head and body movement already provided by the field of view of the camera.

In the absence of task instructions, it has long been thought that the targets of eye movements are selected based on the bottom-up saliency of simple visual features (Itti and Koch, 2001). However, there is growing evidence that objects also play an important role in the selection process (Nuthmann and Henderson, 2010; Pajak and Nuthmann, 2013; Einhäuser, Spain, and Perona, 2008). It has been established in a wide range of psychophysical experiments that objects influence visual attention in multiple ways (e.g., Egly, Driver, and Rafal (1994), Kahneman, Treisman, and Gibbs (1992), and O'Craven, Downing, and Kanwisher (1999), see Scholl (2001) and Cavanagh et al. (2022) for reviews, and Sec. *Objects* in *Methods: ScanDy* for more details). This inspired several computational models of visual attention that incorporate or predict object-based information. Walther and Koch (2006) describe how units of visual information, so-called proto-objects, can be formed based on a saliency map. Other models segment proto-objects based on coherent image regions (Yanulevskaya et al., 2013) or perform perceptual grouping (Russell et al., 2014; Yu et al., 2015) and use these representations to improve saliency estimations in images. Sun and Fisher (2003) present a machine vision system based on object-based attention, which predicts attentional shifts on synthetic and natural images. Consistent with the experimental evidence for the importance of object-based attention, these object-based attention models typically report improvements compared to respective location-based modeling approaches. Due to fundamentally different model architectures and complexities, the comparability between object-based models and location-based baselines is, however, usually rather limited.

In this work, we propose a new modeling framework called *ScanDy* that is capable of simulating realistic **scan**paths for **dy**namic real-world scenes, including saccade timing and smooth pursuit behavior. To achieve this, we utilize the output of video saliency models, combine it with the attentional dynamics of static scanpath models, adapt a saccadic decision-making model to make it applicable to dynamic scenes, and show that predicted scanpaths benefit from object-based attention and selection. *ScanDy* is a mechanistic modeling framework with a modular design. Human scanpaths are simulated based on the features of the scene, with visual sensitivity depending on the current gaze position and the previous scanpath history. In contrast to a deep learning approach, attentional mechanisms are implemented with a small number of well-interpretable parameters based on insights from psychophysical experiments. This allows the implementation of scanpath models, which differ in their assumptions about one or more mechanisms underlying gaze behavior, within the same framework.

We demonstrate in this work how psychophysical hypotheses for scanpath simulation in dynamic real-world scenes can be tested using this framework. We focused on the role of objects for attentional guidance and how object and saliency information are combined. For this, we compared four model implementations within the *ScanDy* framework: Two location-based models, where one is based on low-level saliency (*L.ll* model) and one on high-level saliency (*L.hl* model), and two object-based models, with one incorporating low-level saliency for each object (*O.ll* model) and the other one not using any saliency information but only the center bias (*O.cb* model). Model parameters were fitted such that the simulated scanpaths reproduce the saccade amplitude and fixation duration distributions of free-viewing eye tracking data on videos of the VidCom (Li, Qin, and Itti, 2011) dataset (cf. Sec. *Dataset* in *Methods*). There are currently no established solutions for how to evaluate the model predictions for simulated scanpaths on dynamic scenes (cf. Sec. *Evaluation of simulated scanpaths* in *Discussion*). We assessed the model performance in out-of-training predictions concerning spatial and temporal fixation behavior, including the proportion of fixations detecting, inspecting, and revisiting objects as well as the proportion of fixations exploring the background (cf. Sec. *Functional scanpath evaluation* in *Results*). The comparison of different models within the *ScanDy* framework shows that a model using object-based attention, which selects objects for foveal processing based on visual saliency (i.e., the *O.ll* model), produces exploration behavior that most closely resembles that of humans. We thus conclude that object-level attentional units play an important role in guiding attentional processing. The *ScanDy* framework, together with the code and data for reproducing our results, is open source and can be accessed on GitHub: https://github.com/rederoth/ScanDy.

Methods

ScanDy: A modular framework for scanpath simulation in dynamic real-world scenes

We introduce a new computational framework for **scanp**ath simulation in **dy**namic real-world scenes: *ScanDy*. It predicts the timing and the spatial positions of foveation events (fixation and smooth pursuit) and saccades. Saccade locations and timing are modeled through a sequential decision process between potential targets. The target selection depends on the spread of visual sensitivity around the current gaze position (Sec. *Visual sensitivity (II)*), the scene content quantified by the salience of the potential target (Sec. *Scene features (I)*), and the recent scanpath history (Sec. *Inhibition of return (III)*). The decision process of where and when the eyes should move is modeled with a drift-diffusion model (DDM, Sec. *Decision-making (IV)*) with one accumulation process for each potential saccade target. As soon as the decision threshold of a potential target is reached, a saccade is executed to this target (Sec. *Gaze update (V)*) and the decision process is reset. In between saccades are foveation events, where the gaze position stays at the previously selected target. Depending on the motion of the foveated target in the visual field, these foveation events result in either fixation or smooth pursuit. During all foveation events, we model fixational eye movements (Martinez-Conde, Macknik, and Hubel, 2004; Rolfs, 2009; Rucci and Victor, 2015) as an additive Brownian motion on top of the current gaze position. In the following, we describe the individual components of the framework and their respective realizations in purely location-based (Fig. 1) and object-based (Fig. 2) model variants. We then compare the predicted scanpaths obtained from the two model families, each with two different saliency representations, to evaluate our hypothesis that objects play an essential role in attentional guidance.

Objects

While not being considered at all in location-based models, segmented object masks are the fundamental building block of object-based models. What constitutes a visual object is an open question (Spelke, 1990; Palmeri and Gauthier, 2004) and can strongly depend upon a given task or the scene content (Scholl, Pylyshyn, and Feldman, 2001; Feldman, 2003). In the context of object-based visual selection, there has to exist a pre-attentive neural representation of objects to allow for attention to be directed to them. These pre-attentive units for visual attention have been called proto-objects (Rensink, 2000; Walther and Koch, 2006). The literature on unconscious object representations (cf. Deeprose and Andrade (2006), Tanaka et al. (1991), and Ansorge, Kunde, and Kiefer (2014)) and unconscious object-based attention (cf. Chou and Yeh (2012) and Huang, Yeh, et al. (2015)), however, suggests that the pre-attentive targets of object-based attention can be fully developed object representations with boundaries that do not need to be refined when attended (Cavanagh et al., 2022). Neri (2017) shows that the global structure of a scene, defined by meaningful object boundaries, can be inferred within the first 100 ms and suggests that perception is quickly organized around the inferred scene layout. In addition, eye tracking data across different tasks and types of scenes have shown that observers tend to direct their gaze towards the center of objects as defined by semantic categories rather than towards preliminary proto-objects (Nuthmann and Henderson, 2010; Pajak and Nuthmann, 2013; Borji and Tanner, 2015) and such semantic objects have been shown to predict gaze behavior better than low-level saliency (Einhäuser, Spain, and Perona, 2008). Taken together, the experimental evidence suggests assuming semantic object segmentation masks as representations for object-based attention and selection. Hence, we computed the object masks based on semantics using the Mask R-CNN deep neural network architecture (He et al., 2017) as implemented in the Detectron 2 framework (Wu et al., 2019).

We used the VidCom dataset(Li, Qin, and Itti, 2011), which consists mainly of outdoor videos of people, animals, and traffic. Given this scene content, we defined humans, animals, and vehicles as objects and assigned the remaining portion of the scene to a general background category. We denote the object segmentation masks by O_i, $i \in [0, N_O]$, where N_O is the number of objects appearing in a given video and O_0 is the segmentation mask of the background. Segmented objects were tracked across frames using DeepSort (Wojke, Bewley, and Paulus, 2017) and manually controlled for obvious segmentation or tracking mistakes. As argued for in Cavanagh et al. (2022), we assumed that object masks are fully evolved before being attended and do not depend on any visual information which could have been acquired during the scanpath. Future versions of *ScanDy* may, for comparison, have the option to start using coarse object representations, which will then be refined as

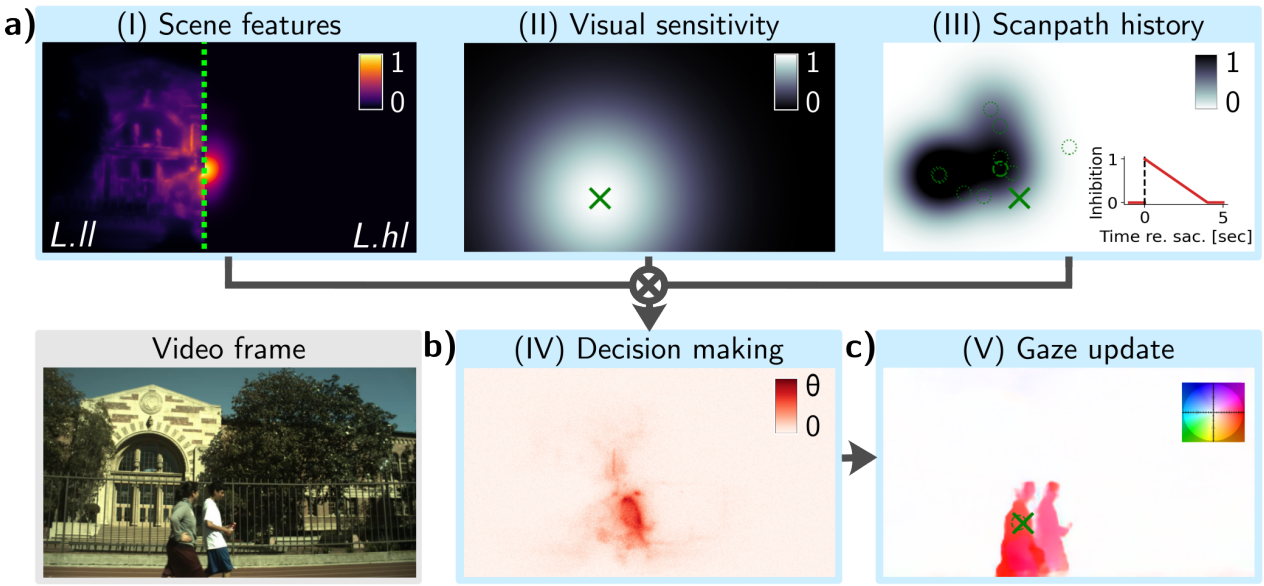

Figure 1. Schematics of a location-based model within the modular *ScanDy* framework, illustrated for an example frame (*field03* of the VidCom dataset, see bottom left). (a) Modules (I-III) are computed simultaneously for each time step. The scene features (I) are quantified by a saliency map. For location-based models, we compared the impact of low-level saliency (Molin, Etienne-Cummings, and Niebur, 2015) (*left*) in model *L.ll* with high-level saliency as predicted by a deep neural network (Min and Corso, 2019) (*right*, maps separated by the green dotted line) in model *L.hl*. The frame-wise saliency maps are multiplied with a generic center bias (Clarke and Tatler, 2014). Color represents relative saliency, normalized by the highest saliency in the image. The visual sensitivity (II) is implemented as a Gaussian of width σ_S centered at the current gaze point (green cross). The location-based inhibition of return mechanism (III) inhibits the previous target locations (shown as dotted circles, latest target (h=1 in Eq. (1)) as a dashed circle). A Gaussian of width σ_I is centered around each previous saccade target location. The amplitude of each Gaussian linearly decreases relative to the time of the respective saccadic decision with slope r (see inset). The inhibition map is the sum of weighted Gaussians around previous target locations, clipped to a maximum of 1. The output maps of these modules (I-III) are multiplicatively combined for the decision-making process (see Eq. (5)). (b) The evidence for each potential saccade target (i.e., each pixel location) is accumulated in the decision variable (cf. Eq. (5)) in module (IV). Each potential target is represented in a drift-diffusion model (DDM) where the drift rate is computed by combining values from maps of modules (I, *left*, model *L.ll*), (II) & (III), and the noise level s. (c) Module (V) updates the gaze position based on the resulting decision variables. If one pixel location in (IV) reaches the DDM decision threshold θ, a saccade to this position is executed. Otherwise, the gaze position is updated (from dashed circle to cross) based on the optical flow calculated using PWC-Net (Sun, Yang, et al., 2018) (plotted with flow field color coding in the top right where color indicates direction and hue indicates velocity). The parameters of the location-based model are listed in Tab. 1.

visual information is gathered during scene exploration. Other object classes – or object representations different from R-CNN computed masks – can also be included in the framework as long as the object masks are temporally consistent, meaning that an object appearing in consecutive frames has the same object ID i.

Scene features (I)

The simulated eye movements are based on the visual information in the scene, which we quantify by frame-wise feature maps F. The default choice for F is a bottom-up saliency computation based on motion, intensity, color, and orientation (Molin, Etienne-Cummings, and Niebur, 2015) (Fig. 1a, I, *left*).

Alternatively, target relevance can be based on high-level features of the scene (Fig. 1a, I, *right*). In this case, F is computed with a deep neural network architecture called TASED-Net (Temporally-Aggregating Spatial Encoder-Decoder Network, (Min and Corso, 2019)), which is trained on predicting the spatial distribution of fixation locations for each frame by using information from previous frames. As a result, TASED-Net prioritizes high-level scene elements, like faces, when compared to feature-based saliency models (Molin, Etienne-Cummings, and Niebur, 2015), and is highly sensitive to salient moving objects (Min and Corso, 2019), since it was implicitly trained to track and attend to objects. All saliency maps $F(x,y)$ used here were normalized to $[0,1]$ for better comparability.

When presenting complex scenes on a computer screen, observers tend to make fixations towards the center of the display. This is partially due to a tendency of interesting objects to be placed in the center during recording (*photographer bias*) (Parkhurst and Niebur, 2003; Tatler, Baddeley, and Gilchrist, 2005) but has been shown to persist irrespective of the scene content (Tatler, 2007; Clarke and Tatler, 2014). Potential explanations for this scene-independent center bias are a learned expectation of interesting objects in the center (as a consequence of repeated exposure to photographer-biased stimuli), the maximal utilization of peripheral vision when looking at the center, a preference of humans for a straight-ahead position of the eyes, and fixational starting markers in the scene center in experimental paradigms combined with an oculomotor bias for short saccades (Tseng et al., 2009). Hence, we explicitly include this scene-independent center bias in our model. Following prior work (Clarke and Tatler, 2014; Nuthmann, Einhäuser, and Schütz, 2017), we quantified the center bias as an anisotropic Gaussian $G_{cb}(x,y) = \exp\left(-\frac{(x-x_c)^2}{2\sigma_x^2} - \frac{(y-y_c)^2}{2\sigma_x^2 \cdot \nu}\right)$ with $\sigma_x^2 = 0.22$ (for x-values normalized to $-1 \leq x \leq 1$) and anisotropy ratio $\nu = 0.45$ and located at the image center (x_c, y_c). All saliency maps in this work were multiplied with G_{cb} to account for this central viewing bias.

Visual sensitivity (II)

In the foveated human visual system, visual acuity decreases with increasing eccentricity (Basler, 1906). Hence, the extent to which certain features in a scene influence eye movements depends on the current gaze position (x_0, y_0), for which we calculate a sensitivity map S at each time step. We account for higher visual acuity in the fovea compared to the periphery with an isotropic Gaussian $G_S = \frac{1}{2\pi\sigma_S^2} \exp\left(-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma_S^2}\right)$. The standard deviation σ_S is a free model parameter with a strong influence on the size of saccades in the simulated scanpaths.

Visual acuity is, however, just a first-order approximation of the actual sensitivity across our visual field. We can selectively attend to objects or locations (Eriksen and Colegate, 1971; Posner, 1980). If humans are presented with a cue within an object, they respond faster to a target within the same object compared to a target outside of it, matched for the distance to the cue (Egly, Driver, and Rafal, 1994; Malcolm and Shomstein, 2015). This suggests a perceptual benefit of the currently attended object with an object-based distribution of spatial attention (Egeth and Yantis, 1997; Scholl, 2001). In the object-based model, we approximate the spread of covert attention across a currently foveated object with a uniform sensitivity, replacing the part of the Gaussian G_S that falls within this object.

This sensitivity map can be extended to also include covert and perisaccadic attentional shifts, as modeled by Schwetlick, Rothkegel, et al. (2020). For simplicity and to keep the number of parameters low, the models described in the following do not include additional attentional mechanisms in the computation of the sensitivity map.

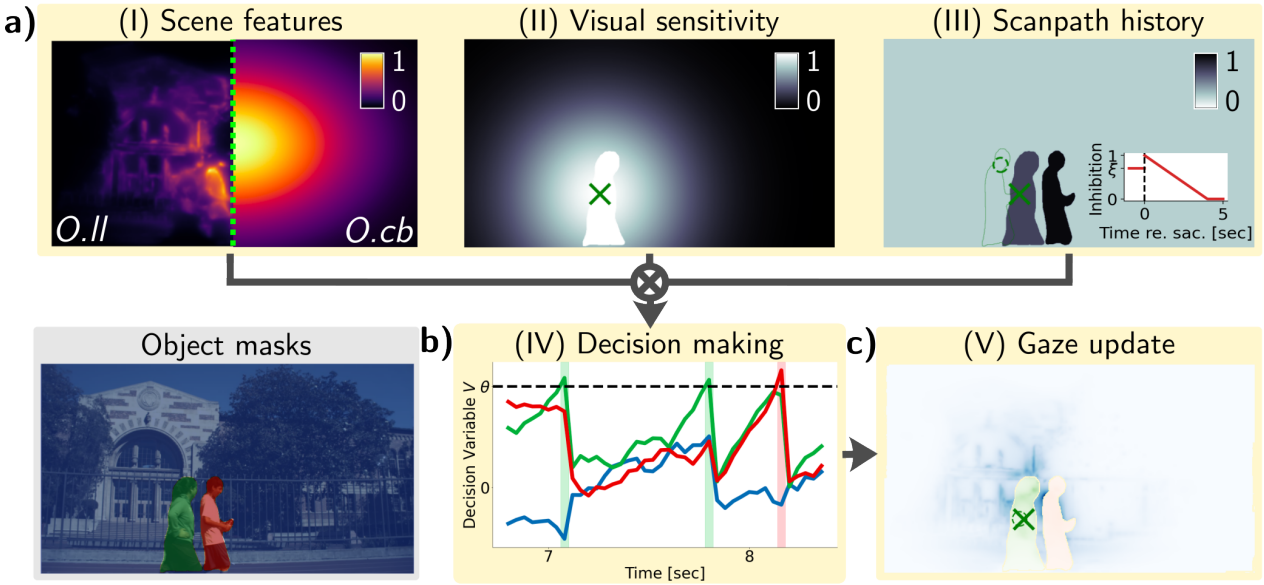

Figure 2. Schematic of an object-based model within *ScanDy*, in analogy to Fig. 1 and for the same video frame. The object masks (bottom left, 2 persons and background are shown) are based on a semantic segmentation using the Mask R-CNN deep neural network (implementation by Wu et al. (2019)). (a) Any appropriate feature map could be used to encode the scene content in module (I). For the model comparison, we used the same low-level saliency (Molin, Etienne-Cummings, and Niebur, 2015) as in Fig. 1 for the object-based model *O.ll* (I, *left*) but in a second implementation, model *O.cb*, did not include any kind of scene features and only used the generic center bias (Clarke and Tatler, 2014) (I, *right*). In addition to the location-based Gaussian of size σ_S, we account for a higher sensitivity across the currently foveated object (Egly, Driver, and Rafal, 1994; Malcolm and Shomstein, 2015) by setting the sensitivity in module (II) across the foveated object mask to one. Instead of inhibiting locations, module (III) inhibits previously foveated object masks (cf. Tipper, Driver, and Weaver (1991)). While foveated, an object is inhibited by value ξ (cf. left object). As soon as the gaze position moves outside of the object mask, the inhibition is set to one (cf. right person), which then, over time, decreases again linearly to zero with slope r (see inset). The scanpath history is identical to Fig. 1 (III), with the dashed circle marking the previous saccade target location. At the time of the previous saccade, the right person was at the location of the green object contour, but in the meantime moved to the right and was followed with smooth pursuit before a saccade was initiated to the now foveated object (green cross). The output maps of modules (I-III) are again combined (see panel (c) for pixel-wise multiplication result of (I-III), with different color maps for each object), but the visual information is now summed across each object mask and normalized by the logarithm of the object size (see Eq. (4)). The resulting value for each object is the drift rate for the drift-diffusion process (see Eq. (5)). (b) In object-based target selection, the evidence for the saccadic decision-making is accumulated for each object in the scene, quantified by a DDM with threshold θ and noise level s in module (IV). (c) The gaze position update (V) follows the movement of the foveated object mask (from dashed circle to cross). If the decision threshold is crossed, a saccade to the target object is executed. The exact landing position within the target object is probabilistic and proportional to the activity of the combined maps of modules (I-III) (see (6), combined maps shown in a different color for each object mask). The object-based models have the same number of free parameters as the location-based ones. They are listed in Tab. 2.

Inhibition of return (III)

Visual processing of stimuli, which have recently been the focus of attention, are often temporarily inhibited (Posner, Cohen, et al., 1984; Klein, 2000; Klein and Ivanoff, 2008). This inhibition of return (IOR) is commonly interpreted to encourage the exploration of previously unattended parts of the scene (Itti and Koch, 2001; Klein, 1988; Klein and MacInnes, 1999; Posner, Rafal, et al., 1985).

The spatial extent of IOR has been shown to be highest at the attended location and graded with distance (Maylor and Hockey, 1985; Bennett and Pratt, 2001; Berlucchi et al., 1989; Klein, Christie, and Morris, 2005). IOR in the location-based models is, therefore, implemented by a Gaussian-shaped inhibition $G_I(x,y)^{(h)} = \frac{1}{2\pi\sigma_I^2}\exp\left(-\frac{(x-x_h)^2+(y-y_h)^2}{2\sigma_I^2}\right)$ around the (pixel-)locations (x_h, y_h) of previous saccade targets, where $h \in [1, N_F]$ denotes the index of previous saccade targets and N_F the number of previous foveations. The time course of how IOR affects gaze behavior over time is, despite the vast literature on the effect, not well established (Samuel and Kat, 2003). For simplicity, we assume that the effect decreases linearly with time. Hence, the location-based inhibition map I, bound between full ($I = 1$) and no inhibition ($I = 0$), at time t is given by

$$I(x,y,t) = \min\left(1, \sum_{h=1}^{N_F} G_I(x,y)^{(h)} \cdot \max\left(0, (1-(t-t_h)r)\right)\right), \tag{1}$$

where r determines the slope of the linear decrease with time and (x_h, y_h, t_h) the respective target location and time. If r is close to zero, all N_F previous foveation targets contribute to the inhibition, resulting in previously selected locations not being revisited. The standard deviation σ_I of the Gaussian influences the average size of saccades and can, even though objects are not considered in this implementation, influence how often saccades are made within the same object.

Various paradigms suggest that IOR can be bound to objects in that it tracks and follows moving objects (Tipper, Driver, and Weaver, 1991; Gibson and Egeth, 1994; Ro and Rafal, 1999; Becker and Egeth, 2000). In addition, IOR was found to spread across an object if attention was allocated to parts of the object (Jordan and Tipper, 1999; Reppa and Charles Leek, 2003). In the object-based models, we hence do not calculate the inhibition for each location as in Eq. (1) but do so for each object in the scene. As long as an object is already foveated, meaning that the current gaze position (x_0, y_0) lies within the object segmentation mask O_i, the selection of this object for the next saccade is inhibited with the within-object factor $\xi \in [0,1]$. To account for the typically much larger size and the heterogeneity of the background, we do not inhibit saccades within the background (within-object inhibition $\xi = 0$ for the background object). The inhibition for each object is set to one as soon as it is not foveated anymore and then linearly decreases with slope r towards zero, as in the location-based IOR in (1). Consequently, the inhibition I_i for an object at a given time is calculated as

$$I_i(t) = \begin{cases} \xi, & \text{if } (x_0, y_0) \in O_i \\ \max\left(0, 1 - r \cdot (t - t_h^{(i)})\right), & \text{else,} \end{cases} \tag{2}$$

where $t_h^{(i)}$ is the point in time when the gaze position was last on the object mask O_i.

The feature map F, the visual sensitivity map S, and the inhibition I of targets (represented by objects or pixel locations) are multiplicatively combined to determine the decision of where to saccade next.

Decision-making (IV)

We describe human scanpaths as a sequential process of saccadic decisions of either moving towards potential targets or staying at the currently foveated target (Ludwig, 2011; Tatler, Brockmole, and Carpenter, 2017). Such a perceptual decision-making process is typically described as a noisy accumulation of evidence from a given stimulus, which can be quantified using drift-diffusion models (DDM) (Ratcliff, 1978; Bogacz et al., 2006; Ratcliff and McKoon, 2008). The DDM is a well-established tool for explaining the cognitive and neural processes during decision-making (Gold, Shadlen, et al., 2007; Forstmann, Ratcliff, and Wagenmakers, 2016; Ratcliff, Smith, et al., 2016). It is based on the assumption that evidence for each choice is accumulated

over time (drift), while each drift process is perturbed by random fluctuations (diffusion); a decision is reached as soon as sufficient evidence for one alternative is accumulated (Bogacz et al., 2006). The DDM, by design, does not only model the perceptual choice but also the time until a decision is reached.

Since DDMs are typically applied to two-alternative forced choice tasks, we extend the evidence accumulation model to multiple options. Each potential saccade target is represented by a drift process with a time-dependent rate $\mu_i(t)$ and a noise strength s. As soon as one process, representing one potential target, reaches the decision threshold θ, a saccade is executed towards this target.

The drift rate of a target μ_i is determined for each frame as the product of the scene features F, the visual sensitivity S, and the inhibition I at the respective target based on the scanpath history. In the location-based models, we treat every pixel location $i \in [1, X \cdot Y]$ as a potential saccade target. Hence, the drift rates are computed as

$$\mu_i(t) = \mu(x,y,t) = F(x,y,t) \cdot S(x,y,t) \cdot (1 - I(x,y,t)). \tag{3}$$

In the object-based models, the objects in the scene define the potential saccade targets. Each object mask O_i has the same dimension as a video frame, with entries within the object mask being 1 and outside the object being 0. To account for the different sizes of objects, we adapt the common practice to scale each object's perceptual size logarithmically (cf. Nuthmann, Einhäuser, and Schütz (2017)). The drift rate for each object μ_i is then given by

$$\mu_i(t) = \frac{\left(\sum_{x,y} O_i(x,y,t) \cdot F(x,y,t) \cdot S(x,y,t)\right)}{\sum_{x,y} O_i(x,y,t)} \cdot (1 - I_i(t)) \cdot \log \sum_{x,y} O_i(x,y,t). \tag{4}$$

Each target in this framework accumulates evidence in favor of moving the gaze from the current position to the respective target. The evidence is based on the target's drift rate μ_i and random fluctuations $\varepsilon \sim \mathcal{N}(0, 1)$ from the diffusion term scaled with the noise level s. The accumulated evidence of a target is given by its decision variable V_i, which is updated as

$$V_i(t + \Delta t) = V_i(t) + \mu_i(t)\Delta t + s\varepsilon\sqrt{\Delta t}. \tag{5}$$

In the following, we set the time resolution measured in frames for the update to $\Delta t = 1$. As soon as the decision variable of one target V_i reaches the decision threshold θ, a saccade is triggered towards this target and the decision variables for all targets are reset to zero. The different number of potential saccade targets in a location-based compared to an object-based model (the number of objects is much smaller than the number of pixels for any given scene) can be accounted for by choosing the free parameters θ and s appropriately.

Gaze update (V)

The gaze position in our modeling framework is updated at every time step (i.e., video frame). As long as all decision processes are below the DDM threshold, no decision is made and no saccade is triggered. In between saccades, the gaze position of humans also changes due to fixational eye movements and, since we consider dynamic scenes, smooth pursuit, which describes the eye movements when the gaze is fixated on a moving object. In the object-based models, this is implemented by calculating the shift of the center of mass of the currently foveated object between two frames and updating the gaze position accordingly. Consequently, the gaze follows the currently foveated target object in a smooth pursuit motion or stays fixated if the object does not move. In the location-based models, we do not have object information available. To simulate realistic gaze behavior and produce comparable scanpaths, we achieve smooth pursuit by updating the gaze position according to the optical flow, calculated using PWC-net (Sun, Yang, et al., 2018). In addition, we model fixational eye movements like oculomotor drift or microsaccades (Martinez-Conde, Macknik, and Hubel, 2004; Rolfs, 2009; Rucci and Victor, 2015) by updating the gaze position in every time step according to a 2D random walk with $\Delta x, \Delta y \in \mathcal{N}(0, \sigma_D)$. To match the fixational eye movements of human observers across frames to the frame rate of 30 fps in the VidCom dataset, we choose $\sigma_D = 0.125$ degrees visual angle (dva). If the decision threshold θ in the DDM is reached for a target, a saccade is triggered, and the gaze location at the next

Table 1. Free parameters of the location-based models and their values which were determined by evolutionary optimization. Columns "Fit (last gen.)" show the mean and standard deviation across all 32 individuals from the last generation of the evolutionary optimization (for a plot of the full parameter space, see the *Supplemental Material*). Columns "Fit (top ind.)" show the parameter values that lead to the best fitness among all individuals, which are the parameter values used for the model comparison shown in Fig. 4. The parameter for the size of the oculomotor drift is fixed to $\sigma_D = 0.125$ dva in all models.

Parameter	Description	Fit (last gen.)	Fit (top ind.)	Fit (last gen.)	Fit (top ind.)
		L.ll (low-level features)		*L.hl* (high-level features)	
σ_s	Range of sensitivity [dva]	12.52 ± 0.49	12.24	9.41 ± 0.78	8.45
θ	DDM decision threshold	0.396 ± 0.041	0.314	1.583 ± 0.076	1.547
s	DDM noise level	0.013 ± 0.002	0.011	0.088 ± 0.004	0.086
$1/r$	Slope of IOR decrease	168.2 ± 32.1	194.3	285.1 ± 13.2	285.4
σ_I	Range of IOR [dva]	7.22 ± 0.27	7.37	4.1 ± 0.11	4.1

Table 2. Parameters of the object-based models and their values which were determined by evolutionary optimization. For details, see caption of Tab. 1.

Parameter	Description	Fit (last gen.)	Fit (top ind.)	Fit (last gen.)	Fit (top ind.)
		O.ll (low-level features)		*O.cb* (center bias only)	
σ_s	Range of sensitivity [dva]	13.36 ± 0.39	13.23	6.19 ± 0.1	6.27
θ	DDM decision threshold	1.809 ± 0.065	1.776	2.904 ± 0.059	2.920
s	DDM noise level	0.245 ± 0.004	0.248	0.223 ± 0.017	0.227
$1/r$	Slope of IOR decrease	225.4 ± 53.4	226.3	150.5 ± 58.0	97.3
ξ	Object-based inhibition	0.766 ± 0.021	0.780	0.918 ± 0.010	0.920

time step is set to be on the corresponding saccade target. In a location-based model, the new gaze position equals the pixel location of the target. The exact landing position of a saccade at time t_0 in an object-based model is sampled from all pixel locations within the object mask O_i of the target object, where the probability $p_i(x,y)$ of each pixel to be the next gaze position is proportional to the scene features F and the visual sensitivity S (inhibition is uniform within each object), with

$$p_i(x,y) \sim O_i(x,y,t_0) \cdot F(x,y,t_0) \cdot S(x,y,t_0). \tag{6}$$

Model parameters and evaluation

Both model families have five free parameters, which are listed together with the fitted parameter values in Tab. 1 for the location-based models and in Tab. 2 for the object-based models. The object-based model parameters are the range σ_S of the visual sensitivity (cf. module (II)), the amount of inhibition ξ of the currently foveated object, the slope r of the decrease of inhibition of return over time (both (III), see Eq. (2)), the DDM threshold θ, and the noise level s of the diffusion process (both (IV), see Eq. (5)). In the location-based implementation, only one parameter changes. Since the inhibition I in Eq. (1) is not bounded by the outlines of objects in the scene, we specify the range σ_I of the Gaussian-shaped inhibition of return around the previous fixation locations.

The influence of model parameters and the predicted scanpaths were explored in extensive grid searches (see Figs. S1 and S2 in the *Supplemental Material*). The decision threshold θ of the DDM directly influences how long it takes to reach a decision and controls the mean duration of foveation events. With a low DDM noise level s, multiple simulations for the same scene lead to similar scanpaths, while a higher value of s leads to a higher scanpath variability across different simulations. The Gaussian spread σ_S of visual sensitivity determines the effective field of view around the current gaze position and strongly influences the average saccade amplitude. The slope r of the decrease of inhibition over time determines the frequency of saccades that return to a previous target. The strength ξ of the inhibition in the object-based models and the size σ_I of the inhibition in the location-based models determine how often saccades are made within the same object as well as the frequency of saccades with short amplitudes.

Our framework was designed in a way that makes it easy to implement and prototype models (cf. Sec. *Framework architecture*) that reflect competing hypotheses. In this work, we compared four different models (*L.ll*, *L.hl*, *O.ll*, *O.cb*), which have different assumptions about the role of objects and saliency information for attentional guidance. To evaluate these, we need an objective way of determining the optimal values of the free parameters for each model. Here we made use of the evolutionary optimization algorithm implemented in *neurolib* (Cakan, Jajcay, and Obermayer, 2021). We used a Gaussian mutation algorithm with adaptive step size (Eiben, Smith, et al., 2003). We optimized the similarity between simulated scanpaths and the human ground truth data with respect to two summary statistics: the distributions of saccade amplitudes and foveation durations. The corresponding fitness function $\mathscr{F}$ maximized by the evolutionary optimization is given by

$$\mathscr{F} = -\frac{d_{SA} + d_{FD}}{2}, \tag{7}$$

where the distance measures d_{SA} for the saccade amplitudes and d_{FD} for the foveation durations are given by the two-sample Kolmogorov-Smirnov (KS) statistic $D = \sup_x |F(x) - G(x)|$ of the respective distribution F compared to the ground truth distribution G from the human data. The parameters of each of the four models were optimized independently across 50 generations, using an initial population size of 64 and an ongoing population size of 32 individuals with the default settings of the evolutionary optimization (Gaussian mutation and rank selection with adaptive step size). The initial parameter ranges for each model are reported in the *Supplemental Material*. We optimized the fitness function, Eq. (7), simultaneously for all videos in the training set. We evaluated out-of-domain, i.e., on a different metric than what we optimized the parameters for, how realistic the simulated scanpaths are for a given input video by analyzing which objects are selected for foveal processing and what function these foveations fulfill, as proposed in Linka and de Haas (2021).

Framework architecture

We adopted the software structure (Fig. 3a) of the *neurolib* framework (Cakan, Jajcay, and Obermayer, 2021).Scanpath models in the *ScanDy* framework inherit from the `Model` base class. Its functionality includes initializing and running model simulations and the evaluation and visualization of the resulting scanpaths. Derived classes are the `ObjectModel` and `LocationModel`. Models that assume object-based selection use the `ObjectFile` class for each object within the scene. Analogous to the concept of object files in visual cognition (Kahneman, Treisman, and Gibbs, 1992), an instance of this class is used as a representation in which information about an object is linked and integrated. Independent from the specifics of the model, the simulated scanpaths have the same format as the eye tracking data from human participants. The human data is handled by the `Dataset` class, together with the corresponding video data, the precomputed object segmentation masks, the saliency maps, and the optical flow data for each scene. Which of these precomputed maps are used as inputs to the scanpath models depends on the specific hypotheses to be tested. The model parameters are handled by the `ParameterSpace` class of *neurolib*, which can be used for parameter explorations using the `BoxSearch` class or model optimizations using the `Evolution` class.

In the following, we briefly describe three potential use cases (Fig. 3b), which are provided as examples with the code, detailed explanations and visualizations in Jupyter Notebooks on *ScanDy*'s GitHub page. To simulate a scanpath for a single video using a predefined model (Fig. 3b, red), we first load the dataset that contains the precomputed saliency maps and object segmentations. Next, we either initialize an object- or location-based model and select the saliency map and reasonable values for the free model parameters (cf. Tabs. 1 and 2). With the initialized model, we only have to specify for which video(s) and how many scanpaths (different random seeds) should be generated in the run command. We can plot the resulting gaze positions on the video frame or analyze the scanpath and the output of the individual modules of the model using built-in functions.

In our second example (Fig. 3b, blue), we want to compare the simulated scanpaths of different models quantitatively. For fair model comparison, we must ensure that better model performance is not just due to better model parametrization. Consequently, instead of evaluating a model for a single set of parameters, we define a fitness function, as in Eq. (7), and specify a range for each free parameter that should be explored during the optimization. During the evolutionary fitting procedure,

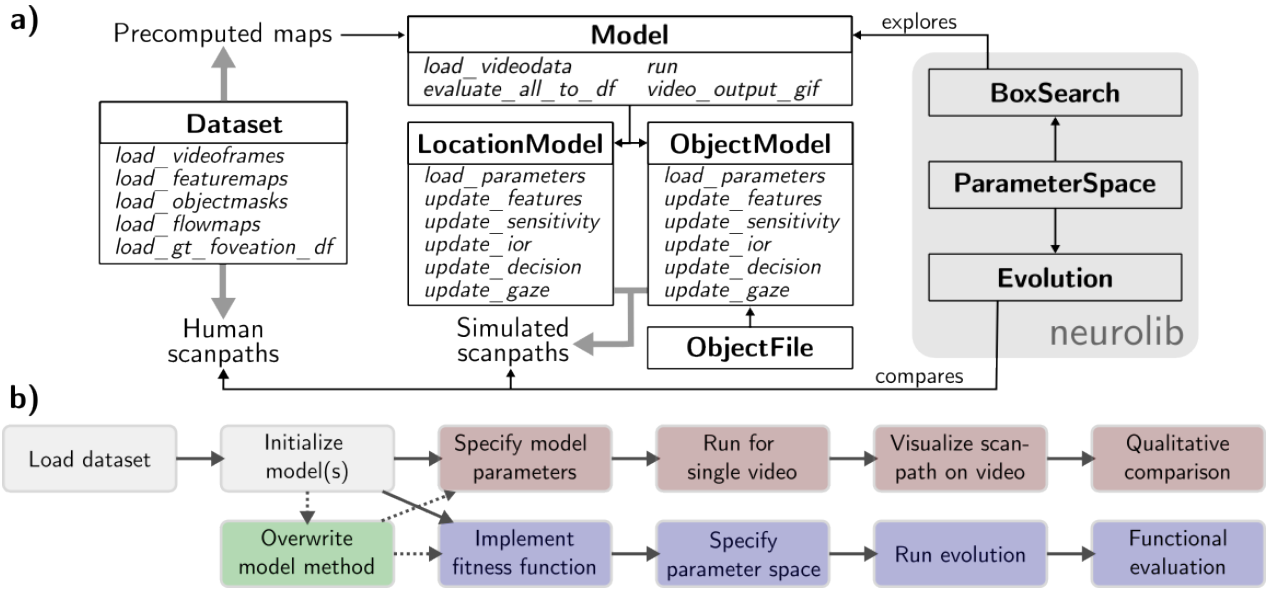

Figure 3. Software architecture of the *ScanDy* framework with example use cases. (a) Every box is a class with the name in bold and the most important methods in italic. The `Dataset` class provides the human scanpath data and makes the precomputed maps of the video data available for the models. At the core of the framework is the `Model` base class, from which all models inherit their functionality. The derived model classes have methods that correspond to the modules described in Figs. 1 and 2. Object-based models might use the `ObjectFile` class to integrate object information in an efficient way. The classes on the gray background are native to neurolib and are used within *ScanDy* for parameter exploration and optimization. (b) Possible use cases of our framework include, in increasing complexity, generating scanpaths for a single video (red), comparing how well different models capture human gaze behavior (blue), and extending on existing models to test new hypotheses (green, common starting points in light gray). We provide Jupyter Notebooks of these and more examples on GitHub.

the simulated scanpaths for different parameters are compared to the human data, finding parameter sets within the specified space that optimize our fitness function. The models with optimized parameters can then be compared in different ways to infer which underlying model assumptions best describe the attentional process in humans. For example, we can calculate the time spent in different foveation categories for a model (cf. Sec. *Functional scanpath evaluation* in *Results*).

In the third example (Fig. 3b, green), we demonstrate the framework's modularity and the ease of modifying existing models to reflect new hypotheses. As an example, we can assume the inhibition of return (III) mechanism and the target selection in the decision-making process (IV) to be location-based and only the spread of visual sensitivity (II) to depend on objects in the scene. This is achieved by initializing a location-based model and overwriting its `update_sensitivity` method where the visual sensitivity across the currently foveated object is set to one as in module (II) of an object-based model. The modified model still has all functionalities and can be analyzed as one of the predefined models (cf. example one and two).

Dataset

We compared the simulated scanpaths with human eye tracking data on high-resolution videos of natural scenes from the VidCom dataset (link to the video and eye tracking data: http://ilab.usc.edu/vagba/dataset/VidCom/ (Li, Qin, and Itti, 2011) that were recorded under free-viewing conditions (i.e., no task instructions). We chose this dataset due to its high-quality human eye movement data (14 participants recorded with a 240 Hz infrared video-based eye tracker). The full dataset comprises 50 videos of everyday scenes, mainly showing individuals or groups of people, animals, or traffic scenes, each with 300 frames at 30 fps and without scene cuts. The scenes have high ecological validity, making them good candidates to explore and evaluate our modeling framework. About half of the videos in the VidCom dataset show a large number of objects with partially strong occlusion (e.g., people or cars behind fences or nets), which leads to failures of our object detection and tracking algorithms, and the videos had to be excluded. If more than 100 object IDs were assigned in a video, even after a first manual clean-up, the object masks were considered to be not useful and the video was excluded. The remaining 23 clips were randomly split into 10 videos for training, i.e. optimizing the free model parameters, and 13 videos to provide a sufficient statistic for testing how realistic the simulated scanpaths are. A list of the video names, the number of human participants for each video, and if it was used in the training or test set can be found in the *Supplemental Material*.

We used the Deep EM classifier (Startsev, Agtzidis, and Dorr, 2019) for gaze event classification in human eye tracking data. Since our model predictions did not differentiate between fixation and smooth pursuit events, we combined them to foveation events. We evaluated for each point in time if the gaze position is on the background or within a 1 dva radius (to account for tracking inaccuracies or drift) of an object mask. To allow further analysis (see Sec. *Functional scanpath evaluation* in *Results*), we assigned every foveation event the ID of the object mask it spent the most time on. We excluded foveation events shorter than 33.3 ms, corresponding to the displayed time of an individual video frame (30 fps). We further excluded all saccades with amplitudes smaller than 0.5 dva, which we consider as microsaccades and hence as a component of fixational eye movements (Rolfs, 2009). The resulting summary statistics of foveation duration and saccade amplitude of the extracted eye movements are shown in Figs. 4a and 4b.

Results

ScanDy is designed to implement and test hypotheses about attention allocation and eye movement behavior when viewing dynamic real-world scenes. Depending on the hypothesis, competing model assumptions for individual mechanisms can be implemented within the framework. Here, we explore the hypothesis that objects play an essential role in attentional guidance, by systematically comparing predictions of models that incorporate different combinations of object and saliency information to the ground truth of human eye tracking data. Our first model *L.ll* is location-based and uses low-level feature information (Molin, Etienne-Cummings, and Niebur, 2015) quantifying concepts from saliency and the feature integration theory (Treisman and Gelade, 1980) of visual attention. To account for the possibility that visual attention and target selection are, in fact, location-based but that the low-level features used in the first model do not allow for a sufficiently good saliency prediction, we

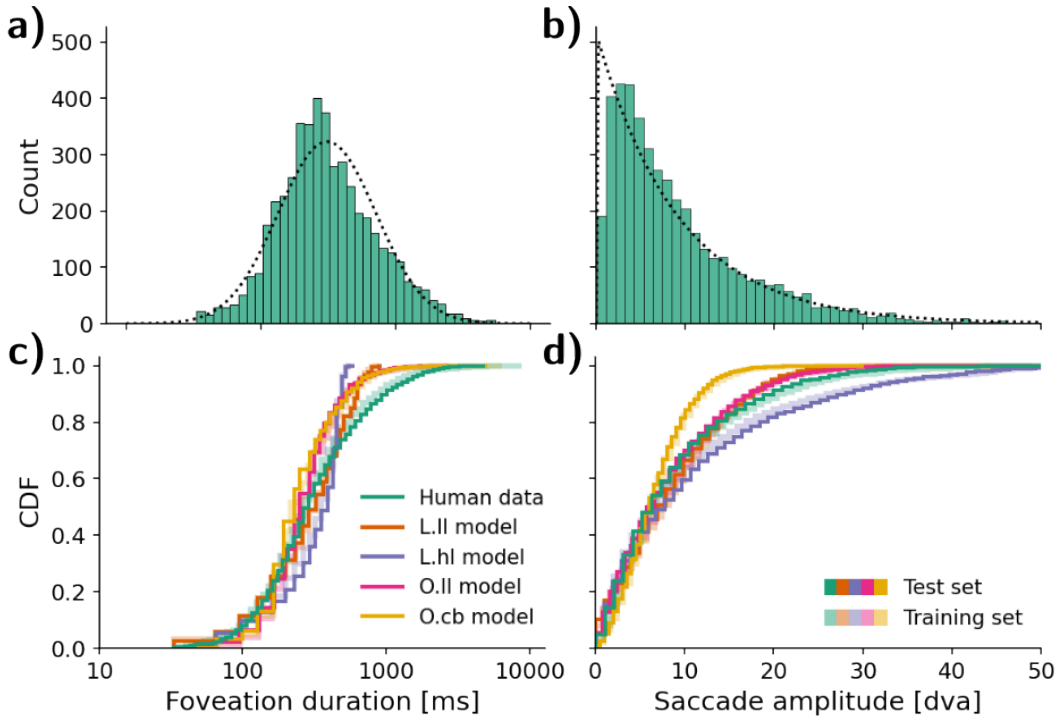

Figure 4. Scanpath summary statistics of the human eye tracking data across all videos compared to the simulated model scanpaths. For each video in the dataset (10 in training, 13 in test data), we simulate 12 scanpaths (same parameters but with different random seeds) to roughly match the number of ground truth scanpaths. (a) Ground truth foveation duration distribution from all human participants across all videos. The dotted curve is a fitted log-normal distribution with $\mu = 5.735$ and $\sigma = 0.838$ (equiv. to an expected value of $e^{\mu+\sigma^2/2} = 439.6$ ms). (b) Ground truth saccade amplitude distribution. The dotted curve is a fitted exponential distribution with scale $\tau = 1/\lambda = 9.227$ (equiv. to an expected value of $1/\lambda = 9.227$ dva). (c) Cumulative distribution functions (CDFs) of foveation durations are compared between human data (green) and the results of the location-based models with low-level features *L.ll* (brown), and high-level features *L.hl* (purple), the object-based model with low-level features *O.ll* (pink), and center bias only *O.cb* (yellow). The human data and modeling results from videos in the test set are represented by opaque lines, while transparent lines represent the corresponding training data and results. (d) Same as (c) but for saccade amplitudes. Model parameters were optimized using the evolutionary optimization algorithm. Parameters are listed in Tab. 1 for the location-based and in Tab. 2 for the object-based models.

introduce a second location-based model *L.hl* that uses high-level saliency as predicted by TASED-Net (Min and Corso, 2019). Our third model *O.ll* is object-based and uses low-level saliency to compute the relevance of each object as a potential saccade target. To better disentangle low-level saliency and object information, we add a fourth model *O.cb* that is object-based but does not use any scene features. It corresponds to the hypothesis that saliency is only relevant for eye movements through an object detection mechanism.

In the first part of the results, we show that all of these different models are capable of producing realistic scanpath statistics measured by the distributions of saccade amplitude and foveation duration. In the second part, we show that, out of the four models, *O.ll* leads to the most realistic gaze behavior. This is assessed by analyzing how the simulated scanpaths move the gaze towards objects in the scene, based on functional foveation categories (Linka and de Haas, 2021).

Scanpath summary statistics

Eye movement events in human scanpaths on dynamic scenes have distinct statistics. The distribution of durations of all foveation events from all participants across all videos approximately follows a log-normal distribution (Fig. 4) with a median duration of 287.5 ms and, due to a few very long smooth pursuit events, a mean duration of 456.4 ms (expected value of the fitted distribution is 439.6 ms). The second important statistic of ground truth scanpaths is the distribution of saccade

amplitudes (Fig. 4b). The distribution for saccades shows a dip at very small amplitudes, which is an "artifact" of the chosen delineation between saccades and fixational eye movements like microsaccades (see Sec. *Dataset* in *Methods*). With higher saccade amplitude, the distribution is well approximated by an exponential distribution with a mean amplitude of 9.29 dva (expected value of the fitted distribution is 9.23 dva).

Although we optimized the free model parameters to fit these two summary statistics (see Sec. *Model parameters and evaluation* in *Methods*), there was no guarantee that the models with such a low number of parameters could actually reproduce these distributions. The log-normal shape of the foveation duration distribution is not explicitly implemented in the framework but had to result in the way how the evidence for moving the eyes is accumulated in the DDM (see Sec. *Decision making (IV)* in *Methods: ScanDy*). Similarly, the exponential distribution of saccade amplitudes can only be reproduced if the way of implementing the visual sensitivity *(II)* and the inhibition of return mechanism *(III)* allow for it.

The models with optimized parameters closely correspond to the distribution of the human foveation durations (Fig. 4c). The cumulative distribution functions (CDFs) (plotted for the parameter configuration with the highest fitness for each model) provide an intuition on the distance between distributions as measured by the KS-statistic used in the fitness function Eq. (7). (The KS-statistic between two distributions corresponds to the largest absolute difference between the two CDFs.) For both summary statistics, the CDFs of model results on the test data (opaque lines) closely correspond to the corresponding results on the training data (transparent lines). This shows that the basic scanpath characteristics of the models are, just as in the human data, robust to the exact scene content.

We observe that the distribution of human foveation durations is wider and has a longer tail than the model scanpaths. The *L.hl* model provides the worst fit and has a maximum foveation duration of 600 ms. In the human data, $\sim 22.5\%$ of foveation events are longer than 600 ms, which is mainly due to long smooth pursuit events. The exponential distribution of the saccade amplitudes is well approximated by all models except for *O.cb* (Fig. 4d). Since the center bias always attributes high relevance to the middle of the screen, irrespective of the scene content, this model is strongly discouraged from exploring parts of the scene with higher eccentricities, leading to shorter saccades.

The parameter space of the last generation of the optimization for each model (Figs. S3-S6 in the *Supplemental Material*) shows that most parameters converge to a common value with low variance. This indicates stability and a desired robustness of the model predictions to small changes in the parameters. Since our fitness function $\mathscr{F}$ in Eq. (7) does not include information about the choices of saccade targets, how fast a target is revisited only has a subordinate effect on the fitness. The time course of IOR, parameterized by r, therefore, is not well constrained by the optimization objective. The fitness function $\mathscr{F}$ allows for a trade-off between a good approximation of the foveation duration distribution (low d_{FD}) or saccade amplitude (low d_{SA}). Depending on the model, we observe that the evolutionary parameter optimization can achieve a better overall fitness $\mathscr{F}$ by converging to parameters that primarily improve the fit in d_{FD} (e.g., *O.cb*, with mean and standard deviation across the last generation of $d_{FD} = 0.123 \pm 0.008$ and $d_{SA} = 0.189 \pm 0.008$) or in d_{SA} (e.g., *L.hl*, with $d_{FD} = 0.210 \pm 0.008$ and $d_{SA} = 0.068 \pm 0.008$), or balance out both contributions (e.g., *L.ll*, with $d_{FD} = 0.11 \pm 0.009$ and $d_{SA} = 0.092 \pm 0.011$). Parameter fits from the *O.ll* model achieve the best overall fitness ($d_{FD} = 0.133 \pm 0.010$ and $d_{SA} = 0.055 \pm 0.010$).

We conclude that both the object- and location-based models are capable of reproducing basic scanpath statistics.

Functional scanpath evaluation

Next, we assessed to which extent different model implementations can reflect human exploration behavior during scene viewing. This requires a scanpath metric that is applicable to dynamic scenes. It should consider the timing of eye movements and smooth pursuit events and be sensitive to the function that an eye movement may subserve in the context of the previous scanpath history (see Sec. *Evaluation of simulated scanpaths* in *Discussion* for a detailed analysis of other potential metrics).

Here we evaluate the scanpaths of the fitted models based on four different functional categories of foveation events (Linka and de Haas, 2021): An eye movement can either explore the background of a scene (*Background*), uncover a new object for the first time for foveal processing (*Detection*), further explore details of a previously detected object by making a within-object saccade (*Inspection*), or return to a previously uncovered object (*Revisit*). Depending on the foveated object ID (cf. Sec. *Dataset*

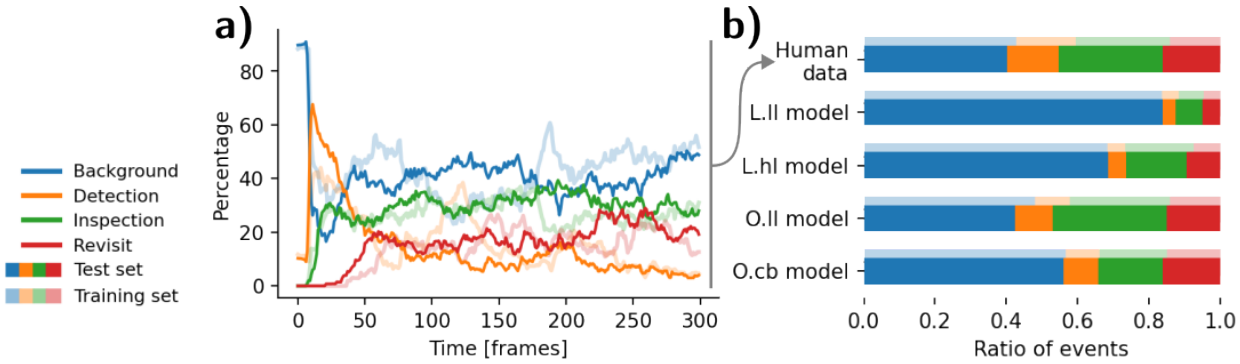

Figure 5. Functional decomposition of scanpaths in the foveation categories "Background" (blue), "Detection" (orange), "Inspection" (green), and "Revisits" (red). Data plotted in opaque (transparent) colors represents the test (training) set. (a) Percentage of foveation events in each category across all human scanpaths as a function of time. Plotted is the proportion of scanpaths that are in each respective category. (Time spent during saccades or noise in the eye tracking data are not considered and proportions are normalized to 100%.) (b) Proportion of time spent in each category averaged across all scanpaths from the human observers (data in panel (a) averaged over time) and for each of the four models averaged across all parameter configurations of the last generation of the optimization process.

in *Methods*), we assigned each foveation event to one of these four categories. This allowed us to characterize human scanpaths by analyzing what percentage of foveation time in the ground truth scanpaths was spent on each category (Fig. 5a). The first foveation predominantly explored the background since participants were instructed to look at the screen center at the beginning of each video, which corresponded to the scene background in most cases. Participants then quickly moved away from the background and started exploring the scene by detecting objects. The percentage of detection events decreases over time as more and more objects in the scene are detected, and already detected objects are more often inspected and revisited. With longer presentation times, the ratio of time spent exploring the background vs. inspecting and revisiting objects is nearly constant. Both the test (opaque lines) and training data (transparent lines) show this general trend, while specific events in the videos cause fluctuations. This overall trend is preserved in the time course of the foveation category percentage for the models (Fig. S7 in the *Supplemental Material*), even though they show strong differences in viewing tendencies.

For each foveation category, we calculated the ratios of time spent across the whole video duration, allowing for a comparison of the human data and the four different models (Fig. 5b). The simulated scanpaths demonstrate a strong similarity in the balance of different foveation categories between the training and test data. Together with the consistent prediction of basic summary statistics (Figs. 4c and 4d), this suggests that the models are able to effectively adapt to unseen videos, despite the wide variation in scene content. We find that the *L.ll* model spends more than 80% of the time exploring the background, which is about twice the time human participants spent on it. The models *L.hl*, *O.ll*, and – surprisingly – also *O.cb*, match the time spent in each category more closely. A qualitative comparison of the simulated scanpaths of the fitted models *L.ll* and *O.ll* additionally demonstrates the more human-like target selection of the object-based model (simulated scanpaths as GIFs shown in Figs. S8-S11 in the *Supplemental Material*)

To facilitate comparison, we expressed the time spent in each category for all four models relative to the human participants (Fig. 6). It is immediately apparent from the figure that all models show a significantly lower value for the time spent on detection events compared to the human data. The main reason for this discrepancy is that in the human eye tracking data, more than 20% of the stimulus presentation time is not spent on foveation events but on saccades, blinks, or is classified as noise due to low data quality. In contrast, scanpaths simulated with the *ScanDy* framework do not assign any time to saccade or noise events and hence spend then complete stimulus presentation time on foveation events. This effectively longer foveation time, however, does not contribute to more detection events, because each object in a scene can only be detected once. We do not find a qualitative difference in results on the training set (transparent) compared to the previously unseen videos of the test set

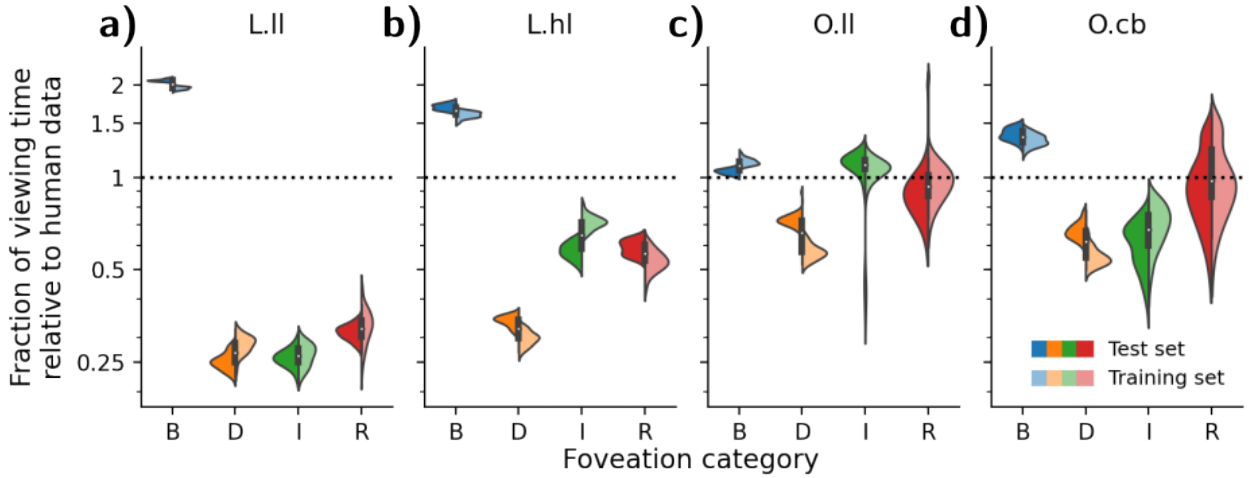

Figure 6. Violin plots of the ratio of the fraction of time spent in each foveation category predicted by the models and the fraction of time spent in each foveation category by the human observers. We show side-by-side the results for the videos of the test (left, opaque) and the training data (right, transparent) on a logarithmic scale. A ratio of one (dotted line) corresponds to a perfect prediction of how humans balance their exploration behavior. The four categories (**B**ackground, **D**etection, **I**nspection, **R**evisit) are shown on the x-axis. Individual data points for each model correspond to the 32 best parameter configurations (i.e., the last generation) of the evolutionary optimization and show the average value across twelve scanpaths per video in the respective dataset. Panels show the location-based models with low-level features *L.ll* (a), and with high-level features *L.hl* (b), and the object-based models with low-level features *O.ll* (c), and with the center bias only *O.cb* (d).

(Fig. 6)

The *L.ll* model (Fig. 6a) spends most time exploring the background and only half the time exploring objects in the scene (see also Fig. 5). Without object representations, this model cannot extract the parts of the scene that humans spend the most time looking at. Replacing the saliency map based on low-level features (Molin, Etienne-Cummings, and Niebur, 2015) with the TASED-Net (Min and Corso, 2019) saliency prediction (i.e., model *L.hl*) leads to a significant improvement in the plausibility of the generated scanpaths (Fig. 6b). This model spends more time exploring the objects, especially as inspections and revisits. The time foveating the background is reduced compared to the implementation with low-level features, though still $\sim 70\%$ higher than in the human data. This model spending less time foveating objects than the human scanpaths might be surprising since the high-level saliency model is highly sensitive to salient objects (Min and Corso, 2019) and tends to only ascribe small saliency values to the background. The high saliency values at locations that coincide with the object masks lead to a large number of short foveations within objects. The optimization objective, however, forces the model to also make long saccades away from these objects following the saccade amplitude distribution (Fig. 4b), which – at least in part – leads to foveations on the background. If the gaze position moves far away from clusters of high saliency, the Gaussian decay of sensitivity means that it takes a long time until the decision threshold is reached again. Detailed inspection of the number and duration of foveation events in the background (data not shown) confirms that this model does only a few but very long foveation events in the background. The detection events in the *L.hl* model are more strongly underrepresented compared to inspections and revisits, since the high-level saliency leads to a large number of very short foveations on objects and each object can only be detected once. In summary, this shows that although the high-level saliency map predicts well where people look on average, the location-based target selection leads to significantly different scanpaths in the *L.hl* model compared to the human ground truth.

In fact, saliency seems to play only a secondary role in the simulation of realistic scanpaths. Even without a saliency map, the object-based model *O.cb* produces a reasonable foveation behavior (Fig. 6d). Unless objects are placed directly in the center, this model by construction assigns the majority of fixation time to the background of the scene, leading to an overrepresentation of $\sim 39\%$ on average compared to the human data. The ratio of inspections is $\sim 37\%$ lower compared to the ground truth, but the revisits match, on average, the human gaze behavior. Detection events are about 30% reduced, which is similar to what we

would expect due to the roughly 20% larger effective foveation time due to not spending time on saccades and noise events. Although the "relevance" of objects in this model is only determined by their distance to the center of the display, it leads to a more balanced exploration behavior compared to both location-based models, *L.ll* and *L.hl*, which use state-of-the-art saliency models.

The average exploration of the background in the *O.ll* model is $\sim 5\%$ higher than in the human data (Fig. 6c). This slight overrepresentation is clearly less pronounced compared to the other models and is likely due to a small overestimation of foveation duration when the gaze is far away from all foreground objects. The time spent revisiting events in both object-based models varies strongly between the different parameter fits (suggesting that the optimization objective in Eq. (7) does not adequately constrain the inhibition of return parameters) but is on average close to the human data. In conclusion, we obtain the most realistic gaze behavior by adding low-level saliency information to the object-based model and using it to prioritize between potential target objects.

Discussion

Advantages and limitations of our mechanistic modeling framework for scanpaths in dynamic scenes

We present a general computational modeling framework for simulating human exploration behavior for ecologically valid scenes. By implementing psychophysically uncovered mechanisms of attentional and oculomotor control, *ScanDy* allows to generate sequences of eye movements for any visual scene. Recent years have shown a growing interest in the simulation of time-ordered fixation sequences for static scenes (Tatler, Brockmole, and Carpenter, 2017; Malem-Shinitski et al., 2020; Schwetlick, Rothkegel, et al., 2020; Schwetlick, Backhaus, and Engbert, 2022; Kucharsky et al., 2021; Kümmerer, Bethge, and Wallis, 2022), as well as the frame-wise prediction of where humans tend to look on average when observing a dynamic scene (Molin, Etienne-Cummings, and Niebur, 2015; Min and Corso, 2019; Droste, Jiao, and Noble, 2020; Wang, Liu, et al., 2021). We are currently not aware of another computational model that is able to simulate time-resolved gaze positions for the full duration of dynamic scenes, analogous to human eye tracking data.

Scanpaths of multiple observers on dynamic scenes with gaze positions for each point in time are rich data, making their simulation a difficult task for any computational model. We achieve realistic gaze behavior in this work with mechanistic, well-interpretable models with only five free parameters. Achieving this simplicity and interpretability requires major simplifications and the here presented modeling framework does by far not contain all attentional mechanisms uncovered by psychophysical experiments.

One strongly simplified assumption we make in modeling the object-based visual sensitivity is that attention spreads instantaneously and uniformly across an object. Neurophysiological and behavioral studies have shown that attention spreads serially inside objects, which can be well described with a "growth-cone" model of perceptual grouping (Pooresmaeili and Roelfsema, 2014). Especially since the currently established growth-cone model does not yet seem to generalize to real-world scenes (Jeurissen, Self, and Roelfsema, 2016), implementing and modifying it within the *ScanDy* framework seems to be a promising approach to investigate the object-based spread of attention.

Another simplifying assumption about visual sensitivity in the models presented here is that visual attention just spreads around the current gaze position. Experimental evidence, however, clearly shows that attentional resources are covertly allocated away from the current gaze position right before (e.g., Rolfs, Jonikaitis, et al. (2011)) and after (e.g., Golomb et al. (2011)) a saccade is executed. Extending the *SceneWalk* model (Engbert et al., 2015) by such perisaccadic covert attention shifts improved the predictions of scanpaths, especially with respect to angular statistics (Schwetlick, Rothkegel, et al., 2020). The *ScanDy* framework can be extended analogously by including perisaccadic shifts in the sensitivity map S (see Sec. *Visual sensitivity (II)* in *Methods: ScanDy*).

In this work, we model the saccadic decision-making process and assume that a selected target is accurately reached within one time step, i.e. the duration of one video frame. For simplicity, we explicitly do not attempt to model the saccade programming and oculomotor control to execute this movement. There have been various models proposed in the literature

for the velocity profile of a saccade based on its properties like amplitude and turning angle (Fuhl and Kasneci, 2018; Specht et al., 2017; Dai et al., 2016; Han et al., 2013; Schweitzer and Rolfs, 2022). Including this in our framework could be a valuable addition and might contribute to our understanding of perceptual continuity. We currently assume that objects can be tracked across time irrespective of the large and rapid shift of objects across the retina caused by saccades. A continuous and stable perception of the environment despite these shifts (Rolfs, 2015) could potentially be explained within our framework by modeling either attentional updating across saccades (Rolfs, Jonikaitis, et al., 2011) or the incidental visual consequences of saccades (Rolfs and Schweitzer, 2022), during which an object's retinal trace leads to motion streaks that can facilitate object correspondence across consecutive fixations(Schweitzer and Rolfs, 2021).

As a consequence of the small number of parameters, the model optimization within *ScanDy* is highly data efficient. This allows, in principle, for fitting individual differences by inferring model parameters for each human participant separately and systematically analyzing the variability in scanpaths (Dorr et al., 2010). The fixation tendencies between observers have been shown to vary strongly between individuals (de Haas et al., 2019; Broda and de Haas, 2022). We hope that with larger, high-quality eye tracking datasets of dynamic real-world scenes, *ScanDy* can become a useful tool to investigate individual differences in visual exploration behavior.

So far, we only attempt to simulate scanpaths during free-viewing and hence do not include any top-down attentional control. In addition to bottom-up attentional guidance, there have been proposals for cognitive models of top-down visual guidance through task-driven attention (Navalpakkam and Itti, 2005; Borji, Sihite, and Itti, 2011). For future work, we expect that models within *ScanDy* could also be optimized to reproduce gaze behavior in visual search tasks by including guidance maps as feature map F (see Sec. *Scene features (I)* in *Methods: ScanDy*) based on the target's specific features or appearance (Wolfe, Cave, and Franzel, 1989; Wolfe, 1994; Zelinsky, 2008).

Our current model realizations within the *ScanDy* framework should be seen as easily modifiable baseline models. Depending on the specific interest in certain attentional mechanisms, models can be extended within the same framework. Analogous to our interest in the impact of object representations on attentional selection, we hope that the community can use *ScanDy* to implement different model variants based on competing assumptions and compare the resulting simulated scanpaths to evaluate hypotheses about gaze control and attentional selection.

Evaluation of simulated scanpaths

The comparison of model scanpaths with human data is an essential way to validate our hypotheses and model assumptions. Scanpaths on static scenes s_s usually consist of a short, ordered list of fixation positions (x,y), with $s_s \in \mathbb{R}^{(2,N_F)}$, and N_F being the number of fixations during the presentation time of the image. The saccade timing is typically not considered, since the scene content at each fixation does not change. This is, by definition, different in dynamic scenes. In videos, it is always necessary to specify when the gaze was at a certain location since potential object and camera movement could influence what there is to see at this position. As a further complication, the gaze position can not only change due to saccades and (often ignored) fixational eye movements, as is the case for static scenes. Smooth pursuit plays an important role when observing dynamic scenes and in everyday life (Land, 2006). In this work, we combine fixation and smooth pursuit events, as both keep the current target foveated and are bounded by saccadic decisions. This means, however, that a foveation event in dynamic scenes does not have a single location, but that this location can (and often does) change over time. We, therefore, characterize a scanpath s_d in dynamic scenes as a list of all gaze positions (x,y,t) with a given time resolution, which means that $s_d \in \mathbb{R}^{(3,N_T)}$. Here we choose the number N_T, with $N_T \gg N_F$, of time steps equal to the number of frames in the video. The dimensionality of s_d, which is much higher compared to the dimensionality of s_s, gives rise to an immense variability in human scanpaths when observing dynamic scenes. Hence we do not expect the models to exactly reproduce these scanpaths, but to capture characteristic statistical features of it.

Thus far, scanpath simulation on videos is an under-researched topic, and this is reflected in the lack of scanpath evaluation methods that are applicable to dynamic scenes. To the best of our knowledge, there is no established metric to evaluate the realism of simulated scanpaths which captures both their spatial and temporal aspects and which does not assume static input

scenes. In the following, we discuss metrics from video saliency prediction and from scanpath prediction on static scenes and elaborate on why they do not capture important aspects of the scanpath or are hardly applicable to dynamic scenes.

The overwhelming majority of work on modeling visual attention has treated attentional selection as a static problem, where gaze positions of multiple observers are described by a spatial distribution. There are multiple established methods to compare gaze data in the form of spatial distribution maps, as summarized by Kümmerer, Wallis, and Bethge (2015) for saliency predictions on images. These methods, however, typically ignore that attention is an inherently sequential process and do not consider any temporal aspects of the scanpath. There are extensions to also study spatiotemporal effects modeling fixation locations as arising from point processes (Barthelmé et al., 2013; Engbert et al., 2015), e.g., by predicting the conditional probability of a saccade target given previous fixation locations. Calculating and evaluating the spatiotemporal likelihood function for all scanpaths and updating them in each frame would, in practice, already be computationally infeasible. But also conceptually, the duration of foveation events is typically not considered in these approaches (although an extension to labeled point processes by annotating points with their duration might be possible), and it is not clear how point processes might be generalized to include the change in gaze position during smooth pursuit events.

To evaluate video saliency models (Wang, Shen, et al., 2018) or for quantifying the variability of human eye movements in dynamic scenes (Dorr et al., 2010), frame-wise spatial distribution metrics have been used. The ground truth for these metrics for one video is computed by smoothing over all gaze positions from all human observers with a spatial or spatiotemporal Gaussian and creating a spatial ground truth distribution map for each frame. Due to averaging across different scanpaths, the resulting maps only provide information about where people are most likely to look, but not how individuals decide to move their eyes. The best-performing model under such a metric, like Normalized Scanpath Saliency (NSS), would consequently just stay foveated on the most salient part of the scene for the full duration of the video. Although humans might do that *on average*, in this work, we are interested in the way how humans actually move their eyes to explore the scene.

A well-established way to evaluate scanpaths on static scenes is using string-based metrics (e.g. Noton and Stark (1971)). The basic idea is to convert scanpaths to strings by defining a number of spatial regions in the scene, for example, in the form of a checkerboard, which are represented by letters. Each fixation is then assigned the letter of the spatial region it lies in, such that the full scanpath can be represented by a string. Fixation durations can be taken into account by sampling the gaze position in fixed temporal bins. The comparison between two scanpaths can then be quantified using string edit metrics like the Levenshtein distance (Levenshtein et al., 1966) or the ScanMatch method (Cristino et al., 2010). In dynamic real-world scenes, it is, however, not clear how the spatial regions should be defined. If the region boundaries are static, this means that the scene content within each region can change over time, making the encoding uninformative about what humans actually looked at. This problem can only be avoided if the region boundaries change over time to capture the movement in the scene, for example, by defining them based on objects. In this work, we chose a related approach to quantify how realistic the simulated scanpaths are.

To obtain realistic scanpaths, we prioritize reproducing the global statistics of human scanpaths in the model results. We show that the simulated scanpaths capture the human foveation duration and saccade amplitude distributions. Additional statistics, like saccade angle distributions, become especially interesting if additional modifications within the *ScanDy* framework are explored (for example, perisaccadic attention shifts as in Schwetlick, Rothkegel, et al. (2020)). Such global statistics are, however, agnostic about where and when saccades are initiated *in relation to the scene content*. Since we use global scanpath statistics to fit the free model parameters, we expect all models to be able to reproduce the foveation duration and saccade amplitude distributions to a certain extent. A good free-viewing scanpath model should, when being tuned to reproduce these statistics, have implicit properties that lead to realistic exploration behavior given a visual scene.

In this contribution, we focus on investigating the role that objects play in attentional selection. It is, therefore, natural to use an evaluation metric sensitive to the extent and the particular way objects are visually explored. The approach for analyzing scanpaths proposed by Linka and de Haas (2021) provides an intuitive way to quantify how and with what underlying function objects in the scene are selected. Although initially intended for static scenes, it naturally generalizes to dynamic scenes because

it is based on objects. Importantly, we extend the three originally proposed categories (Detection, Inspection, and Revisit) by a fourth Background category in order to avoid a biased evaluation that would only consider foveation events on objects. For us, this approach combines the benefits of string-based methods and dynamic region boundaries while additionally providing information about the extent to which objects are detected, inspected, or revisited. For applications where it is important also to analyze which objects are foveated in which order, this metric only provides limited insights. We hope that with growing interest in simulating scanpaths for videos, more refined evaluation benchmarks will be established.

The role of objects in attentional selection

Our main interest was to use this computational framework to assess the importance of object-level attentional units for the guidance of human eye movements. Experimental research on object-based attention has repeatedly shown that objects play a central role in attentional selection (Scholl, 2001; Cavanagh et al., 2022) (cf. Sec. *Objects* in *Methods: ScanDy*). These insights did, apart from some notable exceptions (e.g., Yanulevskaya et al. (2013), Sun and Fisher (2003), and Yu et al. (2015)), not yet translate into computational models, where the idea of location-/pixel-based saliency maps as the basis for attentional selection is still dominating. Our results, however, clearly show that models with object-based attention and selection lead to more realistic scanpaths.

We have only compared two extreme views so far by contrasting two model families that are either completely object-based or do not have any object representations. It could, for example, also be that while the spread of visual sensitivity and target selection should be modeled based on objects, it is beneficial to assume a spatiotopic implementation of IOR (as suggested by Krüger and Hunt (2013)). Exploring such hybrid models is straightforward to do within the *ScanDy* framework (see *Framework architecture* in *Methods: ScanDy*) and might be fruitful to explore in the future.

Our modeling results strongly suggest that objects are an important unit of perceptual selection and play a crucial role in guiding human gaze behavior. How exactly these object representations are formed in the brain is still an open question (Cichy, Pantazis, and Oliva, 2014; Bonner and Epstein, 2021). We follow the argumentation of Cavanagh et al. (2022) and assume that object-based attention can act on fully evolved object representations already before they are selected as saccade targets. An interesting alternative to providing predefined object segmentation masks would be an implementation where object representations and attentional selection inform and influence each other. Optimizing object segmentation to provide realistic scanpaths would allow us to avoid defining what constitutes a visual object and could result in representations somewhere between locations and fully formed semantic objects.

Conclusion

With *ScanDy*, we provide a modular and mechanistic framework for predicting scanpaths in dynamic real-world scenes. Despite its simplicity and the low number of parameters, our results show that models within this framework can reasonably reproduce spatial and temporal aspects of human scanpaths when watching natural videos. Comparing different models implementing competing model assumptions shows that the most human-like visual exploration behavior is obtained when object-based attention and object-based target selection are combined with saliency-based prioritization of saccade targets. *ScanDy* is an open-source project https://github.com/rederoth/ScanDy and due to its modular implementation, it can easily be extended and experimented with. We hope its availability will contribute to the field shifting further towards realistic environments and motivate more research on attention allocation in dynamic real-world scenes.

Author contributions

Nicolas Roth: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft Preparation, Writing - Review & Editing, Visualization **Martin Rolfs:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition **Olaf Hellwich:** Conceptualization, Writing - Review & Editing, Funding acquisition **Klaus Obermayer:** Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135. M.R. was supported by the Heisenberg program of the Deutsche Forschungsgemeinschaft (DFG grants RO3579/8-1 and RO3579/12-1).

References

1. Buswell GT. How people look at pictures: a study of the psychology and perception in art. 1935
2. Yarbus AL. Eye movements and vision. Springer New York, NY, 1967
3. Itti L and Koch C. Computational modelling of visual attention. Nature reviews neuroscience 2001; 2:194–203
4. Borji A and Itti L. State-of-the-art in visual attention modeling. IEEE transactions on pattern analysis and machine intelligence 2012; 35:185–207
5. Itti L, Koch C, and Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence 1998; 20:1254–9
6. Koch C and Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 1985; 4:219–27
7. Engbert R, Trukenbrod HA, Barthelmé S, and Wichmann FA. Spatial statistics and attentional dynamics in scene viewing. Journal of vision 2015; 15:14–4
8. Schütt HH, Rothkegel LO, Trukenbrod HA, Reich S, Wichmann FA, and Engbert R. Likelihood-based parameter estimation and comparison of dynamical cognitive models. Psychological Review 2017; 124:505
9. Schwetlick L, Rothkegel LOM, Trukenbrod HA, and Engbert R. Modeling the effects of perisaccadic attention on gaze statistics during scene viewing. Communications biology 2020; 3:1–11
10. Schwetlick L, Backhaus D, and Engbert R. A dynamical scan-path model for task-dependence during scene viewing. Psychological Review 2022
11. Malem-Shinitski N, Opper M, Reich S, Schwetlick L, Seelig SA, and Engbert R. A mathematical model of local and global attention in natural scene viewing. PLoS Computational Biology 2020; 16:e1007880
12. Kümmerer M, Bethge M, and Wallis TS. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. Journal of Vision 2022; 22:7–7
13. Kümmerer M and Bethge M. State-of-the-art in human scanpath prediction. arXiv preprint arXiv:2102.12239 2021
14. Henderson JM, Weeks Jr PA, and Hollingworth A. The effects of semantic consistency on eye movements during complex scene viewing. Journal of experimental psychology: Human perception and performance 1999; 25:210
15. Underwood G and Foulsham T. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. Quarterly Journal of Experimental Psychology 2006; 59:1931–49
16. Nuthmann A, Smith TJ, Engbert R, and Henderson JM. CRISP: a computational model of fixation durations in scene viewing. Psychological review 2010; 117:382
17. Kucharsky Š, Renswoude D van, Raijmakers M, and Visser I. WALD-EM: Wald accumulation for locations and durations of eye movements. Psychological Review 2021; 128:667
18. Tatler BW, Brockmole JR, and Carpenter RH. LATEST: A model of saccadic decisions in space and time. Psychological review 2017; 124:267

19. Noorani I and Carpenter R. The LATER model of reaction time and decision. Neuroscience & Biobehavioral Reviews 2016; 64:229–51

20. Goldstein RB, Woods RL, and Peli E. Where people look when watching movies: Do all viewers look at the same place? Computers in biology and medicine 2007; 37:957–64

21. Dorr M, Martinetz T, Gegenfurtner KR, and Barth E. Variability of eye movements when viewing dynamic natural scenes. Journal of vision 2010; 10:28–8

22. Smith TJ and Mital PK. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. Journal of vision 2013; 13:16–6

23. Carmi R and Itti L. Visual causes versus correlates of attentional selection in dynamic scenes. Vision research 2006; 46:4333–45

24. Itti L. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition 2005; 12:1093–123

25. Mital PK, Smith TJ, Hill RL, and Henderson JM. Clustering of gaze during dynamic scene viewing is predicted by motion. Cognitive computation 2011; 3:5–24

26. Niebur E and Koch C. Control of selective visual attention: Modeling the" where" pathway. Advances in neural information processing systems 1995; 8

27. Le Meur O, Le Callet P, and Barba D. Predicting visual fixations on video based on low-level visual features. Vision research 2007; 47:2483–98

28. Molin JL, Etienne-Cummings R, and Niebur E. How is motion integrated into a proto-object based visual saliency model? *2015 49th Annual Conference on Information Sciences and Systems (CISS)*. IEEE. 2015 :1–6

29. Marszalek M, Laptev I, and Schmid C. Actions in context. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009 :2929–36

30. Wang W, Shen J, Guo F, Cheng MM, and Borji A. Revisiting video saliency: A large-scale benchmark and a new model. *Proceedings of the IEEE Conference on computer vision and pattern recognition*. 2018 :4894–903

31. Linardos P, Mohedano E, Nieto JJ, O'Connor NE, Giro-i-Nieto X, and McGuinness K. Simple vs complex temporal recurrences for video saliency prediction. arXiv preprint arXiv:1907.01869 2019

32. Droste R, Jiao J, and Noble JA. Unified image and video saliency modeling. *European Conference on Computer Vision*. Springer. 2020 :419–35

33. Min K and Corso JJ. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019 :2394–403

34. Jain S, Yarlagadda P, Jyoti S, Karthik S, Subramanian R, and Gandhi V. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021 :3520–7

35. Wang Z, Liu Z, Li G, Wang Y, Zhang T, Xu L, and Wang J. Spatio-Temporal Self-Attention Network for Video Saliency Prediction. IEEE Transactions on Multimedia 2021

36. Li Y, Fathi A, and Rehg JM. Learning to predict gaze in egocentric video. *Proceedings of the IEEE international conference on computer vision*. 2013 :3216–23

37. Huang Y, Cai M, Li Z, and Sato Y. Predicting gaze in egocentric video by learning task-dependent attention transition. *Proceedings of the European conference on computer vision (ECCV)*. 2018 :754–69

38. Tavakoli HR, Rahtu E, Kannala J, and Borji A. Digging deeper into egocentric gaze prediction. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019 :273–82

39. Nuthmann A and Henderson JM. Object-based attentional selection in scene viewing. Journal of vision 2010; 10:20–0

40. Pajak M and Nuthmann A. Object-based saccadic selection during scene perception: Evidence from viewing position effects. Journal of vision 2013; 13:2–2

41. Einhäuser W, Spain M, and Perona P. Objects predict fixations better than early saliency. Journal of vision 2008; 8:18–8

42. Egly R, Driver J, and Rafal RD. Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. Journal of Experimental Psychology: General 1994; 123:161

43. Kahneman D, Treisman A, and Gibbs BJ. The reviewing of object files: Object-specific integration of information. Cognitive psychology 1992; 24:175–219

44. O'Craven KM, Downing PE, and Kanwisher N. fMRI evidence for objects as the units of attentional selection. Nature 1999; 401:584–7

45. Scholl BJ. Objects and attention: The state of the art. Cognition 2001; 80:1–46

46. Cavanagh P, Caplovitz GP, Lytchenko TK, Maechler M, Peter UT, and Sheinberg D. Object-Based Attention. 2022

47. Walther D and Koch C. Modeling attention to salient proto-objects. Neural networks 2006; 19:1395–407

48. Yanulevskaya V, Uijlings J, Geusebroek JM, Sebe N, and Smeulders A. A proto-object-based computational model for visual saliency. Journal of vision 2013; 13:27–7

49. Russell AF, Mihalaş S, Heydt R von der, Niebur E, and Etienne-Cummings R. A model of proto-object based saliency. Vision research 2014; 94:1–15

50. Yu JG, Xia GS, Gao C, and Samal A. A computational model for object-based visual saliency: Spreading attention along gestalt cues. IEEE Transactions on Multimedia 2015; 18:273–86

51. Sun Y and Fisher R. Object-based visual attention for computer vision. Artificial intelligence 2003; 146:77–123

52. Li Z, Qin S, and Itti L. Visual attention guided bit allocation in video compression. Image and Vision Computing 2011; 29:1–14

53. Martinez-Conde S, Macknik SL, and Hubel DH. The role of fixational eye movements in visual perception. Nature reviews neuroscience 2004; 5:229–40

54. Rolfs M. Microsaccades: small steps on a long way. Vision research 2009; 49:2415–41

55. Rucci M and Victor JD. The unsteady eye: an information-processing stage, not a bug. Trends in neurosciences 2015; 38:195–206

56. Clarke AD and Tatler BW. Deriving an appropriate baseline for describing fixation behaviour. Vision research 2014; 102:41–51

57. Sun D, Yang X, Liu MY, and Kautz J. PWC-net: CNNs for optical flow using pyramid, warping, and cost volume. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018 :8934–43

58. Spelke ES. Principles of object perception. Cognitive science 1990; 14:29–56

59. Palmeri TJ and Gauthier I. Visual object understanding. Nature Reviews Neuroscience 2004; 5:291–303

60. Scholl BJ, Pylyshyn ZW, and Feldman J. What is a visual object? Evidence from target merging in multiple object tracking. Cognition 2001; 80:159–77

61. Feldman J. What is a visual object? Trends in Cognitive Sciences 2003; 7:252–6

62. Rensink RA. The dynamic representation of scenes. Visual cognition 2000; 7:17–42

63. Deeprose C and Andrade J. Is priming during anesthesia unconscious? Consciousness and cognition 2006; 15:1–23

64. Tanaka K, Saito Ha, Fukada Y, and Moriya M. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. Journal of neurophysiology 1991; 66:170–89

65. Ansorge U, Kunde W, and Kiefer M. Unconscious vision and executive control: How unconscious processing and conscious action control interact. Consciousness and cognition 2014; 27:268–87

66. Chou WL and Yeh SL. Object-based attention occurs regardless of object awareness. Psychonomic bulletin & review 2012; 19:225–31

67. Huang TH, Yeh SL, Yang YH, Liao HI, Tsai YY, Chang PJ, and Chen HH. Method and experiments of subliminal cueing for real-world images. Multimedia Tools and Applications 2015; 74:10111–35

68. Neri P. Object segmentation controls image reconstruction from natural scenes. PLoS biology 2017; 15:e1002611

69. Borji A and Tanner J. Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations. IEEE transactions on neural networks and learning systems 2015; 27:1214–26

70. He K, Gkioxari G, Dollár P, and Girshick R. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*. 2017 :2961–9

71. Wu Y, Kirillov A, Massa F, Lo WY, and Girshick R. Detectron2. https://github.com/facebookresearch/detectron2. 2019

72. Wojke N, Bewley A, and Paulus D. Simple online and realtime tracking with a deep association metric. *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017 :3645–9

73. Malcolm GL and Shomstein S. Object-based attention in real-world scenes. Journal of Experimental Psychology: General 2015; 144:257

74. Tipper SP, Driver J, and Weaver B. Object-centred inhibition of return of visual attention. The Quarterly Journal of Experimental Psychology 1991; 43:289–98

75. Parkhurst D and Niebur E. Scene content selected by active vision. Spatial vision 2003; 16:125–54

76. Tatler BW, Baddeley RJ, and Gilchrist ID. Visual correlates of fixation selection: Effects of scale and time. Vision research 2005; 45:643–59

77. Tatler BW. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of vision 2007; 7:4–4

78. Tseng PH, Carmi R, Cameron IG, Munoz DP, and Itti L. Quantifying center bias of observers in free viewing of dynamic natural scenes. Journal of vision 2009; 9:4–4

79. Nuthmann A, Einhäuser W, and Schütz I. How well can saliency models predict fixation selection in scenes beyond central bias? A new approach to model evaluation using generalized linear mixed models. Frontiers in human neuroscience 2017; 11:491

80. Basler A. Über das Sehen von Bewegungen. Archiv für die gesamte Physiologie des Menschen und der Tiere 1906; 115:582–601

81. Eriksen CW and Colegate RL. Selective attention and serial processing in briefly presented visual displays. Perception & Psychophysics 1971; 10:321–6

82. Posner MI. Orienting of attention. Quarterly journal of experimental psychology 1980; 32:3–25

83. Egeth HE and Yantis S. Visual attention: Control, representation, and time course. Annual review of psychology 1997; 48:269–97

84. Posner MI, Cohen Y, et al. Components of visual orienting. Attention and performance X: Control of language processes 1984; 32:531–56

85. Klein RM. Inhibition of return. Trends in cognitive sciences 2000; 4:138–47

86. Klein RM and Ivanoff J. Inhibition of return. Scholarpedia 2008; 3:3650

87. Klein R. Inhibitory tagging system facilitates visual search. Nature 1988; 334:430–1

88. Klein RM and MacInnes WJ. Inhibition of return is a foraging facilitator in visual search. Psychological science 1999; 10:346–52

89. Posner MI, Rafal RD, Choate LS, and Vaughan J. Inhibition of return: Neural basis and function. Cognitive neuropsychology 1985; 2:211–28

90. Maylor EA and Hockey R. Inhibitory component of externally controlled covert orienting in visual space. Journal of Experimental Psychology: Human Perception and Performance 1985; 11:777

91. Bennett PJ and Pratt J. The spatial distribution of inhibition of return. Psychological Science 2001; 12:76–80

92. Berlucchi G, Tassinari G, Marzi C, and Di Stefano M. Spatial distribution of the inhibitory effect of peripheral non-informative cues on simple reaction time to non-fixated visual targets. Neuropsychologia 1989; 27:201–21

93. Klein RM, Christie J, and Morris EP. Vector averaging of inhibition of return. Psychonomic Bulletin & Review 2005; 12:295–300

94. Samuel AG and Kat D. Inhibition of return: A graphical meta-analysis of its time course and an empirical test of its temporal and spatial properties. Psychonomic bulletin & review 2003; 10:897–906

95. Gibson BS and Egeth H. Inhibition of return to object-based and environment-based locations. Perception & Psychophysics 1994; 55:323–39

96. Ro T and Rafal RD. Components of reflexive visual orienting to moving objects. Perception & Psychophysics 1999; 61:826–36

97. Becker L and Egeth H. Mixed reference frames for dynamic inhibition of return. Journal of Experimental Psychology: Human Perception and Performance 2000; 26:1167

98. Jordan H and Tipper SP. Spread of inhibition across an object's surface. British Journal of Psychology 1999; 90:495–507

99. Reppa I and Charles Leek E. The modulation of inhibition of return by objectinternal structure: Implications for theories of object-based attentional selection. Psychonomic Bulletin & Review 2003; 10:493–502

100. Ludwig CJ. Saccadic decision-making. 2011

101. Ratcliff R. A theory of memory retrieval. Psychological review 1978; 85:59

102. Bogacz R, Brown E, Moehlis J, Holmes P, and Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review 2006; 113:700

103. Ratcliff R and McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural computation 2008; 20:873–922

104. Gold JI, Shadlen MN, et al. The neural basis of decision making. Annual review of neuroscience 2007; 30:535–74

105. Forstmann BU, Ratcliff R, and Wagenmakers EJ. Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. Annual review of psychology 2016; 67:641

106. Ratcliff R, Smith PL, Brown SD, and McKoon G. Diffusion decision model: Current issues and history. Trends in cognitive sciences 2016; 20:260–81

107. Cakan C, Jajcay N, and Obermayer K. neurolib: a simulation framework for whole-brain neural mass modeling. Cognitive Computation 2021 :1–21

108. Eiben AE, Smith JE, et al. Introduction to evolutionary computing. Vol. 53. Springer, 2003

109. Linka M and de Haas B. Detection, Inspection and Re-Inspection: A functional approach to gaze behavior towards complex scenes. Journal of Vision 2021; 21:1971–1

110. Startsev M, Agtzidis I, and Dorr M. 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. Behavior Research Methods 2019; 51:556–72

111. Treisman AM and Gelade G. A feature-integration theory of attention. Cognitive psychology 1980; 12:97–136

112. Pooresmaeili A and Roelfsema PR. A growth-cone model for the spread of object-based attention during contour grouping. Current Biology 2014; 24:2869–77

113. Jeurissen D, Self MW, and Roelfsema PR. Serial grouping of 2D-image regions with object-based attention in humans. Elife 2016; 5:e14320

114. Rolfs M, Jonikaitis D, Deubel H, and Cavanagh P. Predictive remapping of attention across eye movements. Nature neuroscience 2011; 14:252–6

115. Golomb JD, Marino AC, Chun MM, and Mazer JA. Attention doesn't slide: spatiotopic updating after eye movements instantiates a new, discrete attentional locus. Attention, Perception, & Psychophysics 2011; 73:7–14

116. Fuhl W and Kasneci E. Eye movement velocity and gaze data generator for evaluation, robustness testing and assess of eye tracking software and visualization tools. arXiv preprint arXiv:1808.09296 2018

117. Specht JI, Dimieri L, Urdapilleta E, and Gasaneo G. Minimal dynamical description of eye movements. The European Physical Journal B 2017; 90:1–12

118. Dai W, Selesnick I, Rizzo JR, Rucker J, and Hudson T. A parametric model for saccadic eye movement. *2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE. 2016 :1–6

119. Han P, Saunders DR, Woods RL, and Luo G. Trajectory prediction of saccadic eye movements using a compressed exponential model. Journal of vision 2013; 13:27–7

120. Schweitzer R and Rolfs M. Definition, Modeling, and Detection of Saccades in the Face of Post-saccadic Oscillations. *Eye Tracking: Background, Methods, and Applications*. Springer, 2022 :69–95

121. Rolfs M. Attention in active vision: A perspective on perceptual continuity across saccades. Perception 2015; 44:900–19

122. Rolfs M and Schweitzer R. Coupling perception to action through incidental sensory consequences of motor behaviour. Nature Reviews Psychology 2022; 1:112–23

123. Schweitzer R and Rolfs M. Intrasaccadic motion streaks jump-start gaze correction. Science Advances 2021; 7:eabf2218

124. de Haas B, Iakovidis AL, Schwarzkopf DS, and Gegenfurtner KR. Individual differences in visual salience vary along semantic dimensions. Proceedings of the National Academy of Sciences 2019; 116:11687–92

125. Broda MD and de Haas B. Individual fixation tendencies in person viewing generalize from images to videos. i-Perception 2022; 13:20416695221128844

126. Navalpakkam V and Itti L. Modeling the influence of task on attention. Vision research 2005; 45:205–31

127. Borji A, Sihite DN, and Itti L. Computational modeling of top-down visual attention in interactive environments. *BMVC*. Vol. 85. 2011 :1–12

128. Wolfe JM, Cave KR, and Franzel SL. Guided search: an alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human perception and performance 1989; 15:419

129. Wolfe JM. Guided search 2.0 a revised model of visual search. Psychonomic bulletin & review 1994; 1:202–38

130. Zelinsky GJ. A theory of eye movements during target acquisition. Psychological review 2008; 115:787

131. Land MF. Eye movements and the control of actions in everyday life. Progress in retinal and eye research 2006; 25:296–324

132. Kümmerer M, Wallis TS, and Bethge M. Information-theoretic model comparison unifies saliency metrics. Proceedings of the National Academy of Sciences 2015; 112:16054–9

133. Barthelmé S, Trukenbrod H, Engbert R, and Wichmann F. Modeling fixation locations using spatial point processes. Journal of vision 2013; 13:1–1

134. Noton D and Stark L. Scanpaths in eye movements during pattern perception. Science 1971; 171:308–11

135. Levenshtein VI et al. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*. Vol. 10. 8. Soviet Union. 1966 :707–10

136. Cristino F, Mathôt S, Theeuwes J, and Gilchrist ID. ScanMatch: A novel method for comparing fixation sequences. Behavior research methods 2010; 42:692–700

137. Krüger HM and Hunt AR. Inhibition of return across eye and object movements: The role of prediction. Journal of Experimental Psychology: Human Perception and Performance 2013; 39:735

138. Cichy RM, Pantazis D, and Oliva A. Resolving human object recognition in space and time. Nature neuroscience 2014; 17:455–62

139. Bonner MF and Epstein RA. Object representations in the human brain reflect the co-occurrence statistics of vision and language. Nature communications 2021; 12:4081