# Human scanpaths during active vision

**M.Sc. Thesis**

Buxin Liao

261092132

Supervisor: *Prof. Suresh Krishna*

Jun 8, 2024

# Table of Contents

# Abstract

One of the fundamental questions in visual perception is how and to what extent visual mechanisms integrate information across fixations, and whether visual processing within each fixation is independent. This thesis explores the temporal features of visual search during goal-directed searches of pictures depicting complex everyday scenes. Particular attention is paid to the evidence of information integration and cross-saccade transfer in these more naturalistic search scenarios.

This thesis combines the approaches of previous studies that used either temporal features or scanpath datasets singly, using two publicly available eye-tracking datasets, COCO-Search18 and COCO-FreeView. The aim is to innovate the field of eye-tracking research. It has been discovered that short-latency second saccades, defined as the second saccade occurring before an inter-saccade interval of less than 125 milliseconds, play a crucial role in goal-directed visual search of natural images. They occur frequently, accounting for an average of 45% of all saccades. These short-latency second saccades more often focus on the search target compared to saccades that follow an inter-saccade interval of more than 125 ms (regular-latency saccades). Short-latency saccades show superior target fixation and discrimination capabilities, particularly when they originate further from the target and move closer. Additionally, these saccades are more likely to occur when they start farther from the target.. Short-latency second saccades occur more frequently during goal-directed searches when a search target is present.

The analysis extends to saccades using computer vision models, a method not previously employed. The results indicate that first saccades directed towards the search target typically end at more salient locations compared to those directed towards non-target areas. Short-latency second saccades often follow first saccades that target less salient, non-target areas. Short-latency second saccades tend to end at less salient points than those reached by regular-latency second saccades. Furthermore, short-latency second saccades are not triggered by hypometric first saccades.

The results indicate that human searchers use satisficing strategies when actively searching for well-defined targets in images of complex everyday scenes. The use of short-saccadic and cross-saccadic information transfer minimizes the cost of additional saccades to interfering stimuli; this would not be possible if perception had to be reinitiated with each new fixation. In addition, bottom-up features influence the adoption of different strategies for different saccades during visual search. Short-latency saccades facilitate the rapid determination of the salience of the current fixation point, allowing for rapid responses. The integration and transfer of information across saccades plays a critical role in the effectiveness of visual search.

# Abstract

L'une des questions fondamentales de la perception visuelle est de savoir comment et dans quelle mesure les mécanismes visuels intègrent les informations entre les fixations, et si le traitement visuel se fait indépendamment entre chaque fixation. Ce mémoire explore les caractéristiques temporelles de la recherche visuelle lors de recherches orientées sur des images représentant des scènes complexes de la vie quotidienne. Une attention particulière est accordée à l'intégration de l'information et au transfert inter-saccades dans ces scénarios de recherche plus réalistes.

Ce mémoire combine les approches d'études précédentes qui ont utilisé soit des caractéristiques temporelles, soit des ensembles de données de parcours de balayage, en utilisant deux ensembles de données oculométriques libres d'accès, COCO-Search18 et COCO-FreeView. L'objectif est d'innover dans le domaine de la recherche sur l'oculométrie. Nous avons constaté que les saccades (que nous appellerons secondes saccades) qui se produisent moins de 125 millisecondes après la précédente saccade jouent un rôle crucial dans la recherche visuelle ciblée d'images naturelles. Elles sont fréquentes et représentent en moyenne 45 % de l'ensemble des saccades. Ces secondes saccades se focalisent plus fréquemment sur la cible de recherche que les saccades qui surviennent après un intervalle intersaccade de plus de 125 millisecondes (saccades régulières). Ces saccades montrent une meilleure capacité de fixation et de discrimination de la cible, surtout lorsqu'elles partent de loin et s'approchent de celle-ci. Les saccades rapides sont plus fréquentes lors des recherches orientées vers un but lorsqu'une

cible est présente.

L'analyse s'étend aux saccades utilisant des modèles de vision par ordinateur, une méthode qui n'avait pas été employée auparavant.Les résultats indiquent que les premières saccades orientées vers la cible de recherche tendent à se terminer dans des zones plus saillantes que celles dirigées vers des zones non ciblées. Les secondes saccades suivent souvent les premières saccades qui se dirigent vers des zones moins saillantes et non ciblées. Elles ont en effet tendance à se terminer sur des zones de la scène observée moins saillantes que celles atteintes par les secondes saccades à latence régulière . En outre, les secondes saccades courtes ne sont pas déclenchées par des premières saccades de portée insuffisante ou des premières saccades hypométriques.

Les résultats indiquent que les chercheurs humains utilisent des stratégies de satisfaction lorsqu'ils recherchent activement des cibles bien définies dans des images de scènes quotidiennes complexes. L'utilisation du transfert d'informations en saccades courtes et croisées minimise le coût des saccades supplémentaires vers des stimuli interférents ; cela ne serait pas possible si la perception devait être réinitialisée à chaque nouvelle fixation. En outre, les caractéristiques ascendantes influencent l'adoption de différentes stratégies pour différentes saccades au cours de la recherche visuelle. Les saccades de courte durée facilitent une détermination de l'importance du point de fixation actuel avec un faible délai, ce qui permet des réponses rapides. L'intégration et le transfert d'informations entre les saccades jouent un rôle essentiel dans l'efficacité de la recherche visuelle.

## Acknowledgements

This thesis will be submitted as part of a Dual Master's degree program between McGill University and the University of Electronic Science and Technology of China. Therefore, there may be two matching versions of this thesis, one written in Mandarin and one in English.

I would like to thank Professor Suresh Krishna and the postdoctoral fellow Katarzyna Jurewicz for helping me and guiding me in this process. Without their help, this project could not be going smoothly.

# Chapter 1
# Introduction

Visual search is a common activity in our daily routines. It involves finding objects or information through visual means. This can include finding water when thirsty or locating traffic lights while driving. Visual search can be characterized by four distinct features: (1) Objectivity: the ability to distinguish the target object among many distractors, (2) Invariance: the ability to precisely find the object regardless of changes in appearance or location, (3) Efficiency: the ability to find the target without wasting time searching through the whole scene, and (4) Zero-shot training: the ability to find new targets with zero exposure in advance (Zhang et al., 2018).

An understanding of active vision can inform our understanding of the neural mechanisms underlying visual search, as well as inspire the development of more effective target search algorithms in the field of computer vision. In recent years, numerous papers have focused on tasks such as object classification and object segmentation. The essence of these tasks is to identify the required objects within a search scene, perform label classification or pixel-level segmentation. In these studies, the majority of which focus on the positional characteristics of humans in visual search, there is a paucity of attention paid to the temporal characteristics of the human visual search process. Furthermore, there is a dearth of research utilizing computer vision models on this topic. The spatial and temporal characteristics are the two most fundamental characteristics of human visual search. This thesis primarily explores and analyses

the temporal characteristics of visual search, combining computer vision features to uncover additional hidden features in human visual search. The temporal characteristics of the visual search process can be revealed to provide a more comprehensive understanding of the neural mechanisms underlying human visual search.

To this end, this thesis conducts an in-depth study on the temporal characteristics of humans in target-directed visual search in the dataset based on public eye movement datasets and bottom-up deep learning models. These studies facilitate a deeper comprehension of the human visual system, which can inform the development of more effective target search algorithms in robot vision by more accurately simulating human visual mechanisms.

## 1.1 Visual search

The subject of human attention in visual search has been widely studied. Attention can be divided into two mechanisms: bottom-up attention, which is purely guided by external driving factors that are often very salient, and top-down attention, which is directed by internal attention based on prior knowledge, autonomous planning, and current goals.

Bottom-up attention is triggered during basic visual information processing along the visual cortical pathways. From the primary visual cortex (V1), feedforward signals are transmitted to multiple cortical areas and branch into two main visual pathways: a ventral pathway for processing object- and feature-based visual information and a dorsal pathway for processing information related to space and movement. The visual cortex is organized into two pathways: the ventral pathway, which includes V1, V2, V3, and V4, and the dorsal pathway, which includes V1, V2, and V3. As input rises from early to later stages of the pathway, the

receptive fields become larger, and the functional properties become more complex. Bottom-up information then propagates from the visual cortex to the prefrontal cortex (PFC).

Top-down visual attention is a spontaneous process that involves the internal selection of specific locations, features, or objects relevant to current behavioral goals. This process increases the signal strength of neurons at the cortical level, ultimately regulating vision. Connections between cortical areas are often reciprocal, allowing feedback signals to flow from higher-level visual areas to lower-level visual areas (Katsuki & Constantinidis, 2014).

### 1.1.1 Eye movements

The human eye contains a structure known as the fovea, which is a part of the retina that provides the highest visual acuity. This structure requires humans to focus on an object of interest to obtain necessary information, resulting in eye movements. Visual search studies involve recording eye movements using eye tracking technology. Researchers primarily analyze fixations and saccades in the data. A fixation refers to maintaining one's gaze over a specific area of the visual field, while a saccade is a quick movement between two fixations (Salvucci & Goldberg, 2000). The combination of fixations and saccades creates scanpaths.

After the stimuli appear on the screen, the information takes time to reach the visual cortex and is processed to make final decisions, resulting in saccades. These processes take about 180-250 ms together (Darrien et al., 2001). The time between stimulus onset and saccade onset (or between the end of one saccade and the beginning of the second) is called the saccade latency.

### 1.1.2 Eye movement experiments

A series of related studies provides objective evidence for the integration of visual information by examining the timing between fixations. This property was first discovered in double-step fixation (Becker & Jürgens, 1979). In this trial, stimuli will be presented as either single or double steps in a randomized sequence. The triggering probability of the single-step format is approximately 25%, and the stimulus will appear at an amplitude angle of 15, 30, or 60°. The double-step format involves presenting a stimulus that moves to one location and then quickly appears at another location before the eyes have a chance to respond. There are four possible combinations of these two positions: the stimulus can move twice in the same direction, twice in opposite directions, or the second move can be a shorter distance that returns the stimulus to the first move. The position in front of the stimulus is crucial. The stimulus moves in two opposite directions, with the second movement covering a longer distance. According to the author, goal-guided saccades involve two steps: determining the saccade direction and calculating the movement amplitude. The researchers also discovered that the preparation times for two different saccades may overlap.

Sharika et al. found a similar phenomenon in the modified double-step task. They improved the task by dividing the stimuli into two categories: steps without transfer or steps with target transfer. In the no-transfer step test, a green box appears at one of the four corners of the screen, followed by a red box at one of the remaining triangles. During the target-switch step, two stimuli, one green and one red, are presented, as in the no-shift step. When the first saccade is performed, the red stimulus moves to a new location. The study found that

corrections of the saccade may already be made before the error is detected (Sharika et al., 2008).

Error correction is necessary to compensate for the inevitable errors that occur when not all eye movements go to search targets. Counter-eye movement tasks have also shown that error correction can occur unconsciously (Mokler & Fischer, 1999). In this task, the subject fixated on a red point in the center while two square stimuli were presented on either side. After 100 ms of central fixation, the target to which the subject would move their eyes in the next saccade was briefly illuminated for 40 ms. Then, after another 100 ms, the stimulus appeared in the opposite direction to the previously presented cue. The participants were instructed to press a button to indicate any errors made during the experiment. The researchers analyzed the reaction time of the participants to detect errors. They concluded that erroneous saccades are more likely to be missed by the participants themselves when only non-retinal information of intended fixation is used to update the perceptual space in the lateral parietal cortex.

Experiments have also been conducted by recording eye movements in subjects when stimuli suddenly appear (Theeuwes et al., 1998). In the experiment, six gray circles with identical numbers inside initially appear on the screen. After one second, all but one circle turns red, and the numbers inside them change to letters. The positions of the red circles change, while the positions of the gray circles remain the same. Participants were instructed to quickly identify the letter 'c' inside the circle that remained the same color. The study authors suggest that the brain processes two types of eye movements simultaneously: one is a voluntary movement towards a stationary target, while the other is a reflexive movement triggered by a changing stimulus.

These studies found that the fixation interval between the first and second fixations can be longer than the usual minimum fixation latency of about 125 ms (McPeek & Keller, 2002). When the fixation duration is short, there is insufficient time for the entire sequence of visual processing, fixation target selection, fixation preparation, and execution. Thus, it can be concluded that short-latency second fixations are based on visual information obtained prior to the initial fixation. The transfer of information between fixations is supported by research. A study using categorical images found that in sparse visual search, the second fixation is primarily driven by information obtained before the first fixation (Caspi et al., 2004). However, the latency to the second fixation in this study was much longer than conventional fixation latencies.

## 1.2 Scanpaths datasets

Recently, with the development of deep learning, there has been a lot of new research on visual search. Large datasets of eye movements have been created for convenient research. These datasets can be either free-viewing data, with no specific task given to the subject, or search tasks, which instruct subjects to find the target within an image.

The COCO-Search18 dataset comprises eye movement data collected from 10 subjects searching for an object from one of the 18 target object categories in a set of everyday scene images (Chen et al., 2021). The eye movement data was collected from 10 students at Stony Brook University. The images were selected from Microsoft COCO, a dataset of natural scene photographs that includes labels for 80 object categories (Lin et al., 2014). The COCO-Search18 dataset includes 6202 search-scenes. The same image can be used to search for

different categories. For instance, an image used to search for a microwave can also be used to search for an oven. Search scenes are stored as images in datasets and may be repeated for different search objects. This thesis refers to them as search scenes when the object is a subject or experiment, and as images when the object is a dataset. In this thesis, objects similar to the target object are referred to as distractors. The search scenes are divided into two equally sized groups: 'target-present (TP)' which always contains a single search object in each image, and 'target-absent (TA)' which contains no search target in an image.

The search images and categories were meticulously filtered to ensure that the task was neither too easy nor too difficult. The 18 categories belong to Microsoft COCO and do not include animals or people, as humans tend to fixate on those (Treisman & Gelade, 1980). Additional criteria were applied when selecting TP or TA images. To ensure the quality of the dataset, certain criteria were used to exclude TP and TA images. TP images that did not meet the following criteria were excluded: a certain ratio of the target object, multiple appearances of the objects from the target category, appearance of the target object in the center area of the image, a certain ratio of the image, or less than 100 images per target category. Similarly, TA images that did not meet the following criteria were excluded: images with the target object or with less than 2 instances similar to target categories (based on the hierarchy of the Microsoft COCO labels).

The stimuli were presented to the subjects at a viewing distance of 47 cm from the center point of the 22-inch monitor. The monitor had a pixel resolution of 1680×1050 pixels, spanning a horizontal and vertical visual angle of 54° and 35°, respectively. The paradigm began by displaying example objects from the target category on the screen. Examples of objects that

did not belong to this category were also shown to ensure subjects could better discriminate the target category. The target category may or may not appear in the search scene later shown. Following the display of the target category, a fixation dot appeared at the center of the screen. The subject initiated each trial by pressing a button on a game-pad controller and fixating on the center point. After pressing the button, the search scene appeared and the subject had to indicate the presence or absence of the target category object by pressing the right or left button on the controller, respectively.

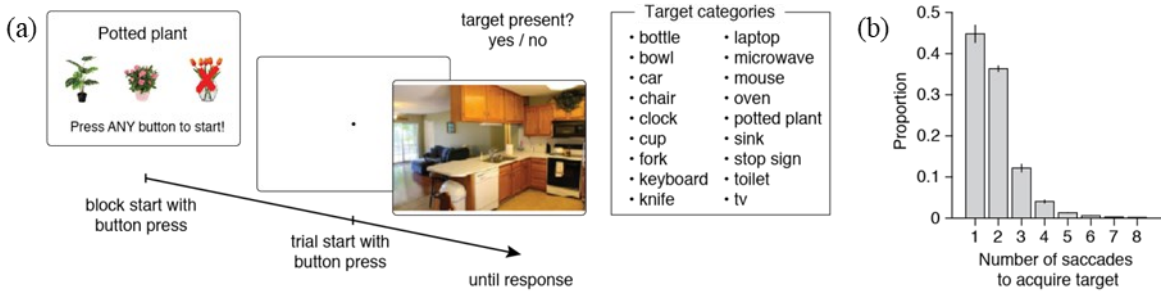


Figure 1-1 COCO-Search18 experiment process. Figure from Jurewicz, Liao and Krishna (Jurewicz et al., 2023). (a) Experimental paradigm of COCO-Search18. The screen displays 2 images related to the search category and one image that is not of that type. The image that appears is different from the object that subsequently appears in the search scene. The 18 object types to be found in COCO-Search18 are marked in the right box. (b) Statistics show the number of saccades used by subjects to find the target object. the target object, with about 81% of the objects being found within two saccades.

Recently, the authors of COCO-Search18 released a complementary dataset called COCO-FreeView (Chen et al., 2022). In this experiment, 10 different participants viewed the same images from COCO-Search18 without a specific task. The experiment included 3 viewing conditions: target-present, target-absent, and free-view, which allowed for the study and comparison of eye movements in different tasks (Chen et al., 2022).

The dataset known as the 'Waldo dataset' features cartoon search scenes, each containing a character named Waldo who is always dressed in red and white clothes and a Santa hat. The scenes are filled with various people and objects, making them confusing, and the task is to locate Waldo who occupies a very small proportion of the scene, as depicted in the picture. During the trial, participants were presented with a search scenario and made eye movements to locate Waldo, whose image was displayed on the screen. The target locations were evenly distributed throughout the image (Zhang et al., 2018).

The MIT1003 dataset is widely used and consists of 1003 images viewed by 15 subjects. The images have a maximum size of 1024 pixels and the task for this dataset was free viewing. At the end of each session, subjects were shown 100 images and asked to identify which ones they had seen before. The subjects' ages ranged from 18 to 35 years old. Each picture is displayed on the screen in its original proportions and remains visible for 3 seconds. A one-second all-gray image will appear between every two pictures. The experimenter will need to compare the subjects each time they view 50 pictures freely. Eye recorders were calibrated to ensure the accuracy of their recordings (Judd et al., 2009). The fixation points of the subjects on each picture were recorded. To obtain a continuous saliency map, the experimenter convolved the fixation points of all subjects on each picture with a Gaussian convolution kernel. A saliency map characterizes the level of saliency of a fixation point on an image (Kümmerer et al., 2018). The MIT saliency benchmark has become the standard for subsequent models to verify their accuracy due to its early release date and large amount of data. This test uses the MIT dataset as input into the model to be tested, and the final performance score is obtained by comparing the saliency map generated by the model with the ground truth saliency map of

the MIT dataset.

## 1.3 Computer-vision models on scanpaths

Since AlexNet's high performance on ImageNet datasets (Krizhevsky et al., 2012), deep learning has become widely used in various fields of vision science. The neurons in deep learning networks are inspired by the neural networks in human brains, and the development of deep learning networks may help us better understand the cognitive processes involved in visual search.

Image segmentation is the process of delineating the region that contains an object. Foveating a region using the fovea can aid in segmentation by providing more detailed information from the given region of an image. To model this process, the scanpaths must be analyzed in depth. For deep learning, predicting the scanpaths can help in finding the object more quickly (Ding et al., 2022). The common method for predicting scanpaths is to generate a fixation density map (FDM). The FDM provides probabilities of foveation for each pixel, which can be interpreted as saliency scores (Samiei & Clark, 2022). The term 'saliency map' is commonly used to refer to these predictions, which are calculated using FDMs (Kümmerer et al., 2018).

Outputs of the network need evaluation to illustrate how well they perform. For classification or pixel segmentation tasks, direct comparison between the ground truth and the output is applied like softmax function. Scanpaths prediction is different from those tasks above. Each fixation of the scanpaths focuses on the subject’s current area of interest. Simply seeing how accurate the position of the prediction is to the ground truth is not feasible. The prediction

and the ground truth for one fixation could refer to the same object, but they may not point to the same coordinates. More metrics are used to evaluate how accurate the scanpath is compared to the ground truth. Since computer-generated scanpaths are most often predicted from the FDM, evaluation on FDM is another method. For other evaluation indicators, the fixation density map needs to be calculated and transformed to obtain a new map for comparison. Kümmerer et al. suggested using saliency map as the terminology here (Kümmerer et al., 2018).

Normalized Scanpath Saliency (NSS) is the metrics based on FDMs. The FDM generated by the model is normalized to have zero mean and unit standard deviation. Salience values of fixations are extracted and averaged as the NSS value (Peters et al., 2005). This metrics allows to valuate performance of models since the saliency score is compared against the distribution of FDM, the higher the NSS value is, the higher the correspondence between the FDM and the scanpaths is.

Pearson's correlation coefficient (CC) is the metrics to evaluate the coefficient between the model-generated FDM and the FDM generated from human behavior. The coefficient is denoted as: $CC(s,h) = \frac{\mathrm{conv}(s,h)}{\sigma_s \sigma_h}$ where $\sigma_s, \sigma_h$ is the deviation of two fixation density maps - from a simulation and a human, respectively. This metrics illustrate the correspondence between two FDMs, and it can be used to compare FDMs generated by different models.

Information Gain (IG) is the increase in information compared to the baseline model. In this case baseline model is the center bias model. Center bias is the known tendency of people to look towards the center of the image (Tatler, 2007). Recent research also points out that people tend to save eye and head motor effort by minimizing the time of viewing eccentrically of the image (Burlingham et al., 2024).

Sequence score (SS) is the method used for evaluating how similar the scanpaths are. The image is evenly divided into multiple areas. Each area of the image is encoded with different characters. The fixation will be marked as the corresponding character according to its location, thus transforming the fixation sequence into a set of characters. Using Needleman-Wunsch string match algorithm (Needleman & Wunsch, 1970), a score of the fixation sequence comparing to the ground truth sequence is calculated (Borji et al., 2013).

SS is more focused on the locations of the scanpaths but cannot capture the semantic meaning of each fixation. Another modification on SS which is the Semantic Sequence Score (SemSS) is introduced to counteract this disadvantage. Yang et al. is doing analysis on target-absent trials from COCO-Search18, and they transform the fixation sequence to a sequence of object categories by using the segmentation annotation from Microsoft COCO. Instead of marking the fixation with one character, things the fixation is on were marked. Then the conventional string comparing algorithm was applied like in the original SS method (Yang et al., 2022).

Multimatch evaluates scanpaths along multiple factors. It compares scanpaths after aligning the scanpaths to have the most similarity. The compared factors include the difference in shape, the difference of angular distance between saccades, the difference in length, Euclidean distance of aligned fixations and the duration of fixations. Multimatch achieves comparing scanpaths in multiple ways (Jarodzka et al., 2010).

In light of the aforementioned evidence, previous researchers have developed a multitude of models for predicting human scanpaths. Kümmerer et al. focused on simulating the human fixation point in free viewing. Based on this, they developed four versions of deep learning

models, named the Deep Gaze series. The original Deep Gaze network only included the processing of search scenes. Subsequent researchers improved the network structure and added a network for analyzing the scanpath. Currently, several versions of the Deep Gaze series of models are ranked at the top of the Tübingen Benchmark and are considered to be excellent models for free viewing (Kümmerer et al., 2022; Kümmerer et al., 2014; Kümmerer et al., 2017; Linardos et al., 2021).

Zelinsky et al. are the main contributors to the COCO-Search18 dataset and the COCO-FreeView dataset. They proposed two versions of the model for predicting scanpaths (Yang et al., 2020; Yang et al., 2023).

IRL is the model first proposed by Zelinsky et, al. This model uses an inverse reinforcement learning network to try to simulate the way the human eye sees things. The model works by processing the image into two images with high-resolution and low-resolution features, similar to the foveal features that the human eye has, and then through reinforcement learning. At the same time, the model uses Generative Adversarial Imitation Learning (GAIL) to use the model through reinforcement learning as the generator (generator) and trains the discriminator (discriminator) to distinguish whether the scan path is the ground truth value or generated by the generator itself. Systems reward models when their behavior resembles human behavior (Yang et al., 2020).

The authors used GAIL's strategy to train a discriminator and a generator simultaneously, where the generator tries to generate a saccade path graph that is as similar as possible to a human saccade path graph, and the discriminator tries to distinguish whether the image is human or computer-generated. In other words, the generator tries to fool the discriminator so

that the discriminator cannot tell which one is real and which one is fake. The discriminator is expected to label 1 as true and 0 as false. This is very similar to the cross-entropy loss (BCE) function, so the discriminator by maximizing the following function: $\mathcal{L}_{\mathrm{D}} = \mathbb{E}_r\left[\ln\left(D(B,a)\right)\right] + \mathbb{E}_f\left[\ln\left(1 - D(B,a)\right)\right] - \lambda\mathbb{E}_r[||\nabla D(B,a)||^2]$, where $\mathbb{E}_r$ and $\mathbb{E}_f$ represent the expectations, or cross-entropy, of real data and false data respectively. The last part is for faster convergence. Because the cross-entropy function is always negative, these three terms are always negative, so $\mathcal{L}_{\mathrm{D}}$ is always negative, so eventually $\mathcal{L}_{\mathrm{D}}$ will converge close to 0 to achieve the target effect. Since the expected value of the real data is $\mathbb{E}_r\left[\ln\left(D(B,a)\right)\right]$, what the generator tries to do is to make its expected value as close as possible to the real expected value to achieve the effect of fooling the discriminator, so $\mathcal{L}_G = \mathbb{E}_f\left[\ln\left(D(S,a)\right)\right]$, based on this discriminator, the corresponding reward function can be deduced as $r(B,a) = \ln\left(D(B,a)\right)$.

The network used sequence scores and Multimatch to test its performance. They selected multiple benchmark models for comparison, including an arbitrary generator that randomly generates scanpaths, a network that determines the location of the target object through a convolutional neural network, and LSTM. wait. The test performance of the IRL network on these standards is higher than the performance level of the baseline network. In addition to being inferior to human observations, the network's average number of fixations required to fixate the target object was higher than the existing models at the time.

The second network version proposed by Zelinsky uses the most advanced transformer network (Dosovitskiy et al., 2020), which they named HAT (Human Attention Transformer) network. The converter network is a classification network that divides the image into multiple small blocks in advance and inputs the multiple small blocks into the converter network

encoder. The encoder generates keys and queries for the image blocks and numerical value (Value), and finally the label category to which the image belongs is calculated under the processing of the multi-layer perceptron. This network is able to integrate the three experimental conditions of COCO-Search18, namely target-present, target-absent, and free-view. The network contains four parts: feature extraction module, fixation module, aggregation module and fixation prediction module. The feature extraction module inputs the search scene into the convolutional neural network to extract information. The extracted features will be input into the fixation module. The most important part of this module is the working memory component. The working memory component will be updated through the converter encoder. Two markers are stored in the working memory. These two markers can remember the location of the feature input to the current module and the sequence number of its fixation. Working memory will be updated every time the network generates a new fixation. In the integration module, this module will select markers from working memory for integration and generate corresponding query Q based on these markers. The finally generated query Q and the features extracted in the feature extraction module are jointly input into the last module, the fixation prediction module. The prediction module will determine the time when the model stops making fixations and determines the location of the next fixation. The model diagram is as follows. This model only needs to modify the number of integrated target objects N to 1 to achieve the free-view function.

The performance of the model was tested under three experimental conditions. Information gain, normalized saccade path salience, and sequence scores were used for testing under these conditions. For target-present and target-absent conditions, the researchers also

used semantic sequence scores for testing. The results showed that in addition to sequence scores, there was a significant difference between target-present and target-absent. Except for being lower under the free-view condition, the rest of the scores are higher than those of other models or the baseline model.

Zelinsky's two versions of the model focus on the 18 target objects marked in COCO-Search18 for modeling analysis, but their model cannot be generalized to all visual searches. What they pursue is the state-of-the-art (SOTA) model. These models require prior knowledge for these target objects outside of model calculations, so the model cannot implement the visual search process for other objects that do not belong to these 18 target objects.

Zhang and colleagues focused on applying visual mechanisms to deep learning models to achieve better performance. They launched two versions of the model. The first model proposed by Zhang et al. is IVSN (Invariant Visual Search Network) (Zhang et al., 2018). The model contains two bottom-up networks for search scenes and search objects, where they used VGG16 as the backbone network (Simonyan & Zisserman, 2014). After removing the fully connected layer of VGG, the search scene and search object are independently input into different VGG networks. After top-down adjustment of the search scene using the extracted features of the search object, the model will finally calculate the attention density map. Zhang calls the density map they obtained an attention density map in the article, but in fact it is also a fixation density map. After obtaining the attention density map, return inhibition is used in the subsequent process of finding the object. The experimenter first designated an area in advance to simulate the visual size of the fovea, called a moving box. Then move the frame to select an area on the attention density map to maximize the sum of the values. If the search

object appears in this area, the search ends. If the search object is not in this area, all values in this area will be immediately set to zero. The moving box will continue to move to find the largest area of the updated attention density map. This action will be repeated until the search object is found. The process of selecting an area by the moving frame is also the process of fixation, and the movement of the moving frame will also form a scanpath. The area with the largest value selected by moving the box here is also called “Winner takes all (Winner take all, WTA)”.

The network was used in three experimental conditions of increasing difficulty. At the same time, the experimenter selected subjects to conduct these three sets of experiments at the same time. The first experiment was to find a target object on a gray background, which the experimenters called the object sequence search task. These objects are from the object categories in the MSCOCO dataset. These objects are converted into grayscale values and appear on the screen. A target object and five other similar objects are evenly distributed in a disc shape on a circle with the center point as the center on the screen. The background of the screen during this test is always gray. After 0.5 seconds of central fixation, the search object will appear on the screen. After 1.5 seconds, the search object will disappear. After a 0.5 second pause, the search scene will appear. The trial ends when the subject selects the location of the search object. The search scene in the first experiment lacked natural scenes, so the second experiment made up for the shortcomings of the first experiment, which the experimenters called the natural scene search task. The experimenter used a natural scene in this experiment. The number of distractors in the natural scene was random instead of the 5 set in the first experiment. The experimental paradigm was largely the same as the first experiment, except

that after central fixation, the name of the search object appeared on the screen instead of the example image. The subject needs to use the mouse to accurately indicate the location of the object in the subsequent search scene. If the subject does not respond after 20 seconds, the trial will also end. In order to test IVSN's ability to find search objects without prior knowledge, and to make the test more difficult, the experimenters introduced a third experiment, the finding Waldo test, which the experimenters called the Waldo search task. The experimental paradigm was almost the same as the second experiment, except that the image of Waldo was only displayed at the beginning of each trial and not in every trial. The search scene will appear after central fixation, and the trial will also end when the subject points out the location of Waldo or does not respond for 20 seconds.

The experiment showed that IVSN can achieve results similar to human behavior. As a comparison, the experimenters did two things: They first set up a null model. The object search algorithm of this model uses a given search box to scan the search scene one by one. For the object sequence search task, the empty model determines the six objects that appear on the screen one by one. Secondly, IVSN was not pre-trained at the beginning of different experiments. This means that the IVSN has no prior knowledge of the search objects like the subjects. The experimenter drew a fixation-accuracy accumulation graph based on the results, that is, the cumulative change in the accuracy of finding the object as the number of fixations increases. If a model requires fewer fixations to achieve a certain accurate accumulation rate, then the model's performance is better. The three experiments consistently showed two results: 1) In the first few fixations, IVSN performs better than humans, but as the number of fixations increases, the human performance level will exceed IVSN, but not until the number of fixations

increases. , the performance level of IVSN will exceed that of humans; the performance of both humans and IVSN is far better than that of the empty model. This result shows that IVSN can achieve similar visual search effects by simulating human vision, but IVSN is still different from humans in four points: 1) IVSN has a global vision that humans do not have. IVSN has global knowledge, while human vision relies on the operation of a mechanism that limits the visual field, such as the fovea; 2) The two mechanisms for judging found objects are different. The subject needs to decide after each scan whether the search object has been found. IVSN relies on an external oracle to tell it whether the object has been found; 3) Humans will often re-focus on the location they have stared at before, IVSN Relying on the return inhibition mechanism limits its ability to revisit previous areas; 4) The current IVSN model uses a pre-trained VGG model, and it does not actually have any learning process during the entire experiment. These phenomena illustrate that IVSN is a preliminary attempt to simulate human behavior, but it is not perfect.

Zhang et al. subsequently proposed a second version of the model, the Target and Context-aware Transformer (TCT) (Ding et al., 2022). The attention map of this model is calculated by a chain connected by multiple target and context-aware attention blocks (TCAB). Each module will propose keys, queries, and values based on the search object, that is, the search scenario. After the feature maps of the search object and the search scene are extracted, they will be input into the first attention module at the same time. The key value of the search object will simultaneously adjust the feature maps of the search scene and the search object. The network that extracts search object and search scene features is CRTNet, which is an object classification converter that integrates object information and contextual reasoning with a

cross-attention mechanism (Bomatter et al., 2021). The calculated search scene and search object feature maps will be input into the second target and scene perception attention module. After a limited number of modules, the final search scene feature map will be extracted as an attention density map. In the same way that the IVSN model selects an object by looking at it, the winner-take-all and return-inhibition mechanisms will select the location of the object. The model is trained on COCO-Search18 and two other datasets, NatClutter and SCEGRAM. NatClutter consists of 240 search scene and search object pairs. These search objects and search scenes are semantically related. For example, the kitchen search scene is semantically related to the microwave oven. The search target was the same object as the one given before the trial but the picture was different. SCEGRAM contains 62 indoor target search objects, each target object appears in 1 semantically related search scene and 5 semantically unrelated search scenes. The target presented in the search scenario and the target image shown in the trial were derived from the same object.

Compared with IVSN, TCT requires fewer fixations than IVSN to achieve the same accurate accumulation rate, which shows that TCT performs better than IVSN. In addition, TCT can also find zero-shot trained search objects in the search scene without retraining or fine-tuning.

In addition to the researchers who have launched multiple models for the above systems, there are also some researchers who have proposed their own models. Samiei et al. modified the MSI-Net network, which is a deep learning network with VGG16 as the backbone network (Samiei & Clark, 2022). Search objects and search scenes are input into different VGG16 networks respectively, and the parameters of the two networks can be accessed from each other.

The two sets of data are subsequently input into the ASPP module, which is a semantic segmentation module with convolutional layers with different expansion rates. Then two sets of feature maps obtained from the search scene and the search object are convolved and then passed through three upsampling modules. This module decodes the feature maps to obtain the fixation density map. The network uses parameters pre-trained on ImageNet and fine-tuned on the COCO-Search18 dataset. Normalized saccade path saliency score, KLD, sequence score, information gain, etc. were used to test the performance level of the network. The experimenters expanded the capacity of the dataset by flipping the dataset images horizontally, and the results showed that the fixation density map generated by the model can indicate the location of the search object. At the same time, the experimenters proved through these evaluation indicators that inputting the search object and search scene into independent VGG networks and applying the ASPP module is better than inputting the data into the same network and not using ASPP.

The ground-truth fixation density map of COCO-Search18 can also show many characteristics of the human visual search process. For example, for search objects such as microwave ovens, the fixation density is concentrated at the bottom; for objects such as bottles and bowls, the fixation density is concentrated on objects such as tables. On the surface. Natural scenes contain many distractors, so Samiei further tried to use deep learning to study distractors and target objects. Samiei adopts the Mask-RCNN network structure. Mask-RCNN is a convolutional neural network trained on MSCOCO that is capable of instance segmentation (He et al., 2017). Mask-RCNN is an improved version of Faster-RCNN (Girshick, 2015), and it can achieve more accurate instance segmentation and can calculate the model's confidence

score that the object belongs to each label. Here, the researchers modified the instance segmentation function of Mask-RCNN. In this task, Mask-RCNN will only determine the segmented objects as target objects or distractors. Because the object labels provided by MSCOCO do not completely include all objects that appear in MSCOCO pictures, for example, the label contains “orange” but does not include “lemon”. The researchers therefore used MSCOCO’s annotator to manually label images in COCO-Search18 in which the target objects were “bottle”, “bowl” and “car”. The network detects up to 20 distractors on an image. The results show that Mask-RCNN has an advantage in segmenting search objects in terms of accuracy, with an accuracy of about 76%. However, the network’s accuracy in segmenting distractors is not high, only 56%. The network tends to identify larger objects as distractors, and objects are only considered search objects or distractors if the fixation falls just within their segmentation contours. Still, the model was able to accurately identify the objects in the image that were most distracting to the search object.

Schwetlick et al. chose a method different from deep learning. They tried to use the mathematical model SceneWalk to model the saccade path. The saccade path model they established has two functions: activation flow function A(t) and fixation-based inhibition flow function F(t). The saliency map and fixation point of the search image are input into the model, in which a Gaussian distribution function with the fixation point as the center simulates the human field of view. A saccade process can be decomposed into three stages: the first stage is the main stage, when human behavior is fixation, and the human attention point and the fixation point coincide with each other. After the second stage, which is the pre-saccade stage, is drawn up. The fixation point is selected. At this time, the fixation point remains at the original position,

and the attention point moves to the position of the proposed fixation point. Finally, in the third stage, the post-saccade stage, the attention point will move along the current fixation point and the proposed fixation point. The straight line where the point is located, that is, the retinal attentional trace (RAT) continues to move a certain distance, and the fixation point will move to the proposed fixation point. At this point, a saccade process ends. At the beginning of the next phase, the attention point and the fixation point will be realigned to start the next saccade process. The activation flow function is a Gaussian distribution function with the attention point as the center of the circle, and the suppression flow function is a Gaussian distribution function with the current fixation point as the center of the circle; the activation map is obtained by combining the saliency map and the activation flow function pixel multiplication, and at the same time, a return suppression mechanism is applied , subtract the suppression flow function to obtain the final combined graph. The next proposed fixation point will be the place with the highest value in the combination diagram. At this time, the attention will move to the position of the proposed fixation point. The center of the circle of the activation function will move with the attention point to regenerate the combination diagram. At this time, the retina will be selected. The point with the highest value on the attention trajectory is used as the proposed movement position of the attention point. At this time, the post-saccadic stage is entered, the saccade action begins to be realized, the attention point moves to the proposed attention point position, the stream function update is suppressed, and the attention point moves to the proposed attention point. At this point, the activation function is updated and the combined graph will be updated again at this time. After entering a new stage, the new combination map will guide the location of the next proposed fixation point, thus starting the next round of

scanning operations. It is worth mentioning that the activation flow function and the inhibition flow function are both Gaussian distribution functions, and their changes are linear. From the main stage to the post-saccade stage, the fixation point changes. For the inhibition flow function, the Gaussian distribution The center of the circle has changed. At this time, the function can be regarded as a Gaussian distribution function with two circle center distributions. Each is multiplied by a coefficient to represent its weight. As time increases, the Gaussian distribution function with the original fixation point as the center of the circle is multiplied by The coefficient will gradually become 0, and the coefficient of the Gaussian distribution function with the new fixation point as the center of the circle will become 1. Activating the stream function is still the same changing steps. The calculation of the combined graph is performed at all times, so the combined graph is also linearly transformed.

## 1.4 Rationale, Aims and Hypotheses

This study primarily examines the temporal characteristics of visual search, utilizing a bottom-up model to investigate the diverse modes of visual search at varying times and their impact on visual search. The specific methodology employed is as follows:

1. Investigation of the temporal characteristics of visual search utilizing the COCO-Search18 dataset reveals that humans employ eye movements to present objects of interest to the fovea. Given the constraints on human brain resources, humans have adopted the foveal strategy for object search. Nevertheless, human visual search is highly efficient and typically requires minimal time to locate the target object. For this reason, there are numerous bottom-up significance models for predicting where humans fixate during free viewing. These models,

such as convolutional neural networks, simulate the process of human visual search by building models that mimic visual transmission pathways and have made considerable progress. However, the bottom-up approach is only one component of the human visual mechanism; the top-down mechanism also plays a crucial role in human visual search. Despite this, the number of goal-directed search models is limited. The majority of current models focus on the location of human attention during visual search, with the capacity to simulate the process of human visual search based on human scanpaths. However, the number of models simulating eye movements is limited, due to the relative novelty of research on the temporal characteristics of human visual search. By studying the temporal characteristics of goal-directed visual search, this thesis can provide theoretical support and new ideas for future model improvement.

2. Continuation of study of the characteristics of visual behavior through bottom-up saliency maps. As previously stated, there are a number of bottom-up models that simulate the bottom-up aspects of human vision. For the human visual mechanism, bottom-up and top-down processes operate simultaneously. Bottom-up and top-down mechanisms collaborate in goal-directed visual search to guide human visual search. These models can also be employed in goal-directed search tasks to further elucidate the impact of bottom-up components. In contrast to top-down models, bottom-up models generate corresponding saliency maps based on actual human performance, which are saliency maps with ground truth values. Consequently, the saliency map generated by the model can be compared with the saliency map generated by the actual human performance to assess the performance of the model. This study also examines the bottom-up aspect of goal-directed visual search, utilizing a deep learning network model that is trained on human ground truth values. The objective is to investigate the role of the

bottom-up mechanism in goal-directed visual search.

# Chapter 2
# Exploration of Temporal Features of the Visual Search Using COCO-Search18

Visual search has long been a subject of research due to its intricate yet efficient capacity to identify targets. There is a wealth of literature that has developed models to simulate visual search based on its characteristics and applied them to fields such as robot vision. For instance, the Convolutional Neural Network (CNN) that simulates the human visual cortex has a global view of the input image, that is, every part of the image is completely visible to it. However, humans rely on the mechanism of placing things of interest in the fovea for visual processing due to limited brain resources. Similarly, temporal features are also overlooked in convolutional neural networks, as people are primarily concerned with identifying the target object's location, rather than the temporal aspects of the network's search.

This chapter primarily examines the temporal aspects of visual search, offering novel insights for subsequent model enhancements leveraging the temporal aspects of visual search.

## 2.1 Introduction

Although extensive studies have been made on modelling *where* the human subjects view the images, the temporal properties of *when* people make fixations are still unexplored (Federmeier & Schotter, 2020). While firm statements have been made that fixation durations play an important role in decision making and can change according to various factors, the

details on the timing on fixations within scanpaths remains largely unknown (Henderson & Pierce, 2008; Kirchner, 2017). For the existing datasets, common features are not fully described and tested in these datasets. The authors often only release the data without further analysis.

The first and the second saccades refer to the first and the second saccades after the onset of stimuli. Studies suggest that although the latency of around 200 ms exist before first saccade (Darrien et al., 2001), the latency before the second saccade is quite different. The preparation for the second saccade might even exist during the first saccade (McPeek et al., 2000). Thus, scanpaths can be influences by the information from the current (second) and the previous (first) fixation.

Although there are a considerable number of eye movement datasets, the majority of them focus on free viewing, such as the SALICON and MIT1003 datasets, or do not provide fixation duration, such as the Waldo dataset. The COCO-Search18 dataset has three advantages over the COCO-FreeView dataset: (1) The COCO-Search18 dataset and the COCO-FreeView dataset encompass both target search and free viewing conditions; (2) The COCO-Search18 dataset and the COCO-FreeView dataset employ the EyeLink standard for Eye movement sampling is employed, and the dataset provides detailed information about the subjects' fixation duration. (3) The COCO-Search18 dataset and the COCO-FreeView dataset utilize the same sequence of search scenes, which are derived from the Microsoft COCO dataset. The images in this dataset are derived from authentic scenes, ensuring that the subjects' fixation in the search scene can be as closely aligned with their visual behavior in their daily lives as possible.

## 2.2 Data Preparation

The data format provided by the COCO-Search18 and COCO-FreeView datasets cannot be directly analyzed because each piece of data is labeled according to the trial. This study focuses on each fixation point and the first two fixations at the beginning of the trial. With regard to the COCO-Search18 dataset, the dataset only marks whether each trial is correct or not, that is, whether the subject correctly answers that the target object exists in the search scene. With regard to the fixation point in each trial, the dataset does not indicate whether the fixation point extends beyond the screen or whether it correctly identifies the target object. Furthermore, the coordinates provided for the fixation point are in pixels, which must be converted to angles for analysis. This information can be calculated, necessitating further processing of the dataset for enhanced analysis.

### 2.2.1 Dataset Introduction

This thesis analyzes the COCO-Search18 dataset, which is divided into two datasets: the target-present and the target-absent dataset. For instance, Table 2-1 displays the attributes and their meanings provided by the target-present dataset.

Table 2-1 Attributes provided by the COCO-Search18 target existence dataset and their meanings.

| Attributes | Meaning |
|---|---|
| name | The image name |
| subject | The subject number |
| task | The task of the trial |
| condition | TA or TP |
| bbox | The bounding box of the search target$(x, y, w, h)$ |

Table 2-1 Attributes provided by the COCO-Search18 target existence dataset and their meanings.

| Attributes | Meaning |
|---|---|
| X | X coordinates from the first to the last fixation |
| Y | Y coordinates from the first to the last fixation |
| T | Fixation durations from the first to the last fixation |
| length | The length of the trial |
| correct | Whether the subject answered correctly or wrongly |
| RT | Subject's reaction time |
| split | "train" or "valid" for deep learning use |

Target-absent is consistent with the attributes provided by the dataset for target-present.

### 2.2.2 Dataset Processing

The search scenes and search objects datasets were derived from MSCOCO and subsequently processed for ease of use in the future research. The MSCOCO dataset provides object parameters for each image, including bounding box parameters, first-level category, and second-level category number. The term 'bounding box' refers to the smallest rectangle that can contain the target object. The MSCOCO dataset has a hierarchical structure with rough categories as the first-level and subdivision categories as the second-level. The second-level categories are contained within the first-level categories.

The bounding box is transformed. COCO-Search18 images have a uniform size of 1680×1050 pixels, while images in the MSCOCO dataset do not. According to the method of generating images from MSCOCO to COCO-Search18, the image will be resized to fill a maximum size of 1680×1050 pixels while maintaining its aspect ratio, and the unfilled parts will be filled with black pixels. Here, this thesis calls the image of MSCOCO as the original

image, the image generated by COCO-Search18 as the target image, and the area that needs to be filled with black pixels as the filling area. Here are the steps to calculate the transformed bounding box:

(1) Calculate the ratio required to resize the original image:

$$r = \min\left(\frac{h}{h_0}, \frac{w}{w_0}\right), h = 1050, w = 1680 \quad (2\text{-}1)$$

$h_0, w_0$ are the height and width of the target image respectively, and r is the scaling factor for resizing. Because the image size in the MSCOCO dataset is smaller than the target image size, $r$ must be greater than 1.

(2) Calculate the size of the filling area. Because the resized image will fill the size of the target image with the largest area, the filling area will be evenly distributed on the upper and lower sides of the image with the width of the target image as the width or evenly distributed on the left and right sides of the image with the height of the target image as the height. side. If the filling area is located on the upper and lower sides, the width of the filling area is $w$, and the height can be calculated according to the equation:

$$rh_0 + 2m_h = h \quad (2\text{-}2)$$

If the filling area is located on the left and right sides, it can be seen that the height of the filling area is h, and the width can be calculated according to equation:

$$rw_0 + 2m_w = w \quad (2\text{-}3)$$

$m_h$ and $m_w$ represent the width and height of the padding area calculated based on height and width respectively. To make the original image fill the size of the target image to the maximum, the filling area must be distributed only on the left and right sides or the top and

bottom sides of the image.

(3) Calculate the transformed parameters of the target bounding box. The bounding box $\boldsymbol{B_0}$ in MSCOCO is recorded as: $\boldsymbol{B_0} = (x_{b_0}, y_{b_0}, w_{b_0}, h_{b_0})$. The point in the upper left corner of the bounding box is the origin, the coordinates are expressed as $(x_{b_0}, y_{b_0})$, and the width and height of the bounding box are expressed as $h_{b_0}$ and $w_{b_0}$. The bounding box $\boldsymbol{B}$ in COCO-Search18 can be obtained by multiplying $\boldsymbol{B_0}$ by the scale factor $r$, and adding the width and height of the filled area:

$$\boldsymbol{B} = r\boldsymbol{B_0} + \boldsymbol{M} \tag{2-4}$$

Expanding formula (2-5) we can get:

$$(x_b, y_b, w_b, h_b) = r \cdot \left(x_{b_0}, y_{b_0}, w_{b_0}, h_{b_0}\right) + (m_w, m_h, 0,0) \tag{2-5}$$

Due to the existence of the filled area, the coordinates of the bounding box need to be added to the offset caused by the existence of the filled area. The width and height of the bounding box only need to be multiplied by the scale factor without caring about the existence of the offset.

The attribute names and their meanings of the processed search scenario dataset are as shown in Table 2-2:

Table 2-2 Attribute names and their meanings in the search scenario attribute dataset.

| Attributes | Meaning |
|---|---|
| area | The object area in MSCOCO (pixels) |
| iscrowd | Whether the object is human |
| image_id | The image id |
| image_name | The image name |

Table 2-2 Attribute names and their meanings in the search scenario attribute dataset.

| Attributes | Meaning |
|---|---|
| width | The image width in MSCOCO |
| height | The image height in MSCOCO |
| bbox | The bounding box for the object$(x, y, w, h)$ |
| category_id | The category id the object belongs to |
| supercategory | Object's supercategory |
| category_name | Object's subcategory |
| id | Object's id |
| is_target | Whether the object was used as the search target |
| rel_bbox_area | The relative bounding box area |
| abs_bbox_area | The absolute bounding box area |

In this dataset, the relative area of the bounding box is the area of the bounding box relative to the image size in the MSCOCO dataset. The number of the object is unique, that is, the number of the same type of object is also different.

Here we take the processing target existence dataset as an example: for each trial, X, Y, and T are $1 \times N$ vectors, and they correspond one to one. For the convenience of the subsequent research, the data were processed as follows:

(1) Split X, Y, T data. In the original dataset, each set of data represents the data collected in each trial. Here, each trial is split into multiple sets of data. X, Y, and T in each set of data only contains a numerical value and adds a parameter value "indx_sac" to these data, indicating the serial number of this set of data in the trial, starting from 0. For each trial, if the length of the vector in the original data is N, it means that the subject in this trial made N fixations, and the sequence number will be marked from 0 to N-1. For the convenience of the subsequent

research, X, Y, and T have been rearranged. The COCO-Search18 experiment is recorded via EyeLink at a frequency of 1000 Hz, and the standard EyeLink protocol is used to automatically detect saccades. The X and Y positions of the eye are recorded. If the velocity of the eye movements is above the threshold, the protocol considers this moment as the beginning of a saccade, and if it falls below the threshold, the protocol considers this moment as the end of a saccade. The definition of the first saccade in this thesis is the first saccade after the central fixation, the second fixation is the fixation after the end of the first saccade, the first saccade latency is the time to execute the first saccade. The time is the duration that the subject maintains the central fixation after the stimulus appears, which is the first data of T in each trial. We rename parameter "T" to "latency" to indicate that the time here is the meaning of saccade latency. At this time, the meaning of each set of data changes from "data collected in each trial" to "related attributes of each fixation". Since the coordinates of the first saccade are the second in the X and Y of the original data, the latency of the first saccade is the first value of T in the original data, so the value of T is aligned with the X and Y values in latency. The first latency value in X and Y that is the central fixation is set to the NaN value in MATLAB, and the X and Y coordinate values corresponding to the last latency value are set to NaN values, because by definition, these two values do not have the meaning in these fixations. At this time, there are $N + 1$ groups of data split from each trial, and the serial number of each trial is marked from 0 to N.

(2) Calculation or conversion of amplitude. In visual search, the distance is generally expressed in terms of amplitude. This attribute represents the angle required for the eyes to fixate this point in degrees. Therefore, it is necessary to calculate and convert the Euclidean

distance on the screen into degrees. The resolution of the screen is 1680×1050 pixels, and the aspect ratio of the screen is $1680/1050 = 1.6$. The screen size is 22 inches. The screen size refers to the length of the diagonal. Therefore, the width and height of the screen can be calculated based on trigonometric functions, which are approximately 47.4 cm and 29.6 cm. The subjects will be calibrated during the test so that their eyes are perpendicular to the center point of the screen and the distance to the screen center is 47 cm. At this time, the subject's visual angle can be calculated according to the inverse trigonometric function, which is 54°×35°. Thus amplitude value of any point (x, y) on the screen can be calculated by the equation:

$$\theta_{(x,y)} = \arctan\frac{\left|\left|(x,y),\left(\frac{w}{2},\frac{h}{2}\right)\right|\right|}{\mathrm{c}} \tag{2-6}$$

Where $\left|\left|x,y\right|\right|$ represents the Euclidean distance between x and y, expressed in centimeters, and c is the distance from the eye to the center of the screen, which is 47 cm.

In the same way, the distance between two points can also be converted into an angle. The angle here means the angle required for the eyes to move from one point to another. Then for point $(x_1, y_1)$ and point $(x_2, y_2)$, the distance is:

$$d = \arccos\frac{c^2\cos^2\alpha + c^2\cos^2\beta - l^2\cos\alpha\cos\beta}{2c^2\cos\alpha\cos\beta}, \alpha = \theta_{(x_1,y_1)}, \beta = \theta_{(x_2,y_2)} \tag{1-7}$$

where $l$ is the Euclidean distance between the point $(x_1, y_1)$ and the point $(x_2, y_2)$.

(3) Judgment of fixated objects. If the fixation point falls within the bounding box of the target object, then we say that this fixation endpoint fixates the target object. An attribute "indx_on_tar" has been added, which represents the serial number of the object that was fixated at. Select the first point that fixates on the target object in the trial, and its "indx_sac" value is

$n$. Then this thesis will label all the data in this trial as $-n, -n+1, \ldots N-n$. If no fixation points fixate the target object in this trial, its "indx_on_tar" sequence value is consistent with the "indx_sac" sequence value. The attribute "target_found" is added to indicate whether a fixation point fixated on the target object in this trial. In addition, based on the object label and object bounding box contained in the image provided by MSCOCO, the type of object fixated by each fixation endpoint is determined in turn, and the fixation endpoint fixated within the object bounding box is still defined as the object being fixated. At this time, this fixation endpoint may not fixate any object, or it may fixate multiple objects.

(4) Add other attributes. Added some additional properties for easy access. For example, the angle of the fixation point on the screen is added. This attribute represents the angle between the line drawn by the screen center point and the fixation point and the x-axis. Because for the image, the direction of the x-axis is from left to right, and the direction of the y-axis is from top to bottom, then the four quadrants represented by this coordinate axis are the lower right, lower left, upper left, and upper right areas of the screen. The "times_used" attribute was also added to represent the number of times the search scenario used in this trial appeared in different target object search tasks. The attribute names and meanings corresponding to the processed fixation attribute dataset are as shown in Table 2-3:

Table 2-3 Attribute names and meanings of the saccade attribute dataset under target-present conditions.

| Attributes | Meaning |
|---|---|
| name | The image name |
| subject | The subject number |
| task | The task of the trial |

Table 2-3 Attribute names and meanings of the saccade attribute dataset under target-present conditions.

| Attributes | Meaning |
|---|---|
| condition | TA or TP |
| bbox | The bounding box of the search target $(x, y, w, h)$ |
| area_degrees | The bounding box area of the search target |
| X | The X coordinate for the fixation before the saccade |
| Y | The Y coordinate for the fixation before the saccade |
| length | The length of the trial |
| correct | Whether the subject answered correctly or wrongly |
| RT | Subject's reaction time |
| indx_sac | The index for the fixation |
| distance_from_center | The distance of the fixation from the screen center (degree) |
| latency | The latency for each saccade |
| indx_on_tar | The index for fixations on the search target |
| target_found | Whether the target was found in this trial |
| distance_target_before_sac | The distance to the target before the saccade (cm) |
| distance_target | The distance to the target after the saccade (cm) |
| distance_target_border | The distance to the target border after the saccade (cm) |
| distance_target_before_sac_deg | The distance to the target before the saccade (degree) |
| distance_target_deg | The distance to the target after the saccade (degree) |
| distance_target_border_deg | The distance to the target border after the saccade (degree) |
| angle_target | The angle of the target |
| amplitude | The amplitude of the fixation |
| angle_fixation | The angle of the fixation |
| times_used | The time used as the search scene |
| on_cat | The category being fixated |

Table 2-3 Attribute names and meanings of the saccade attribute dataset under target-present conditions.

| Attributes | Meaning |
|---|---|
| on_cat_id | The category id being fixated |
| numbers_distract | The numbers of the distractor |
| in_information_border | Whether the fixation was in the filling area |

The target-absent dataset was processed in the same way as well as the COCO-FreeView dataset, which complements the COCO-Search18 dataset. The data structure of COCO-FreeView is similar to that of COCO-Search18, so the process will not be described here. As the test condition of COCO-FreeView is free-view, the saccade attribute dataset generated based on this dataset does not have attributes related to the target object.

The fixation points' coordinates can be categorized into three groups: fixation points outside the screen, fixation points in the black filled area, and fixation points in the normal area. Fixation points outside the screen are incorrect, while fixation points in the black filled or normal image area are considered normal points on the screen. Therefore, we selected all fixation points for the subsequent research, except for the fixation point outside the screen, unless otherwise specified. Furthermore, we exclude trials that were marked incorrect where subjects answered wrongly in these trials.

## 2.3 Results

10 subjects participated in each of the experiments, and their performance vary. In later analysis, standard error of mean was used across these subjects when analyzing their performance:

$$\mathrm{SEM} = \frac{s}{\sqrt{n-1}} \tag{2-8}$$

where $s$ is the standard deviation, $n$ is the number of the subjects.

For the purpose of statistical analysis, hierarchical generalized linear models (GLM) were employed, which accounted for between-participant variability in accordance with the specifications of the Matlab 'fitglme' function. In instances where the dependent variable was a probability, the binomial distribution and logit link function were utilized. Conversely, when the dependent variable was latency, the normal distribution and identity link function were applied.

For visualization, the latencies are grouped into 10 bins with equal sizes for each subject. The variable that relates to latencies will be calculated for latencies in each of these bins and averaged across subjects. For distance or amplitude, all of these are converted to degrees for convenience, and they are grouped in 5-degree bins.

In these results, target-present dataset is analyzed and unless specified, target-absent and free-view datasets are not comprised by the analysis.

Figure 2-1 shows the distribution of the first and second saccades. The first saccades show a bimodal distribution separated by the time of 125 ms. The first group within the 125 ms has an average latency of 55 ms, the second group has the average latency of 235 ms, starting from 125 ms to 400 ms. The first group is the pre-emptive saccades. Check on these saccades show they have a very low probability of foveating the target (Figure 2-1B, the single blue dot before 100 ms). These saccades are excluded from further analysis. The second group of the first saccades is the stimulus-driven first saccades, as they are the regular-latency saccades triggered

by the stimuli onset. The figure clearly showed that second saccades have substantially lower average of latencies (131±5 ms) than first saccades (234±5 ms) and this is the case for all subjects. The shaded areas in the figure represent the SEM across subjects. Upon examination of the individual performance data, it was found that all values exhibited a p-value of less than 0.0001, as determined by two-sample t-tests.

The stimulus-driven first saccades start from 125 ms, which is used as the boundary of short or regular latencies for second saccades. For the second saccades, those with less than 125 ms latency are defined as short-latency second saccades, and those more than 125 ms latency are defined as regular-latency second saccades. This is consistent with the finding that second saccade may be prepared during the first saccades whereas the first saccades cannot be prepared in advance due to no prior knowledge about the stimuli (McPeek et al., 2000). The figure also shows a high probability for the short-latency second saccades to foveate the target (Figure 2-1B, 2-1C, short-latency second saccade: 73 ± 2% versus regular-latency second saccade: 62 ± 1%, mean difference 11 ± 1%, $p < 0.0001$).

The first saccades also showed a decreased probability of foveating the target as the latency becomes longer (GLM, effect of saccade latency, beta = -0.003, $p < 0.0001$, Figure 2-1(b), blue points), and the probability of triggering short-latency second saccades is also decreasing (GLM, effect of first saccade latency, beta = -0.002, $p < 0.0001$, first latency bin: 48 ± 1%, last latency bin: 40 ± 1%, Figure 2-2(d)). These phenomena happening together point out the early stage of the first saccades have richer information.

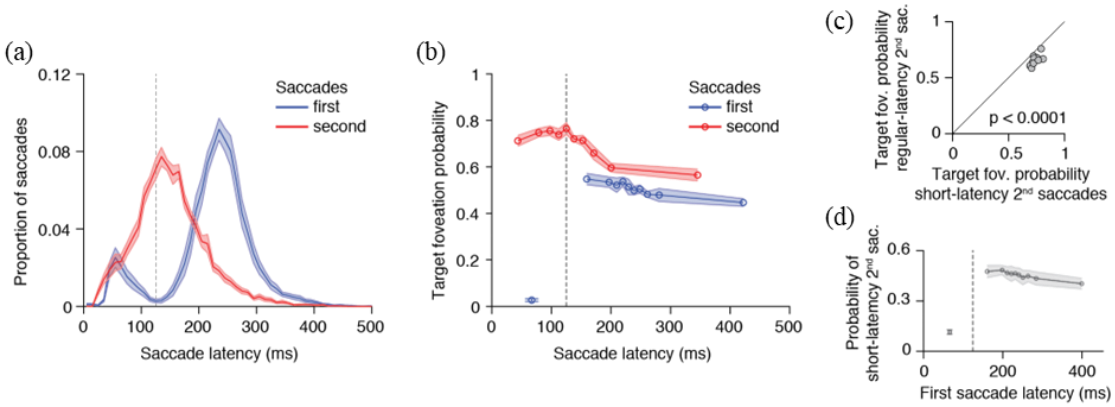


Figure 2-1. Second saccade latencies are short in average and are more target-directed in target-present trials. 45% of second saccades are short-latency saccades that have less than 125 ms latencies. A) Latency distribution of first and second saccades. B) The target foveation probability for first and second saccades. Second saccades are more likely to fixate the target, and those with short-latency have substantially higher probability of target foveation. C) short-latency second saccades have higher probabilities of foveating the target than regular latency second saccades. D) The probability of making short-latency second saccade is higher when first saccades have short latencies.

For the probability of saccades foveating the target, the phenomena of short-latency second saccades maintaining a high target foveation probability exist across the different starting point distances from target (Figure 2-2). As starting point distance from target increases, second saccades drop more slowly in target foveation probability than first saccades. In the second saccades, short-latency second saccades drop more slowly than regular-latency second saccades. This advantage of foveating the target grows as the starting point distance from target increases. The proportion of short-latency second saccades is higher when the distance reaches around 10 degrees. That is, relatively more short-latency second saccades than regular-latency second saccades are triggered at a large distance from target. This is not because the starting points of first saccades that will be followed by either short-latency second saccades and regular-latency second saccades is differently distributed, the proportion of the starting distance

for the first saccades is the same (Figure 2-2).

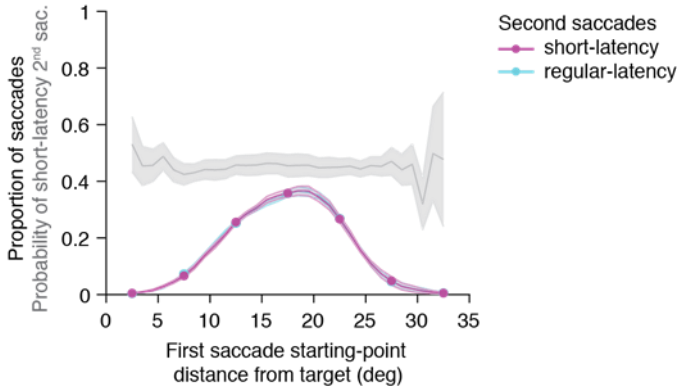


Figure 2-2 The probability of a short-latency second saccade does not depend on the distance between the starting point of the first saccade and the target. The magenta solid line and the cyan solid line show the proportion of short-latency and regular-latency second saccades at different distances between the starting point of the first saccade and the target. The gray line represents the proportion of short-latency second saccades (magenta solid line) that are triggered. It can be seen from the figure that the proportion of short-latency second saccades does not change with the distance from the starting point of the first saccade to the target.

As the pre-saccadic distance from the target increased, the probability of foveating the target decreased in a linear fashion (GLM, main effect of the distance from target, beta = -0.08, $p < 0.0001$). However, short-latency second saccades demonstrated a distinct advantage in foveating the target compared to regular-latency second saccades. This advantage grew more pronounced with increasing distance from the target, reaching a substantial (~25%) level even at distances of approximately 30 degrees away from the target (GLM, interaction between distance and second saccade type, beta = 0.03, $p < 0.0001$, Figure 2-3(a), dashed lines). In other words, despite the brief intersaccadic interval between the first and second saccade, short-latency second saccades demonstrated a markedly reduced decline in target foveation probability as the target became increasingly eccentric on the retina during the second fixation.

Additionally, short-latency saccades exhibited a heightened propensity to be initiated when the first saccade terminated at a greater distance from the target (Figure 2-3(a), dots and solid lines, GLM, probability of short-latency second saccade, main effect of distance from target, beta = 0.03, $p < 0.0001$). In other words, the number of short-latency second saccades initiated at a large distance from the target was greater in proportion to the number of regular-latency saccades. This was not due to a systematic difference in target distance from the initial saccade starting point for short-latency and regular-latency saccades. The control analysis revealed that the proportion of initial distances from the target (prior to the first saccade) was identical for the initial saccades followed by short- and regular-latency second saccades (GLM, effect of first saccade starting-point distance from target, beta = 0.0009, $p = 0.842$). Moreover, our findings indicate that the target-foveation advantage for short-latency saccades remains consistent even after accounting for potential confounding factors, including target eccentricity, target size, number of distractors, and category difficulty.

The probability of foveating the target was found to be greater for short-latency second saccades than for regular-latency second saccades, even when the direction of the second saccade deviated more from the direction of the first saccade (GLM, interaction between change in saccade direction and second saccade type, beta = -0.003, $p = 0.002$). However, the probability of foveating the target generally decreased with the magnitude of direction change (GLM, main effect of the change in direction, beta = -0.008, $p < 0.0001$, Figure 2-, dashed lines). This advantage is significant for cases where the saccade is opposite to the first saccade (direction change of 150-180°). Regarding the frequency of saccade direction changes, it is already known from free-view studies that the most common are saccade sequences in the same

direction or in the opposite direction (Schwetlick et al., 2020).

Interestingly, in the COCO-Search18 data, short-latency saccades were most likely to change the saccade direction, while regular-latency saccades were most likely to continue in the current direction. As the direction change increases, the proportion of short-latency second saccades increases compared to the proportion of regular-latency saccades (Figure 2-3(b), solid line). Therefore, there is no evidence that a short-latency second saccade can complement a first saccade's miss-to-target action on a single trajectory.

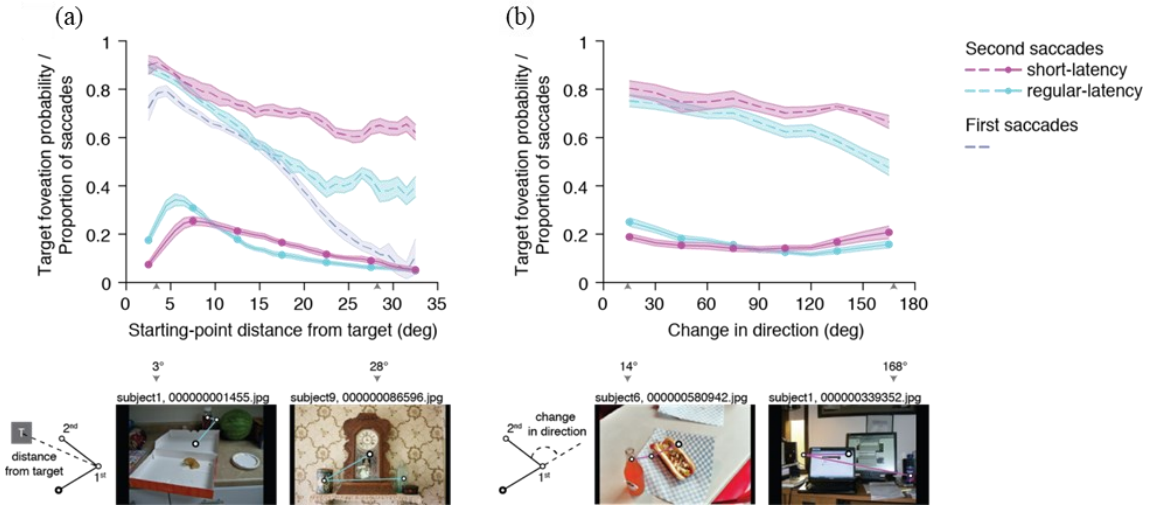


Figure 2-3 A short-latency saccade usually starts farther away from the target and proceeds in the opposite direction from the previous saccade. (a) When the distance between the starting position and the target is large, the short-latency second saccade has a greater target fixation advantage (magenta dotted line) compared with the regular-latency second saccade (cyan dotted line). The target fixation probability for the first saccade used for comparison (light blue dashed line) also decreases with increasing distance from the object. Likewise, the proportion of short-latency second saccades with onsets farther from the target (magenta solid line) is greater than the proportion of regular delayed saccades (cyan solid line). Bottom row: Schematic showing the distance between the start point of the second saccade and the target (dashed line), and examples of saccade paths with the start of the second saccade closer (left) or further away from the target (right). Both examples were correct trials of the "bottle" search target category, in which participants accurately fixated the target on their regular-latency second saccades. (b) When the direction change from the

previous saccade is large, the short-latency second saccade has a greater target fixation advantage (magenta dashed line) compared with the regular-latency second saccade (cyan dashed line). The proportion of general direction change at the onset of the short-latency second saccade (solid magenta line) is greater than that of the regular-latency saccade (solid cyan line). Bottom row: Schematic illustrating the change in direction between the first and second saccades (dashed line), and examples of saccade paths in which the second saccade proceeds in the same direction (left) or in the opposite direction (right) as the first saccade . Both examples were correct trials of the "bottle" search target category, in which participants accurately fixate the target on the short-latency second saccade.

Consistent with the well-known phenomenon of the extracentral preview effect, after participants fixated on a target, they took significantly less time to provide a manual response if the second saccade started closer to the target (94% after the second saccade fixate the target of time fixation remained on the target until the manual response was completed; results did not change qualitatively when remaining trials were excluded). Importantly, when the target was further away from the center of the screen during the second fixation, the increase in verification time was significantly less for short-latency second saccades than for regular-latency second saccades (Figure 2-4). This suggests that, in addition to information about target location, trials with short-latency saccades also benefit from information about target identity, which is conveyed by the first saccade or, in trials containing short-latency saccades, compared with those containing regular-latency saccades. Compared with the experiment, the processing speed of target information was accelerated. Comparison of verification times showed that verification time depended on target eccentricity distance to a greater extent during the second fixation and that verification time depended on target eccentricity distance to a greater extent for short-latency saccades than for regular-latency saccades (GLM, interaction between saccade starting-point distance from the target and verification time, beta = -9.882, $p < 0.0001$).

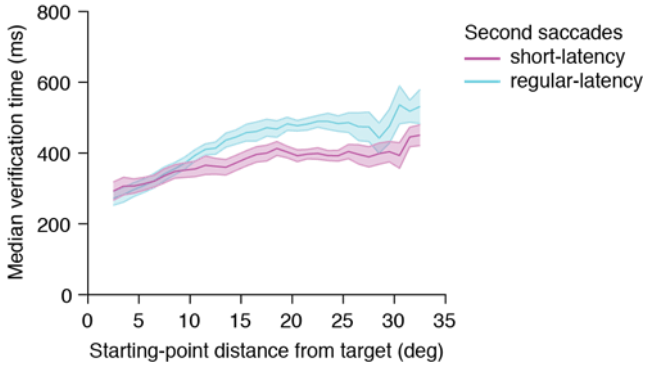


Figure 2-4 Short-latency saccades show target recognition advantages. When subjects fixate the target with a second saccade and the starting point was further away from the target, they took less time to provide the correct response after fixating on the target than with a regular-latency second saccade.

The proportion of short-latency second saccades also varies with the task. Figure 2-5 shows the different proportion of short-latency second saccades as the viewing conditions change from target-present to target-absent and to free-view, where the goal-directed search plays a diminishing role in these 3 tasks, along with the diminishing top-down activity. It is the search task that is mainly engaging the top-down activity, as it requires an active maintenance of and comparisons to the target category. The target-present condition and the target-absent condition are parts of the same task: the subjects need to answer whether the target object is in the image, whereas in the free-view, subjects need only to view the image without looking for a target. The figure also shows that when the object is present in the task, the latencies for first and second saccades are significantly shorter, that is to say, there are more short-latency second saccades in target-present conditions (target-present: 45 ± 3%, target-absent: 22 ± 2%, free view: 18 ± 3%). This very likely points out that short-latency second saccades exist when there is more top-down activity is in the visual search.

As the target was absent from both the free view scenes and the target-free scenes, a direct

comparison of the target-foveation probability was not possible. Instead, we examined the proportion of short-latency second saccades as a function of saccade amplitude. Saccade amplitudes refer to the distance of the saccades. Large saccade amplitudes may refer to the searching process, and when the saccades amplitude is large, the proportion of short-latency second saccades is relatively higher in target-present and target-absent search (Figure 2-5(c), top row). The percentage of second saccades larger than 15 degrees was found to be substantially higher in target-present (29 ± 2%) and target-absent (27 ± 2%) conditions than in free-view trials (GLM, effect of saccade amplitude, target-present: beta = 0.05, $p < 0.0001$, target-absent beta = 0.03, $p < 0.0001$, Figure 2-5(c), gray lines). Moreover, the probability of short-latency second saccades was higher for larger-amplitude saccades in target-present and target-absent conditions, but not in the free-view condition, where the probability of short-latency saccades was actually decreasing with amplitude (beta = -0.02, $p < 0.0001$). To quantify and compare the effect of amplitude on the relative number of short-latency second saccades across the three viewing conditions, we also fit the model separately to data from each individual participant. The average slope of short-latency saccade probability was found to be significantly higher when the target was present (average beta = 0.054 ± 0.006) than when the target was absent from the scene (average beta = 0.032 ± 0.005, $p = 0.004$, paired t-test). Furthermore, both conditions were found to be significantly higher than when participants were free-viewing the scene(average beta = -0.018 ± 0.015, both $p \leq 0.004$, two-sample t-tests).

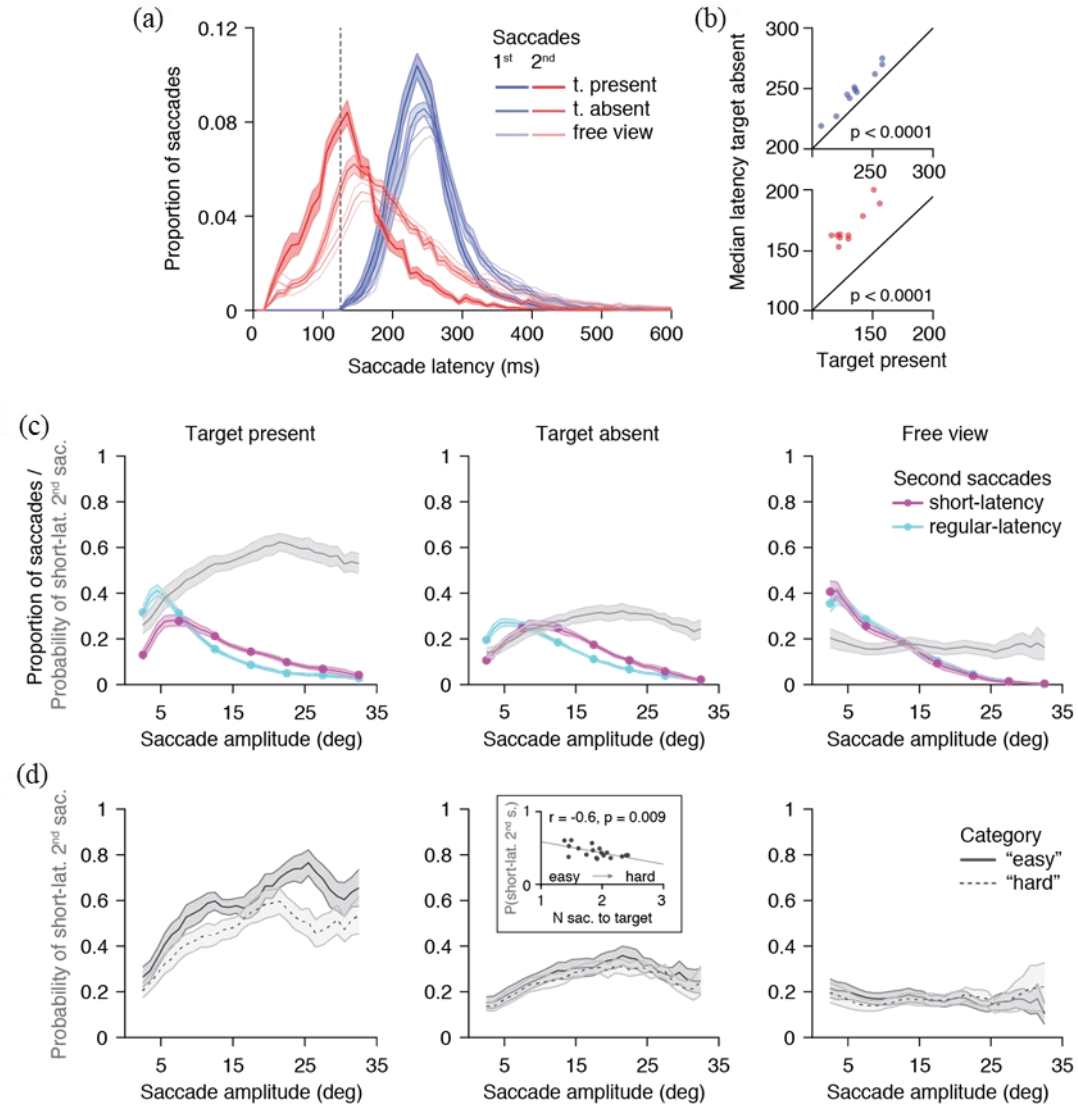


Figure 2-5. Comparison of behavior of first and second saccades in different tasks. A) The proportion of short-latency second saccades decreased as the viewing condition changes from target-present to target-absent and then to free-view while the first saccades decrease much slower. B) The comparison of first and second saccades in target-present and target-absent shows a clear trend of second saccades slowing down. C) Distributions of short-latency and regular-latency saccades. Top row reveals that short-latency saccades are more likely to be triggered in the target-present tasks and with larger amplitudes. Bottom row shows that the onset of short-latency second saccades varies with the difficulty of the task in target-present viewing condition.

The difference of proportion of short-latency second saccades also exists between target-present and target-absent search where the only difference between them is whether the target

object appears in the image. Similarly, if the target category is different, the proportion for the short-latency second saccades may also be different. As target category varies, probabilities of short-latency second saccades may also be different in the trials that contain the target (but not when it is absent in the scene), so we did the further analysis by separating the search target categories into 3 groups by checking the average lengths of fixations to find the target. We define the fixation that is on the bounding box of the target as the successful finding out the target. If multiple fixations were made on the target area, we only take the earliest fixation. The number of fixations made to this particular fixation is the number of the fixations to find the target. The shorter the number, the easier the object is to find in the image. We do not plot the medium difficulty group. The easier group has larger proportion of short-latency second saccades, and the proportion is higher when saccades amplitude is larger in target-present trials (GLM, interaction of saccade amplitude and category difficulty, beta = 0.03, $p <= 0.0001$, Figure 2-5(d)). But such difference is not seen in target-absent trials. These results illustrate that when the target is absent, the short-latency second saccades are not more likely to be triggered in spite of the different difficulty between search categories.

## 2.4 Discussion

This chapter analyzes two large open public datasets, COCO-Search18 and COCO-FreeView, which record the fixation patterns of human subjects while free-viewing or using active visual search for everyday scenes from the MSCOCO dataset. The study refers to these second saccades triggered by short eye-movement intervals (fixation durations) as saccades. The research finds that short-latency second saccades are very frequent during goal-directed

visual search with natural images.

The duration of short-latency saccades does not allow sufficient time for the complete process of visual processing, saccade target selection, saccade preparation, and execution to occur. Therefore, this thesis concludes that short-latency second saccades are based on the first saccades to acquire visual information.

The study revealed that short-latency second saccades, which are defined as saccades with an interval of less than 125 ms between them, occur frequently (accounting for 45% of total saccades) during goal-directed visual search in natural scenes. Previous studies have reported short-latency saccades in tasks such as simple and modified two-step saccade tasks, antisaccade tasks, and tasks with sudden triggering. These tasks were designed to maximize the proportion of error-corrected responses following an initial error (Becker & Jürgens, 1979; Mokler & Fischer, 1999; Sharika et al., 2008; Theeuwes et al., 1998). Most of these studies have involved simple laboratory tasks where the saccade target is highly salient. However, in the sandwich-making task, Hayhoe et al. found that the proportion of short-latency second saccades remained high even under natural fixation conditions (Hayhoe et al., 2003). Short-latency second saccades are more frequent in target-guided visual search in target-present conditions, and both active search and the top-down salience of the search target contribute to this frequency. This study also highlights the importance of this saccade in an active search task in natural scenes. The study suggests that the top-down salience of the search target facilitates its detection during the first saccade, leading to a short-latency second saccade. Additionally, short-latency saccades were more frequent when the target was easier to locate. When the search target is easily found on the first fixation, short-latency saccades occur infrequently because the first

saccade is goal-directed. However, when they do occur, they are highly goal-directed. When the search target is very difficult to find during the first fixation, short-latency saccades may be less frequent. At moderate difficulty, short-latency saccades may be both frequent and goal-directed, which is the finding of this thesis.

It may be tempting to overlook the benefits of short-latency saccades in this thesis, assuming that such high accuracy at such short fixation durations must be due to some form of confusion. However, this thesis demonstrates that the advantage of short-latency saccades is not due to their initial distance from the target or to the decomposition of a saccade into primary and corrective saccades. The second saccade with a short latency is not utilized to correct localization inaccuracies that may occur when fixating on a target. Instead, it compensates for an incorrect initial target selection. The observed effects may be influenced by unobserved factors. This thesis also examined other related variables, including target size, location, and category, which confirmed this thesis's findings.

This thesis states that the boundary between short-latency and regular-latency saccades is set at 125 ms based on the first saccade latency distribution. However, this limit is not considered to be fixed. It is observed that there is a sharp decrease in target fixation probability during the second saccade latency around this limit. The data suggests that the distribution is a weighted mixture of two underlying distributions, one for trials based on information before the first saccade and the other for 'regular-latency' trials collected using the second fixation information. Alternatively, the information from the first and second fixations may be associated with increasing importance of the second fixation on each trial due to the lack of useful cross-saccade transfer in the first fixation, the lack of cross-saccade information

dissipation, or weighting of information from the second fixation, among other possibilities. This thesis notes that short-latency saccades have advantages in target fixation and identification, indicating that information about target location and identity is conveyed between the first two saccades and used on the third fixation target. This finding is consistent with previous research on the benefits of additional visual field preview across multiple saccades (Krishna et al., 2014).

The advantage of short-latency saccades in target fixation may be overlooked when starting farther away from the target. Short-latency saccades are more likely to be elicited when they begin further away from the target (i.e., the target distance before the second saccade is greater). This thesis hypothesizes that these effects result from an interaction between the top-down salience of the search target stimulus and lateral inhibition. Lateral inhibition plays a crucial role in the superior colliculus, frontal eye fields, and lateral ventricular zones (Chakraborty et al., 2022; Huber-Huber et al., 2021). Thus, the selection of the first saccade target may inhibit surrounding stimuli, preventing them from becoming salient during the blank period and triggering a short-latency second saccade. However, stimuli that are farther away from the target of the first saccade are less inhibited. Therefore, they can be selected during the blank period and trigger a short-latency second saccade. It is important to note that when the first saccade ends, the search target stimulus is less inhibited. Recent modeling results suggest that surround suppression, also known as lateral suppression, is crucial in explaining the properties of error-correcting double responses in the evidence accumulation model (Lin et al., 2014). Without surround suppression, short-latency saccades would be more frequent than observed.

Notably, short-latency second saccades are goal-directed. The saccade target is already selected before the execution of the first saccade, and it is too late to cancel them. During this process, competition occurs. If the endpoint of the first saccade is more salient, then the saliency would suppress the trigger for short-latency second saccades. Second saccade latency is the time between the end of the first saccade and the onset of the second saccade. If the important areas require more time to process, the second saccade will have a longer latency, reducing the likelihood of triggering short-latency second saccades.

# Chapter 3

# Extension of Characterization of Visual Behavior Using Saliency Maps

In image segmentation, bottom-up methods divide the image into multiple regions, extract features from each region, and then recognize these regions as objects to segment them based on the features of the objects (Borenstein et al., 2004). Bottom-up is also considered one of the two components of visual attention. For example, a red object on a green background attracts more attention due to the high contrast under such conditions (Gegenfurtner & Kiper, 2003). This thesis explores hidden features from bottom-up saliency maps provided by deep learning models, such as the Deep Gaze networks.

## 3.1 Introduction

In the bottom-up deep learning network, the data utilized for model training originates from human experiments. These experiments entail free viewing without tasks. Each search scene in these data sets contains human fixation on it, and the saliency map for this search scene can be generated based on the distribution of fixation points.

Kümmerer and colleagues have been working on developing Deep Gaze network for over 8 years. The network has now 3 generations and 4 versions (Kümmerer et al., 2022; Kümmerer et al., 2014; Kümmerer et al., 2017; Linardos et al., 2021). Its purposes are saliency and scanpaths prediction, and it is always used based on the free-viewing behavior.

Deep Gaze I is based on AlexNet (Krizhevsky et al., 2012), but all full connected layers are removed. The network is trained on MIT1003 dataset (Judd et al., 2009), with images of the size of $1024 \times 768$ pixels. The FDM is generated using the equation $o(x,y) = \alpha c(x,y) + \sum_k w_k r_k(x,y) * G_\sigma$ where $r_k$ denotes the response which is the output of each layer in the network. After rescaling to the size of the input image, each pixel in the image has multiple responses. After being normalized to have the unit standard deviation, they are added after multiplying with the corresponding weight $w_k$. The output is then convolved with the Gaussian kernel $G_\sigma$. $c(x,y)$ denotes the center bias, and it is added with a weight $\alpha$. The fixation probability map is calculated using the softmax: $p(x,y) = \frac{e^{o(x,y)}}{\sum_{x,y} e^{o(x,y)}}$. The network performance is measured using IG, and the result surpassed the state-of-the-art models at that time. Compared to the existing model (eDN), Deep Gaze I was able to increase the IG by 22% (Kümmerer et al., 2014).

Deep Gaze II replaced AlexNet backbone with the VGG-19 network (Simonyan & Zisserman, 2014). A readout network is added after VGG-19 features. This network outperforms its previous generation, also achieving highest IG among the models that existed that time (Kümmerer et al., 2017).

Deep Gaze IIE was able to combine multiple backbones to have better performance. The letter E stands for ensemble. Confidence calibration was used to explain the reason why this structure gets better performance. Confidence calibration is a way of showing whether the FDM has over-confident or under-confident results. The probabilities on the map are ordered and then divided into several bins, where these bins have the same sum-up probabilities. For a perfect model, each bin will have the same amount of actual fixations since they should have

the same probabilities. If the model is over-confident, fewer fixations will be spotted in high-probability areas than expected based on the model (they will be in low-probability areas). If the model is under-confident, the fixations will be more frequent in high-probability bins than in low-probability bins. The result shows Deep Gaze IIE are close to perfectly calibrated (Linardos et al., 2021).

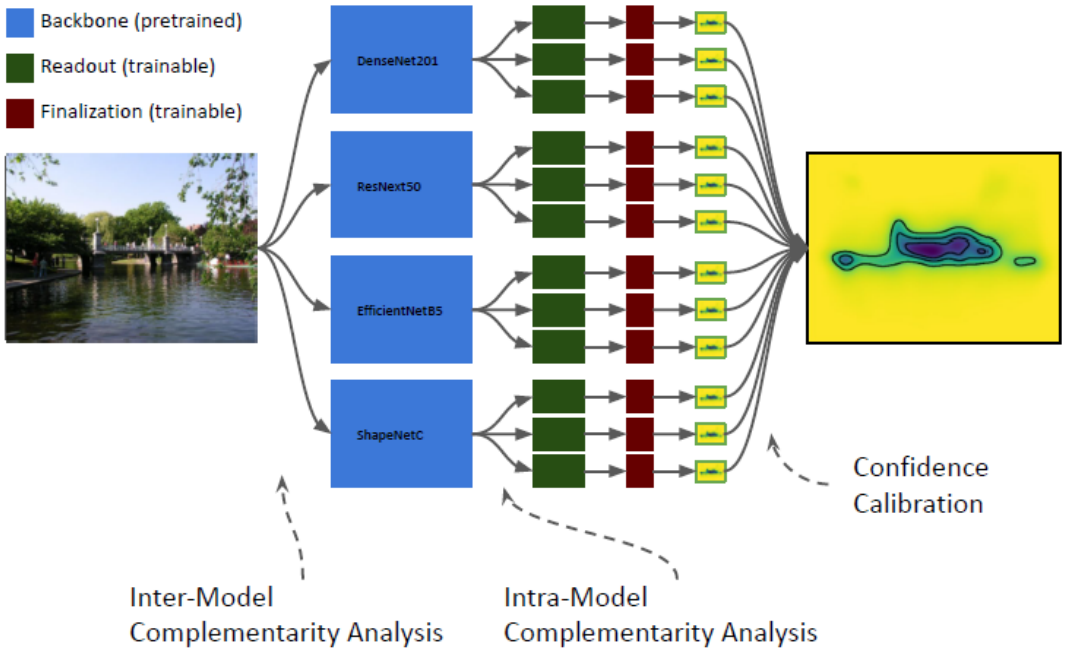


Figure 4-1 Deep Gaze IIE model structure. According to Linardos et al. (Linardos et al., 2021). Compared with Deep Gaze IIE, the VGG network was replaced with the then-state-of-the-art ImageNet network backbone and weighted on them. Confidence calibration can help understand the feasibility of the network.

Deep Gaze III added scanpaths into consideration. The encoding network consist of two parts: spatial priority network for processing images and scanpath network for processing scanpaths. The backbone for spatial priority network is DenseNet (Huang et al., 2017). The scanpaths are traceable to 4 previous fixations. After going through an encoding network, the outputs of spatial priority network and scanpath network are then fed through a fixation

selection network. The final FDM is produced after multiplication with a center bias. The model performance is measured using IG. The model managed to reproduce some features of behavior shown in empirical data. For example, the empirical data displays a tendency of saccades toward horizontal direction and Deep Gaze III also reproduced this phenomena.

In Deep Gaze III, the spatial priority network will yield only one saliency map for later process, the spatial priority map which contains the feature which is relevant to the network. This raised a question of whether the saliency map should be only one or multiple. Each saliency map represents one feature, and single map cannot represent multiple features because reweighting of the map is impossible. For example, one can hypothesize, luminance or more simple features would be key factors at a large distance from the current fixation, and more complex features would matter more at a short distance. Comparison was made between the network with multiple or single priority map. The result showed no improvement when using multiple feature maps. This suggested that the distance plays only a little role in free-viewing tasks (Kümmerer et al., 2022).

This chapter employs Deep Gaze IIE as the computer vision network for research and analysis. The utilization of Deep Gaze IIE for subsequent research and analysis is advantageous in three keyways: (1) Deep Gaze IIE ranks highly on the Tübingen Benchmark, indicating that the performance level of this network is still the most advanced in bottom-up networks. (2) In comparison to its predecessor, Deep Gaze IIE has undergone multiple improvements and has surpassed Deep Gaze I and Deep Gaze II in terms of performance level. In contrast to its subsequent version, Deep Gaze III, Deep Gaze III requires the scan path as an input parameter. This implies that the saliency map generated by Deep Gaze III depends on the scan path,

whereas in this study it only depends on the image. (3) Deep Gaze IIE is trained based on the benchmark truth obtained from experiments with human subjects, and the network has high credibility. The saliency map generated by it can be used as the benchmark truth for subsequent visual research. This thesis uses the saliency map generated by Deep Gaze IIE as a bottom-up clue to explore the characteristics of visual behavior. For the sake of clarity, the Deep Gaze IIE scores in this section are converted back to likelihood values using a power of two. As the quadratic function is monotonically increasing, the higher the converted value, the higher the model considers its significance.

## 3.2 Network Pre-processing

The use of Deep Gaze IIE offers three advantages: 1. It ranks near the top on the Tübingen Benchmark, indicating that its performance level in bottom-up networks remains state-of-the-art. 2. Compared to its predecessors, Deep Gaze IIE has undergone multiple improvements, surpassing both Deep Gaze I and Deep Gaze II in performance. In contrast to its successor, Deep Gaze III, which requires scanpaths as input parameters, the saliency maps generated by Deep Gaze IIE rely solely on the image itself. Additionally, Deep Gaze IIE was trained using ground truth data obtained from human subjects' experiments, ensuring high network reliability. Therefore, the saliency maps generated by Deep Gaze IIE can serve as benchmark ground truth for subsequent visual research. All saliency maps generated by Deep Gaze IIE in this study used images transformed from the COCO-Search18 dataset. The original Microsoft COCO images were not used because they have smaller dimensions, with an average resolution of 500 pixels. Deep Gaze IIE was trained on larger images, usually larger than 1024 pixels, and tests

have shown that it performs worse on small images compared to large images. All original outputs will be used for the subsequent analysis. Figure 4-2 displays an example of saliency maps generated by Deep Gaze IIE using COCO-Search18 images.

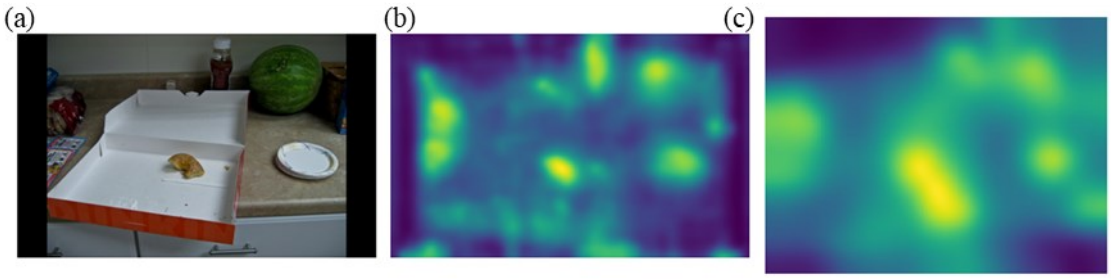


Figure 4-2. Comparison of saliency map outputs for images of different sizes. (a) Input original image. In the COCO-Search18 dataset, the image size is 1680×1050 pixels. In the Microsoft COCO dataset, the image size is 500×375 pixels. (b) Saliency map output for COCO-Search images. (c) Saliency map output for Microsoft COCO images.

After generating the fixation density map, Deep Gaze IIE applies Gaussian transformation and central bias correction to obtain the final saliency map. The code for Deep Gaze IIE provides an option to disable central bias correction. This correction weights the fixation density map after Gaussian blurring to simulate human visual patterns, as humans tend to fixate more towards the center of the screen. The saliency scores in the central region of the weighted images will significantly increase compared to the unweighted images. It is important to note that the saliency maps used in subsequent experiments are those without central bias correction, unless explicitly stated otherwise.

Relevant parameters were added to the fixation attribute dataset based on this. In practical implementation, the saliency map of each search image is computed and saved. The values of the fixation endpoints are determined by both the search image and the coordinates during

subsequent processing. Therefore, the Deep Gaze IIE score at each fixation endpoint is actually the Deep Gaze IIE score at the coordinates within the current search image.

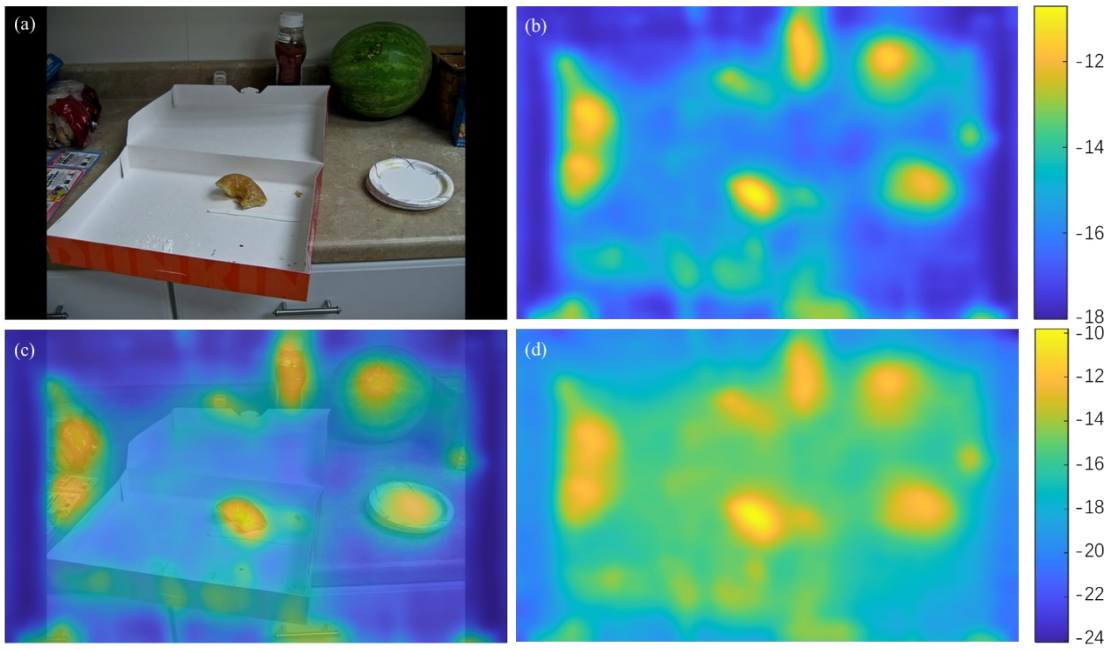


Figure 4-3. Saliency maps of input images and images generated by Deep Gaze IIE. (a) an example image with the search target being 'bottle' is shown in (a). (b) a saliency map generated by Deep Gaze IIE from the input image is presented. The saliency map is without central bias correction. Brighter areas indicate higher saliency. When the color tends towards blue, the model perceives lower saliency scores; conversely, when the color tends towards yellow, the model perceives higher saliency scores. The model predicts the likelihood of fixation at each pixel, and the resulting output image values are represented in a logarithmic scale with base 2. The likelihood ranges between 0 and 1, resulting in negative values. Higher values indicate higher likelihoods due to the monotonicity of the logarithmic function. (c) an overlay of images in (a) and (b) shows that the model perceives the highest saliency score at the 'croissant'. (d) A saliency map of Deep Gaze IIE after central bias correction. The central part of the image has higher saliency scores due to the higher weights.

## 3.3 Results

Figure 4-4 displays the distribution of saliency for first saccades based on saccade latency. The method, parameters and statistical analysis used to process latency are identical to those in

Chapter 2. It is important to note that saccades occurring before 125 milliseconds are considered pre-emptive and are excluded due to their lower probability of fixating on the search object. For the first saccade, whether it fixates or misses the search object, as shown in Figure 4-4, the saliency of the first saccade endpoints that fixate the search object is significantly higher than that of the first saccades that miss the search object (paired t-test, $p < 0.0001$).

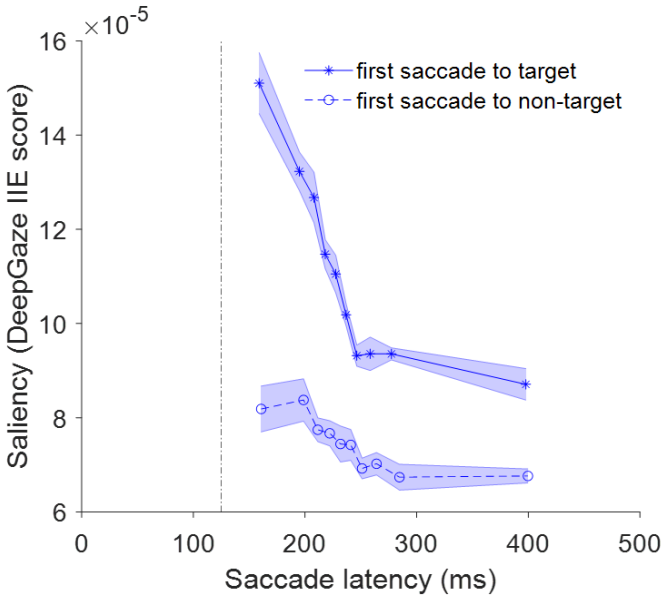


Figure 4-4. The points where the first saccade ends and fixates on non-target areas can be interpreted as instances where the saccade did not fixate on the search object. On average, the saliency for saccade ending points that fixate on search objects is higher than for those that miss the search objects, regardless of the latency of the first saccade. Both first saccades show a similar pattern: those with shorter latencies tend to end in areas with higher saliency scores. The starting points of the first saccade in both graphs are greater than 125 milliseconds because pre-emptive saccades have been excluded from the data.

The results indicate that first saccades tend to end at more salient places than second saccades when searching for a target, as shown in Figure 4-5. Specifically, the ending point of first saccades is more salient than that of second saccades (two-sample t-test, $p = 0.0070$).

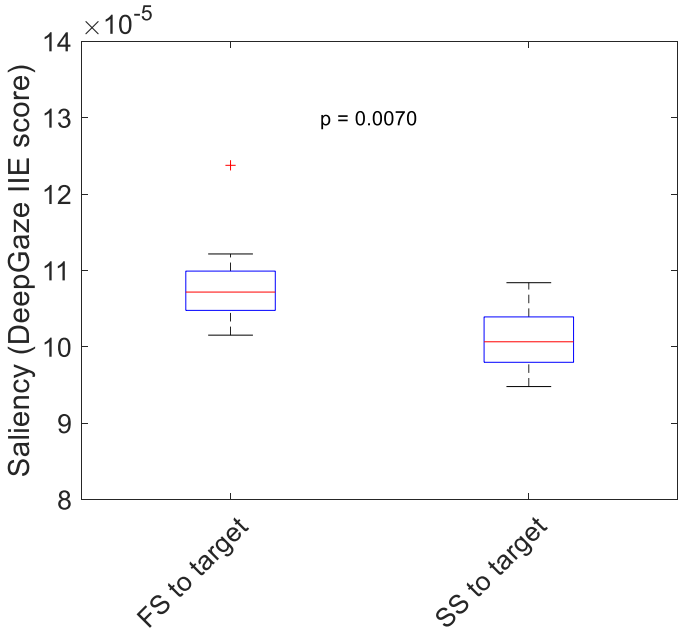


Figure 4-5. Saliency box plot for first saccades that go to the target and second saccades that go to the target. It shows that the ending point of the first saccade is more salient than that of the second saccade.

The section on saccades to non-target areas will be followed by an analysis of saccades to search targets. A saliency box plot is presented for first saccades, which are divided into two groups based on the type of second saccades that follow. The first saccades before short-latency second saccades end at less salient points than those before regular-latency second saccades.

If the first saccade is followed by a short-latency saccade, its Deep Gaze IIE score is lower (two-sample t-test, $p < 0.0001$, Figure 4-6(a)). This suggests that regions with low saliency prompt humans to quickly shift attention during the search process. In contrast, regions with high saliency are more likely to attract and hold attention for longer periods. The first saccade before regular-latency second saccades strongly supports this idea. Upon further analysis, it was found that the saliency of the ending points of first saccades before regular-latency second saccades is significantly higher than those of the first saccade before short-latency second

saccades, regardless of the distance of the first saccade endpoint from the boundary box of the search object and their Deep Gaze IIE scores. A comparable pattern is noted for the magnitude of the first saccade. These findings suggest that when the endpoint of the saccade does not fixate on the search object, the first saccades prior to regular-latency second saccades consistently end at more salient locations than the first saccades prior to short-latency second saccades.

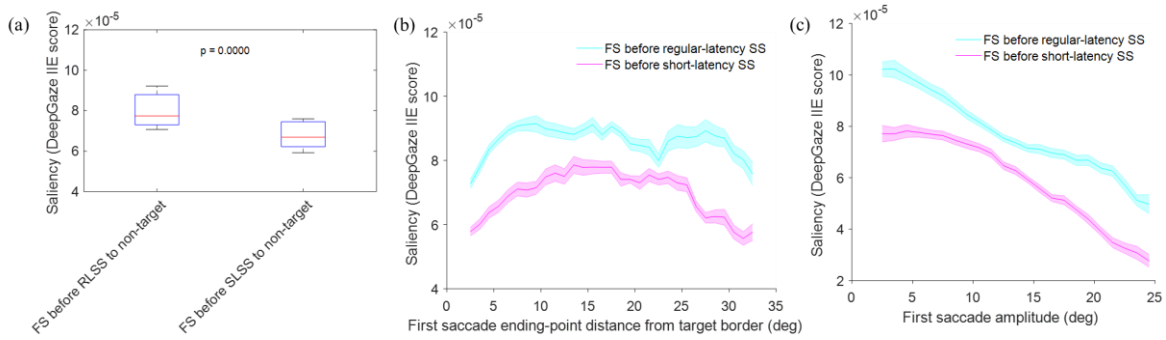


Figure 4-6. Saliency for first saccades before regular-latency second saccades that go to non-target areas and first saccades before short-latency second saccades that go to non-target areas. (a) Saliency box plot for the two types of first saccades. First saccades that do not go to the target area and are followed by regular-latency second saccades end at more salient points than those followed by short-latency second saccades. (b) The saliency distribution is presented as a function of the distance from the target border to the ending point of the first saccade. The study discovered that first saccades preceding regular-latency second saccades terminate at more salient points than those preceding short-latency second saccades, irrespective of the distance of the ending point of the first saccade from the target border. (c) The saliency distribution is shown as a function of the first saccade amplitude. This is consistent with the findings in Figure (b), where the ending point of first saccades before regular-latency second saccades is consistently more salient than those before short-latency second saccades.

The saliency of short-latency second saccades is significantly lower than that of regular-latency second saccades (two-sample t-test, $p = 0.0030$, Figure 4-7). It is important to note that

short-latency second saccades use information integrated after the previous saccade. When they do not fixate on the search object, their Deep Gaze IIE score is lower. This suggests that bottom-up saliency is not the reason for missing the target in the second saccade. If saliency were the influencing factor, the Deep Gaze IIE score of short-latency second saccades would be higher. Short-latency second saccades are goal-directed and not drawn to high saliency areas; in fact, they go to less salient areas.

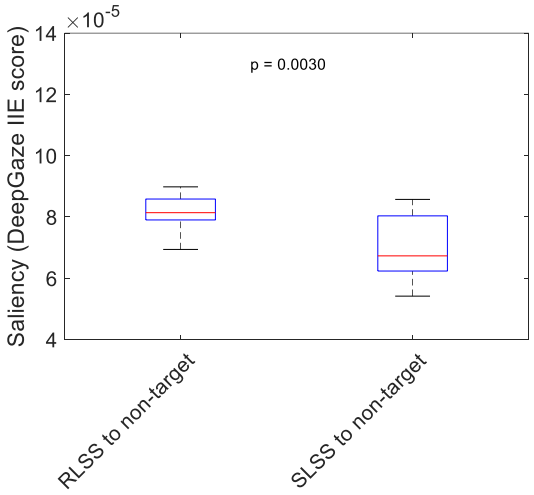


Figure 4-7. Saliency distribution for regular-latency second saccades and short-latency second saccades. The endpoint of regular-latency second saccades is more salient than that of short-latency second saccades.

Figure 4-6(a) shows that the saliency of ending points of first saccades before short-latency second saccade sequences, which do not fixate the search object, is significantly lower than those before regular-latency second saccades, which also do not fixate the search object. The lower Deep Gaze IIE scores for short-latency second saccades compared to regular-latency second saccades can be explained by the sources of information for short-latency second saccades. I propose a hypothesis: the lower Deep Gaze IIE score of the first saccade before short-latency second saccades is due to hypometria. These two sequences are classified based

on the nature of the second saccade; therefore, the first saccades preceding each type of second saccade have their respective properties. If this hypothesis is correct, it suggests that a hypometric Deep Gaze IIE score of the first saccade is more likely to trigger a short-latency second saccade.

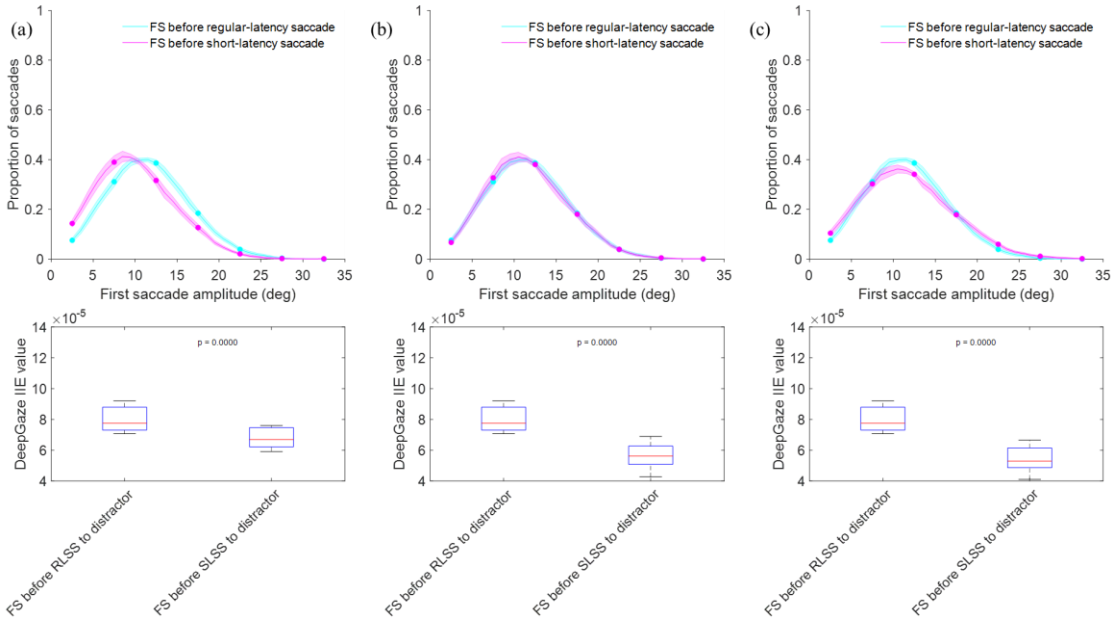


Figure 4-8. Probability distribution of the amplitude of the first saccade. (a), (b), and (c) each contain a probability distribution plot of the first saccade amplitude and box plots of Deep Gaze IIE scores for two types of first saccades. (a) Original plot. The data indicates that the amplitude level of the first saccade before short-latency second saccades is significantly lower than that before regular-latency second saccades. This difference is significant when comparing the peaks of the two types of first saccades. (b) By adding the amplitude of the first saccade before short-latency second saccades to bridge the gap with the amplitude of the first saccade before regular-latency second saccades, the difference in Deep Gaze IIE scores between the two types of first saccades remains significant. (c) The difference in Deep Gaze IIE scores between the two types of first saccades remains significant even after multiplying the amplitude of the first saccade before short-latency second saccades to compensate for the difference.

The study investigates the impact of saccade distance on the two types of first saccades.

Figure 4-8(a) shows the probability distribution of amplitude values for the two types of first saccades based on saccade amplitude. The results indicate that the amplitude values of the first saccade before regular-latency second saccades are higher than those before short-latency second saccades. Both types of first saccades display a similar pattern, with an increase followed by a decrease in amplitude proportion. The difference in amplitude values between these two types of saccades can also be inferred from the significant difference in their peaks. Higher amplitude values indicate a longer saccade distance. As the first saccade deviates from the center of the screen, an increase in amplitude means that the saccade ends at a location further from the center. If we attempt to bridge the gap between these two types of first saccades, and if the hypothesis is correct, the difference in Deep Gaze IIE scores between the two types of first saccades will become insignificant. Figures 4-8(b) and (c) compare the Deep Gaze IIE scores after employing two methods to bridge the gap: (1) adding the absolute average difference between the amplitude values of the first saccade before short-latency second saccades and those before regular-latency second saccades, (2) multiplying by the reciprocal of the ratio between the amplitude values of the first saccade before short-latency second saccades and those before regular-latency second saccades. Both methods aim to compensate for the difference in amplitude values by making the first saccade before short-latency second saccades fixate towards a position farther from the center. Even after bridging the gap, the difference in Deep Gaze IIE scores between the two types of first saccades remains significant (two-sample t-test, $p < 0.0001$ in all three conditions, Figure 4-8). Even with compensation for amplitude differences, the first saccade before short-latency second saccades cannot return to normal levels, contradicting the hypothesis. Both factors affecting the first saccade contradict

the hypothesis, suggesting that the first saccade being hypometric is not the trigger for the second saccade. However, it cannot be denied that the Deep Gaze IIE score of the first saccade before short-latency second saccades is hypometric. This indicates that the phenomenon of being hypometric is caused by other unknown factors.

This chapter examines the relationship between various types of saccades and the Deep Gaze IIE scores generated for each search image using the Deep Gaze IIE model. Saccades directed towards search targets end at more salient points than those directed towards non-target areas. In addition, this study addresses the phenomenon where the saliency of the first saccade ending points before short-latency second saccades is lower than that before regular-latency second saccade ending points. The study of features such as latency and amplitude of these two types of first saccades confirms that short-latency second saccades are not triggered solely by first saccades below average level. When not fixating on search targets, the saliency for endpoints of short-latency second saccades is lower. This suggests that the second saccades are not attracted to more salient areas.

## 3.4 Discussion

This chapter employs the bottom-up model for in-depth analysis, whereas the top-down model is not further studied. The rationale for not selecting the top-down model is that there is no satisfactory ground truth. Top-down models, such as Mask-RCNN, can process the search scene and ultimately output the possible object categories and confidence scores in the search scene. The training of this type of deep learning network is carried out under the condition that each object category of the search image is labeled. There are several object categories in total.

The model will first divide the area where the object may exist, and then score the confidence of each object category in each area. If the confidence score of a category is the highest, then the model believes that this area belongs to this object category. To achieve this, the model often only needs to widen the gap between the confidence scores belonging to the category and the confidence scores that do not belong to the category. This allows the model to maintain good performance in different search scenarios. However, as Linardos proposed in his trust calibration judgment method, the top-down network often deliberately increases the gap between the target category and the non-target category. This makes the network overconfident in the generated saliency map. This is due to the fact that the top-down network lacks the benchmark truth generated by human behavior as training data, which also renders top-down analysis more challenging than bottom-up analysis (Linardos et al., 2021).

The Deep Gaze IIE saliency score indicates that when the saccade fixates the target object, the end point of the first saccade is more significant than the end point of the second saccade. This may be due to the fact that both saccades are saccades that fixate the object. When the first saccade fixates the target object, only one saccade was performed. When the second saccade fixates the object, it must be the case that the previous saccade did not fixate the target object. Humans are affected by bottom-up saliency, which means that they will end up in a more salient area in the first saccade. When the first saccade does not fixate the target object, the saliency score of the target object may not be the most significant in the search scene. Consequently, when the second saccade fixates the search object, the saliency score of its end point will be lower. It is also noteworthy that the end point of the first saccade that fixates the target object is more significant than the end point that does not fixate the target object. This

advantage is more pronounced when the latency is shorter. The shorter the latency, the shorter the processing time. At this juncture, it is evident that visual behavior is influenced by the bottom-up mechanism. This phenomenon also corroborates the notion that humans tend to direct their fixation towards areas with higher saliency scores in the first saccade.

When a saccade does not fixate the target object, the first saccade preceding the short-latency second saccade will terminate in a less salient area. This thesis has demonstrated that short-latency second saccades are not corrective saccades; rather, they are goal-directed saccades. Prior to the initiation of the first saccade, the visual system is likely to have identified the target object. However, as the execution of the first saccade has already been determined, it cannot be cancelled. If the saliency score of the area where the first saccade ends is significantly higher than that of the search object, the visual system is likely to cancel or delay the second saccade due to the current saccade's location being in a more salient area. This illustrates that when the first saccade does not fixate the search object and the saliency score of its location is higher, the first saccade is more likely to be attracted for a longer duration. The latency of the second saccade is defined as the interval between the end of the first saccade and the beginning of the second saccade. If the area attracts the first saccade for a longer duration, then the second saccade is less likely to become a short-latency second saccade. This also indicates that the first saccade before the short-latency second saccade will end in a less salient area compared to the first saccade before the regular-latency second saccade.

When a saccade does not fixate the target object, the saliency score at the endpoint of the short-latency second saccade is significantly lower than that of the regular-latency second saccade. This is in accordance with research findings, which indicate that when the first saccade

misses the target object, the endpoint of the second saccade has a lower saliency score. The information for the short-latency second saccade is derived from the first saccade, which explains why the area the short-latency second saccade moves to will also have a lower saliency score. In contrast, the regular-latency second saccade obtains information that is updated after the first saccade ends, which increases the likelihood that it will move to an area with a higher saliency score.

# Chapter 4
# Conclusions

## 4.1 Key Findings and Significance

This study examines the temporal characteristics of human goal-directed visual search through public datasets and bottom-up saliency models. A multitude of models have demonstrated the capacity to simulate where human vision looks, yet research on temporal characteristics remains relatively scarce. Consequently, this thesis initially studies the temporal characteristics of visual search through public datasets. Furthermore, this thesis employs a bottom-up deep learning model to analyze the bottom-up aspect of visual search. The findings of this study are presented in the following summary:

1. This study employs the COCO-Search18 dataset to investigate the temporal characteristics of visual search. The research focuses on the first and second saccades and categorizes their latencies, revealing that the latency of the second saccade is typically shorter than that of the first saccade. For the second saccade, short-latency second saccades, defined as those with a latency of less than 125 milliseconds, exhibit a higher target fixation rate. Short-latency second saccades account for 45% of all second saccades. This higher target fixation rate is not due to the corrective nature of these saccades; subsequent research has found that short-latency second saccades have a higher goal-directedness towards eccentric objects. The trigger of short-latency second saccades is also conditional, influenced by the presence of the search object. These results indicate that short-latency second saccades are goal-directed.

2. We continue the study of visual behavior characteristics using bottom-up saliency maps. This thesis utilizes saliency maps generated by the top-performing Deep Gaze IIE model to further investigate the previously mentioned public dataset. The findings are as follows: Saccades that fixate the search object and those that do not exhibit different patterns. For the first saccade, the endpoint of saccades that fixate on the search target is typically more salient than the endpoint of saccades that do not. For the first and second saccades that fixate on the target, the endpoint of the first saccade is more salient than that of the second. When a saccade misses the target, the endpoint of short-latency saccades is typically less salient, and the first saccade before a short-latency second saccade also tends to fixate on less salient non-target areas. The endpoint of short-latency second saccades is less salient than that of regular-latency second saccades. Short-latency second saccades are primarily goal-directed. When the first saccade misses the target object, the execution of the second saccade is influenced by the saliency score at the endpoint of the first saccade. If this point has a high saliency score, it will attract attention and cause a longer fixation at this point, reducing the likelihood of triggering a short-latency second saccade. Furthermore, the information utilized for short-latency second saccades is derived from the first saccade. Consequently, if the first saccade fails to identify the search object, the probability of a second saccade terminating in a more salient area is diminished.

The findings of this study indicate that when searching for category-defined goals in photos of complex everyday scenes, human searchers employ a gratification strategy. It was observed that humans do not try to minimize the number of eye movements during search tasks, which is consistent with previous studies that have shown a bias towards exploration (Araujo

et al., 2001; Wu & Kowler, 2013). The timing mode continuously adjusts based on the information it receives and internal models that determine when to delay and when to explore. Timing decisions may consider the marginal benefits of waiting too long versus the benefits of exploration. This ensures that the cost of additional eye movements to distractors is minimized. If perception restarted with each fixation, the information conveyed within short-latency saccades and across saccades would be lost in the update (Findlay et al., 2001). The integration and transfer of information across saccades plays an important role in natural vision.

## 4.2 Future Work

For each endpoint's Deep Gaze IIE score, this study defines it as the saliency score of the pixel at the endpoint, which is the simplest application method. However, errors are inevitable in the data collection process, and deviations in the fixation points collected by the eye tracker may lead to changes in the saliency scores of the pixels. In future research, the saliency score of the saccade endpoint can be obtained by calculating the weighted average of the saliency scores around that point. Additionally, this study defines interference objects as a set of objects in an image that are not search objects from the Microsoft COCO dataset. The Microsoft COCO dataset includes objects that are not listed in the Microsoft COCO object labels, such as 'watermelon', which belongs to the category of fruits.

Regarding bottom-up saliency features, a specific region is susceptible to interference. Chapter 3 consistently highlights that the Deep Gaze IIE fixates on the search target with significantly higher values than on non-target objects. This implies that in the saliency map generated by Deep Gaze IIE, the search target has the highest saliency score. Figure 4-2

demonstrate that the model identifies regions other than the search target 'bottle' as most salient, such as the distractor 'croissant'. However, none of the participants in the experiment were affected by the 'croissant' in this image. There are two possible explanations: (1) The pastry known as a 'croissant' is located at the center of the image. The first fixation begins at the center of the screen, capturing information from this region. (2) Despite the 'croissant' having a high saliency score, its dissimilarity to the bottle inhibits bottom-up saliency during visual search, which is instead influenced by top-down saliency. In order to redefine the definition of bottom-up distractors, they can be defined as areas with higher average saliency scores within a 2×2° square. To obtain numerical characteristics of the search target and distractors in the saliency map generated by Deep Gaze IIE, the saliency scores of the search target can be compared with the saliency scores of the first two or three distractors, excluding the search target.

In addition to bottom-up mechanisms, there are also top-down mechanisms in visual processing. These mechanisms are more target-directed. However, research in this area remains limited due to the lack of datasets with benchmark ground truth data from humans. Several deep learning networks are still attempting to simulate top-down mechanisms (Ding et al., 2022; Yang et al., 2023; Zhang et al., 2018). These networks are mostly improvements on previous networks based on convolutional neural networks or transformer networks, with limited application of visual mechanisms. Subsequent research on top-down mechanisms should begin here. Design experimental paradigms based on top-down mechanisms in advance and conduct in-depth studies on visual search mechanisms using human behavioral data as benchmark ground truth. Additionally, research on different object categories in this thesis can be enhanced through top-down mechanisms.

# References


Araujo, C., Kowler, E., & Pavel, M. (2001). Eye movements during visual search: The costs of choosing the optimal path. *Vision research*, *41*(25-26), 3613-3625.

Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision research*, *19*(9), 967-983.

Bomatter, P., Zhang, M., Karev, D., Madan, S., Tseng, C., & Kreiman, G. (2021). When pigs fly: Contextual reasoning in synthetic and natural scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Borenstein, E., Sharon, E., & Ullman, S. (2004). Combining top-down and bottom-up segmentation. 2004 Conference on Computer Vision and Pattern Recognition Workshop,

Borji, A., Tavakoli, H. R., Sihite, D. N., & Itti, L. (2013). Analysis of scores, datasets, and models in visual saliency prediction. Proceedings of the IEEE international conference on computer vision,

Burlingham, C. S., Sendhilnathan, N., Komogortsev, O., Murdison, T. S., & Proulx, M. J. (2024). Motor "laziness" constrains fixation selection in real-world tasks. *Proceedings of the National Academy of Sciences*, *121*(12), e2302239121.

Caspi, A., Beutter, B. R., & Eckstein, M. P. (2004). The time course of visual information accrual guiding eye movement decisions. *Proceedings of the National Academy of Sciences*, *101*(35), 13086-13090.

Chakraborty, S., Samaras, D., & Zelinsky, G. J. (2022). Weighting the factors affecting

attention guidance during free viewing and visual search: The unexpected role of object recognition uncertainty. *Journal of Vision*, *22*(4), 13-13.

Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., & Zelinsky, G. (2021). Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, *11*(1), 8776.

Chen, Y., Yang, Z., Chakraborty, S., Mondal, S., Ahn, S., Samaras, D., Hoai, M., & Zelinsky, G. (2022). Characterizing target-absent human attention. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Darrien, J. H., Herd, K., Starling, L.-J., Rosenberg, J. R., & Morrison, J. D. (2001). An analysis of the dependence of saccadic latency on target position and target characteristics in human subjects. *BMC neuroscience*, *2*, 1-8.

Ding, Z., Ren, X., David, E., Vo, M., Kreiman, G., & Zhang, M. (2022). Efficient zero-shot visual search via target and context-aware transformer. *arXiv preprint arXiv:2211.13470*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Federmeier, K. D., & Schotter, E. R. (2020). *Gazing Toward the Future: Advances in Eye Movement Theory and Applications*. Academic Press.

Findlay, J. M., Brown, V., & Gilchrist, I. D. (2001). Saccade target selection in visual search: The effect of information from the previous fixation. *Vision research*, *41*(1), 87-95.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual review of neuroscience*, *26*(1), 181-206.

Girshick, R. (2015). Fast r-cnn. Proceedings of the IEEE international conference on computer vision,

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 6-6.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. Proceedings of the IEEE international conference on computer vision,

Henderson, J. M., & Pierce, G. L. (2008). Eye movements during scene viewing: Evidence for mixed control of fixation durations. *Psychonomic Bulletin & Review*, *15*, 566-573.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Huber-Huber, C., Buonocore, A., & Melcher, D. (2021). The extrafoveal preview paradigm as a measure of predictive, active sampling in visual perception. *Journal of Vision*, *21*(7), 12-12.

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. Proceedings of the 2010 symposium on eye-tracking research & applications,

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. 2009 IEEE 12th international conference on computer vision,

Jurewicz, K., Liao, B., & Krishna, S. (2023). Information integration across saccades plays a prominent role during goal-directed viewing of everyday scenes. *bioRxiv preprint bioRxiv:10.31234*.

Katsuki, F., & Constantinidis, C. (2014). Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, *20*(5), 509-521.

Kirchner, J. P. (2017). *Do saccades in human visual search show evidence of change-of-mind?* University of Göttingen].

Krishna, B. S., Ipata, A. E., Bisley, J. W., Gottlieb, J., & Goldberg, M. E. (2014). Extrafoveal preview benefit during free-viewing visual search in the monkey. *Journal of Vision*, *14*(1), 6-6.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kümmerer, M., Bethge, M., & Wallis, T. S. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, *22*(5), 7-7.

Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2018). Saliency benchmarking made easy: Separating models, maps and metrics. Proceedings of the European Conference on Computer Vision (ECCV),

Kümmerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. Proceedings of the IEEE international conference on computer vision,

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part

V 13,

Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. Proceedings of the IEEE/CVF International Conference on Computer Vision,

McPeek, R. M., & Keller, E. L. (2002). Superior colliculus activity related to concurrent processing of saccade goals in a visual search task. *Journal of Neurophysiology*, *87*(4), 1805-1815.

McPeek, R. M., Skavenski, A. A., & Nakayama, K. (2000). Concurrent processing of saccades in visual search. *Vision research*, *40*(18), 2499-2516.

Mokler, A., & Fischer, B. (1999). The recognition and correction of involuntary prosaccades in an antisaccade task. *Experimental Brain Research*, *125*, 511-516.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, *45*(18), 2397-2416.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. Proceedings of the 2000 symposium on Eye tracking research & applications,

Samiei, M., & Clark, J. J. (2022). Predicting Visual Attention and Distraction During Visual Search Using Convolutional Neural Networks. *arXiv preprint arXiv:2210.15093*.

Schwetlick, L., Rothkegel, L. O. M., Trukenbrod, H. A., & Engbert, R. (2020). Modeling the effects of perisaccadic attention on gaze statistics during scene viewing. *Communications*

*biology*, *3*(1), 727.

Sharika, K. M., Ramakrishnan, A., & Murthy, A. (2008). Control of Predictive Error Correction During a Saccadic Double-Step Task. *J Neurophysiol*, *100*, 2757-2770.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 4-4.

Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, *9*(5), 379-385.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Wu, C.-C., & Kowler, E. (2013). Timing of saccadic eye movements during visual search for multiple targets. *Journal of Vision*, *13*(11), 11-11.

Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., & Hoai, M. (2020). Predicting goal-directed human attention using inverse reinforcement learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., & Samaras, D. (2022). Target-absent human attention. European Conference on Computer Vision,

Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., & Samaras, D. (2023). Predicting Human Attention using Computational Attention. *arXiv preprint arXiv:2303.09383*.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any Waldo with zero-shot invariant and efficient visual search. *Nature communications*, *9*(1), 3730.