COMPUTATIONAL MODELS OF BOTTOM-UP AND TOP-DOWN VISUAL ATTENTION

MENGMI ZHANG

(*B.Eng.(Hons.), NUS*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2019

Declaration

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Mengmi Zhang
23rd January 2019

Acknowledgments

My profound gratitude goes to my mentors and research advisors, Dr. Jiashi Feng, Dr. Joo Hwee Lim, Dr. Gabriel Kreiman, Dr. Keng Teck Ma, Dr. Shih-Cheng Yen, and Dr. Qi Zhao, for supporting me during these past three years. Jiashi is one of the most knowledgable persons I have ever seen in machine learning. He has always provided insightful discussions about algorithm design and even the details of mathematical derivations and software implementations. As one of my closest mentors and research advisers, Joo Hwee has been supportive, has motivated and guided me with the broad scientific picture on visual attention, memory and search while inspired me with sufficient freedom to define relevant problem statements and pursue their answers. I am also very impressed with his leadership skills. Whenever I encounter adminstration and project management problems, he is the first person I can ask help for. I am also grateful to Gabriel, who is lively, enthusiastic, responsible and energetic. He is my primary resource for getting my science questions patiently answered and having my science results critically examined. I would also like to thank Keng Teck, who is not only my mentor but also one of my best friends. I salute his rich ideas and genuine opinions about my research. Finally, I am indebted to Shih-Cheng, Qi Zhao and my thesis advisory committee members, Dr. Ong Sim Heng and Dr. Liyuan Li, for helpful career advice and suggestions in general.

A strong support team is important to survive and stay sane in my graduate school. I am lucky enough to have all the amazing friends all over the world who can recognize themselves. In particular, I thank Chunchun Liu, Weinan Xu, Xiatong Zhang, Mona Ragab Sayed, Pan Zhou, Yunpeng Chen, Sarah Yuruo Shan, Jiani Wu, Alex Bowen Yang, Joseph Olson, Jerry Wang, Jiye Kim, Yuchen Xiao, Pranav Misra, Eleonora Iaselli, Yen-Ling Kuo, Kevin Smith and Farah Ahmed Wick who have given me encourage-

ments, fun, supports, advises, help, care, refreshing perspectives, and positive attitudes towards challenges.

My very special thanks go to my parents who have always given me unconditional love and supports.

This work was not possible without supports from Reverse Engineering Visual Intelligence for cognitive Enhancement (REVIVE) programme, Singapore Agency for Science, Technology and Research (ASTAR). I would also like to express my gratitude to Boston Children's Hospital (BCH), Harvard Medical School (HMS) and Center for Brain Minds and Machines (CBMM) in MIT where the latter half of this work was completed.

Contents

Summary

Humans have the remarkable ability of prioritizing the sequence of eye-fixations to survey a complex environment. This ability enables us to react rapidly to environmental changes, and maximize the amount of useful information obtained from visual inputs, despite the limited high-acuity processing capability and memory capacity in the brain.

Some objects automatically pop out in the environment and attract our visual attention in a bottom-up manner. Most classic computational models of the bottom-up visual attention adopt the 2-stage visual attention paradigm for predicting a sequence of eye fixations: computing saliency maps based on feature integration theory, and deciding the order and location of these fixations from saliency maps based on winner-take-all and inhibition of return principles. Both the processed information of the foveated visual input (the what information) and the current fixation location (the where information) have influences on selecting the next fixation location. The 2-stage visual attention paradigm greatly simplifies the temporal dependencies across fixations with two assumptions: first, visual inputs over all spatial locations are processed equally without fovea effect; second, all previously visited locations are inhibited forever without memory decay. Instead of explicitly defining the temporal dynamics in fixation sequences, the first body of work in this thesis describes several computational models which foveate at different parts of the visual inputs and automatically learn to exploit temporal dependencies across fixations in an end-to-end supervised manner from human eye movement data. In addition, we also present the first computational model to anticipate dynamics of fixation sequence in the near future. The experimental results suggest that the eye fixation prediction and anticipation performance of all our models is comparable to or even better than state-of-the-art algorithms.

There is strong neurophysiological evidence supporting that visual attention is more

than a feed-forward spatially filtering process. We next address the question of how visual attention modulates the visual processing pathway in a top-down manner. The second body of work in this thesis is related to pioneering a biologically-inspired computational visual search model that can locate targets without exhaustive sampling and generalize to look for novel targets on static images with zero training on these targets. The model provides an approximation to the mechanisms integrating bottom-up and top-down signals during visual search in natural scenes.

In sum, this thesis describes several computational implementations of integrated bottom-up and top-down visual attention that learn to exploit the temporal dynamics across fixations via supervised training, and modulate the visual processing pathway in a top-down fashion via zero-shot learning. These models not only contribute to the development of artificial intelligence in terms of state-of-the-art prediction of fixation locations and anticipation in images, as well as egocentric and third-person videos, but also provide insights into the mechanisms of human visual attention. The latter was demonstrated by closely approximating the behavior of human eye movements in a series of psychophysics experiments.

List of Tables

List of Figures

List of Symbols

Latin variables

t	Time step or fixation stage
I	Input image and the subscript often denotes input conditions depending on the context
$L(\cdot)$	Loss function

Greek variables

ϕ	Network parameters

Acronyms

2D	Two Dimension
3D	Three Dimension
VA	Visual Attention
RNN	Recurrent Neural Network
2D-CNN	2D Convolutional Neural Network
3D-CNN	3D Convolutional Neural Network
SS	Scanpath Score
AUC	Area Under the Curve
sAUC	Shuffled-AUC
NSS	Normalized Scanpath Saliency
AAE	Average Angular Error
CC	Correlation Coefficient
PR	Precision-Recall Curve
DSNN	Deep Scanpath Neural Net

FNN	Foveated Neural Network
DFG	Deep Future Gaze model
GAN	Generative Adversarial Network
IVSN	Invariant Visual Search Network

Chapter 1

Introduction

Visual attention is a cognitive process of selectively concentrating on regions of the environment while ignoring the rest [1]. Fixation or gaze (where human is looking at), as a perceptual variable, cues overt visual attention. Humans do not analyse an entire scene at once. In order to rapidly react to environmental changes and maximize the amount of information obtained from visual inputs with the constraints of limited computational resources and memory capacity, prioritising fixation sequences is critical for humans. Moreover, the entire scene representation, which is internally and consistently built up in memory by integrating information at the different fixations over time, can guide subsequent eye movements for decision making and reasoning [2].

We can classify visual attention into two distinct functions: the bottom-up attentional mechanism where external stimuli attracts attention in a bottom-up manner due to their inherent features, such as the visual contrast relative to the backgrounds; and the top-down attentional mechanism driven by the desires and current goals, such as searching for a cup of water when a person feels thirsty. A computational model of visual attention address the descriptive process for how bottom-up and top-down attentional guidance is computed. Subsequently, the computational model can be evaluated by comparing with how the humans perform given the similar experimental stimuli. Chapter 2 provides a brief account of a wide range of related works on visual attention in the fields of neuroscience, psychology and artificial intelligence. It is to examine germane research on visual attention and its connection with memory to bring insights into developing biologically-plausible computational visual attention models.

1.1 Bottom-up, Saliency-based Visual Attention Models

While there have been recent works devoted to understanding bottom-up visual attention, *e.g.* [3, 4, 5], the studies often adopt the 2-stage paradigm for predicting a sequence of eye fixations: computing saliency maps based on feature integration theory [6], and deciding the order and location of these fixations from saliency maps based on winner-take-all [7] and inhibition of return principles [3]. In this paradigm, the processed information of the foveated visual inputs (the what information) at the past fixations has been missing while predicting the next fixation location.

Moreover, these studies focus on saliency prediction on static images and it is not clear how these works can generalize to predict gazes on video frames where there is motion information between adjacent frames. Specifically, over normal videos, egocentric videos captured from first-person perspectives in head-centered coordinate system, involve head motion dynamics. The effect of head motion and foreground object motions on gaze dynamics is not fully understood.

Our work also extends the gaze prediction problem to go beyond current video frames [8, 9] and presents the novel and important problem of *gaze anticipation*: predicting future gaze locations within a few seconds ahead. Unbeknown to how the future looks like, gaze anticipation is a more challenging problem over gaze prediction.

There are many applications where gaze anticipation turns out to be useful in enabling the predictive computation. These include but not limited to interactive commercialization [10], human-machine interaction [11], and alter system for constantly monitoring drivers' attention [12]. One particular example is in Virtual Reality (VR) wearables. VR headsets have become increasingly popular nowadays. As one category of egocentric devices, they often have to synthesize virtual realities in real time during interaction from users at the cost of high computation power [13]. Gaze anticipation plays important roles in facilitating these computation-hunger systems to plan ahead and increase their buffer time [14]. Based on anticipated gaze locations within the next few seconds, pre-rendering of the virtual scenes provides smoother presentations in virtual reality and hence better user experience [13]. In interactive e-commerce design [10] with gaze anticipation, remote information servers could also benefit in pre-fetching contextual e-advertisements and prompting to the consumers without noticeable time

delays.

The interplay of head motion, foreground object motions, as well as gaze motions, intrigues us to develop computational bottom-up visual attention models capable of interacting with various types of motions and capturing the temporal dynamics across fixations. In computer vision research, there have recently been great strides using deep learning for various computer vision tasks which achieves fascinating performances. In the first part of the thesis, we have focused on developing deep learning based models which automatically learn to exploit temporal dependencies across fixations in an end-to-end supervised manner trained on human eye movement data. Inspired by the neurophysiological findings, we conducted several experiments involving subsystems of visual attention including the fovea, the bottom-up ventral stream in the visual cortex and the working memory. In addition to the significant boost in fixation prediction and anticipation performances, we also analysed the features from the learnt attention models where some analysis suggest alignments with cognitive findings, such as the changes of spatial bias during the scanpath prediction on static images in a free-viewing task and the role of foveated visual inputs and motion information between adjacent frames in fixation prediction and anticipation.

We summarize the contributions of each chapter in the first part of the thesis as below:

In Chapter 3, we introduce a novel recurrent neural network, Deep Scanpath Neural Network (DSNN), which integrates the information from all the past eye fixations to predict the next fixation location. We evaluate DSNN in three challenging benchmark datasets on static images. DSNN demonstrates an unprecedented scanpath prediction accuracy, while it obtains a competitive predictive accuracy of the saliency map with state-of-the-art models. Our analysis of the learnt model reveals that the recurrent connections in DSNN are effective to improve the predictive visual scanpath accuracy, and it also shows the emergence of a temporally changing spatial bias during the scanpath prediction.

In Chapter 4, we propose a novel deep convolution neural network, Foveated Neural Network (FNN), to predict gaze on current frames in egocentric videos. In the network, we get inspirations from human visual system and introduce a fovea module respon-

sible for sharp central vision. FNN analyzes and encodes the retina-like visual inputs from the region of interest on the previous frame. The hidden representations of the previous frame and the feature maps of the current frame are fused to guide the gaze prediction on the current frame. As additional input to FNN, we introduce the dense optical flow between these adjacent frames which represents motion information. In the experiments, we demonstrate that FNN outperforms the state-of-the-art algorithms in the publicly available egocentric video dataset. The analysis of FNN suggests that both the hidden representations of the foveated visual input from the previous frame and the motion information between adjacent frames contribute to the improved gaze prediction performance in egocentric videos.

In Chapter 5, we extend the conventional gaze prediction problem to go beyond current frames and introduce a new problem of gaze anticipation on future frames. To tackle this problem, we propose a generative adversarial network based model, named as Deep Future Gaze (DFG), encompassing two pathways: DFG-P anticipates gaze prior maps according to the task influences from input frames; DFG-G models both semantic and motion information in future frame generation useful for gaze anticipation. DFG-G consists of two networks: a generator and a discriminator. In the generator, a two-stream spatial-temporal convolution architecture (3D-CNN) generates future frames by explicitly untangling the foreground and background motions. The generator then attaches another 3D-CNN to anticipate gaze on these synthetic frames. The discriminator provides additional feedbacks to the generator by distinguishing the synthetic frames of the generator from the real frames. DFG significantly outperforms all competitive baselines on the publicly available egocentric and third person video datasets. Without any fine-tuning, compared with state-of-the-art methods, DFG also achieves better performance of gaze prediction on current frames in egocentric and third person videos.

1.2 Target-driven Visual Attention Models by Top-down Modulation

Visual search is a versatile paradigm for studying top-down visual attention. In the second part of the thesis, we address the top-down visual attention problem in the context

of visual search.

We encounter visual search challenges in our daily life, such as finding a friend in a party crowd or searching for a car in a parking lot. There are four key properties that visual search must fulfill: (1) selectivity (to distinguish the target from distractors in a cluttered scene), (2) invariance (to localize the target despite changes in its appearance or even in cases when the target appearance is only partially defined), (3) efficiency (to localize the target as fast as possible, without exhaustive sampling), and (4) zero-shot training (to generalize to finding novel targets despite minimal or zero prior exposure to them).

Due to the explosive number of combinations of the target variations and the visual scene complexity, visual search is a computationally difficult task. In most cases, observers do not look for an identical match to the target at the pixel levels. Instead, they intend to find the target varying in color, rotation, occlusion, scale, illumination, and other transformations. Moreover, observers may be interested in looking for any target object belonging to a generic category (*e.g*. finding any spoons, rather than a specific one). Researchers have taken efforts in addressing the invariance problem in visual recognition where object identification is robust to any transformations at pixel levels (*e.g*. [15, 16, 17], among many others). The fundamental challenge in invariant object recognition has led to the development of hierarchical computational models that progressively build transformation-tolerant features selective for object identification.

Compared with the enormous body of work in bottom-up object recognition models, there have been fewer works devoted to the invariance problem in visual search. Behavioral [18, 19, 20] and neurophysiological [21, 22, 23] visual search studies focus on the identical target search. In those experiments, the appearance of the target object has been very well defined in each trial. *e.g*. observers are asked to search for a vertical green bar or an identical match to a picture of a chair. There have been other works investigating the ability to search for rotated faces with respect to a canonical viewpoint [24]; however, the ambiguity in the target appearance is minimal and it circumvents the critical challenge in invariant visual search. In subsequent studies on hybrid visual search, observers have to look for multiple objects but the appearances of these objects are fixed [25]. In [26, 27], scientists have evaluated reaction times during visual search

for generic categories as a function of the number of distractors, but it is not clear how these findings could be generalized to invariant visual search in complex natural scenes.

Template matching methods demonstrate selectivity to distinguish a target from distractors; however, it performs poorly in invariant visual search since it fails to find transformed targets robustly. In computer vision, object detection and image retrieval approaches address object localization problem at the cost of extensively being trained with the sought targets and exhaustively scanning the entire image via sliding windows [28, 29, 30, 31]. Moreover, there is no evidence showing that these computer vision approaches bear resemblance to the neurophysiological mechanism of visual search in human brains. Instead of sequential scanning and class-specific supervised training in heuristic algorithms, observers are capable of performing rapid search by moving their eyes in a target-driven manner, even if the exact appearance of target is unknown or there is merely single-trial exposure to the novel target.

Behavioral [18, 32, 33, 34] and neurophysiological [22, 23, 35, 36] studies suggest that task goals, such as the sought target in a visual search paradigm, guides attention allocation and eye movements when presented with a search image. Goal-dependent modulation originating from frontal cortical structures [23, 37] projects onto visual cortex structures in a top-down fashion [35, 38]. Several computational models have been proposed to describe visual search behavior or the modulation of responses in visual cortex [19, 20, 21, 33, 39, 40, 41, 42, 43]).

In Chapter 6, we propose a zero-shot deep architecture, Invariant Visual Search Network (IVSN), which maps the discriminative power from object recognition models to visual search. IVSN takes two inputs, a target object and a search image, and produces a sequence of fixations. Distinct from heuristic template matching, IVSN can efficiently find the target based on the overall attention map regardless of variations in the sought target within the search image (including changes in scale, color, rotation, different exemplars from the same category). As a performance benchmark, we quantitatively measure human invariant visual search behaviors and introduce four increasingly more complex tasks where we track eye movements while subjects search for a target. IVSN can selectively, invariantly and efficiently find target objects and its performance is consistent with human's performance both on average and at an image-by-image lev-

el. Finally, experimental results also demonstrate that IVSN, like humans, succeeds in efficiently finding the target in a challenging visual search problem for novel objects without any prior training with the target or search image features.

Eye movements reflect rich information about the complex cognitive states of the brain, including thought processes and goals [44, 45, 46, 47, 48, 49, 50, 51]. Additionally, with advanced eye-tracking technologies, it is now possible to monitor eye movements at high spatial and temporal resolution while controlling the task and visual environment. Therefore, eye movements provide a suitable arena to investigate how to infer a person's goals from their actions.

In Chapter 7, we apply our understanding about visual attention mechanisms and address the challenging problem of inferring what the subject is looking for in the context of a visual search task by decoding their error fixations. We define *"error" fixations* as the non-target fixations before the target was found. Given these error fixations, the goal is to decode what the target is. Several studies have shown that the error fixations during visual search are not random: those fixations are more likely to be on objects and locations that are similar to the target [52, 53, 54].

With the advancement of eye-tracking technology in wearable devices, computational models to infer the search target from human eye movements have several important application domains, such as health care, interactive user interfaces, and virtual reality (VR). For example, gaining information of the sought object of interest would be invaluable for VR processors to provide timely feedback to players. As another example, compared with neural decoding methods based on electrode recordings inside human brains, decoding intentions in physically-disabled patients from eye movements is less invasive, has lower cost and significantly fewer potential complications.

To the best of our knowledge, there have been few attempts to build computational models that use eye fixation information for inferring what the search target is on complex natural images. To tackle this challenging problem, we proposed a zero-shot deep network. The network applies knowledge from an object recognition task on a target inference problem *without any retraining*. We designed two sets of visual search experiments with object arrays and natural images, respectively, collected human eye movement data, and evaluated the model on these two datasets given the human error

fixations in the search tasks. The model could successfully decode what the target was without any prior training on the inference task.

Note: Overall, the thesis is a bundle of my PhD works where most of them have been published or the pre-print versions are accessible in my personal website `https://a0091624.wixsite.com/mengmi`. Instead of reading the thesis, I strongly advise readers to look at my publication list and refer to my personal website for relevant paper downloads. In the beginning of each chapter, I have also highlighted which paper the chapter is based on for your convenience.

Chapter 2

Literature

This chapter provides a brief account of a wide range of related works on visual attention (VA) in the fields of neuroscience, psychology and artificial intelligence. It is to examine germane research on visual attention and its connection with memory to bring insights into developing biologically-inspired computational visual attention models.

2.1 Neurobiological Inspirations

In this section, we study great amount of works in neuroscience related to cortical circuit of visual attention. Since the interplay of memory and visual attention plays inevitable roles in both bottom-up and top-down attention modulation, we also examine the relations between memory and visual attention in details in the latter half of this section.

2.1.1 Cortical Circuit of Visual Attention

This section serves as an overview of the cortical circuit for visual attention. We introduce important brain regions responsible for VA. Figure 2.1 shows the network of visual areas in the brain where visual attention modulation and execution or alike are addressed. The circuit is organized topologically. Major references [55, 56, 57] contribute to building this network. [58] presented the similar circuit for visual attention without integration of memory. [59] showed the cortical areas involved in bottom-up and top-down processing. Refer to [60] for a complete hierarchy of 32 visual cortical areas.

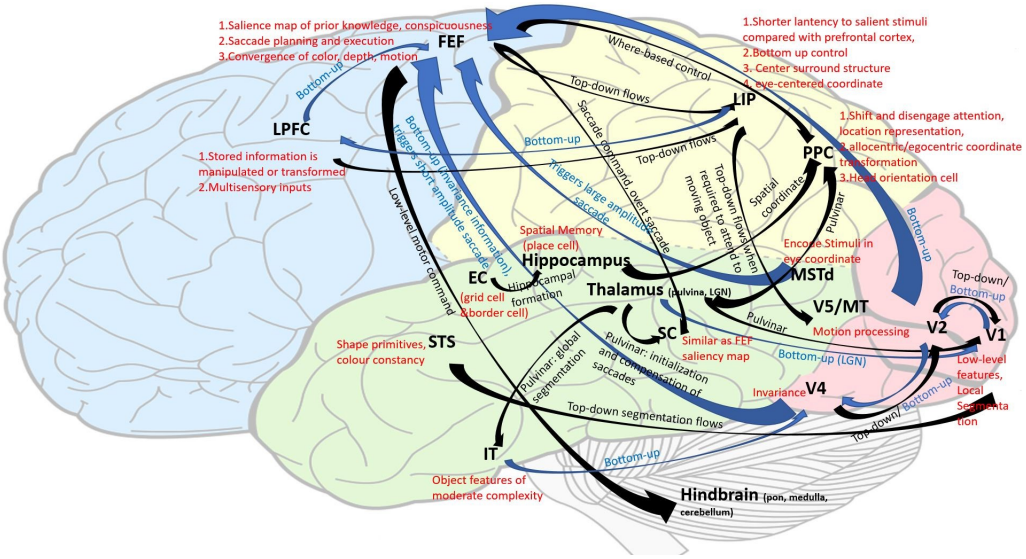


Figure 2.1: Overview of the cortical circuit for visual attention.

Brain has forebrain, midbrain and hindbrain. Forebrain includes thalamus, hippocampus, and cerebrum. Cerebrum has four lobes: parietal, occipital, frontal and temporal lobes. Hindbrain includes cerebellum, pon and medulla oblongata.

V1: primary visual area (occipital lobe). Spatial attention is identified in this visual area as the retinotopic map in this area is precise. It is also a bottleneck for visual information flow to the secondary visual brain areas [7, 61]. Previous works on contextual interactions has shown how their characteristics coincide with the extent and specificity of the long-range horizontal connections that are intrinsic to area V1 [62, 63, 64, 65]. For example, according to stimulus contrast and background foreground relationships, V1 can dynamically adjust the size of the receptive field after the interaction between excitatory and modulatory surround changes. The response of V1 is context dependent, can be changed by altered visual experience and training, and then later substantially modified by behavioral context and the state of attention. Evidence [66, 67] also shows that attentional modulation of the early cortical levels at V1 determine spatial vision thresholds e.g. contrast, orientation and spatial frequency. Over a wide range of stimulus contrasts, the contrast discrimination threshold initial decreases and later increases with increasing stimulus contrast. It also has dynamic nature of the sharp tuning of responses to orientation and spatial frequency and the relative constancy of orientation and spatial frequency tuning [68, 69]. Visual chunks are often locally segmented at V1.

Later on, the top-down attention modulates the inter and intra object binding problem and affects the global segmentation process in V1 [56] until the attended object is fully segmented and found. V1 computes spatially and temporally in parallel local low level features of objects within diverse retinotopic sub maps or modules [70, 71].

LPFC: lateral prefrontal cortex (frontal lobe). Dorsal LPFC activates when stored information is actively manipulated or transformed. In particular, the intraparietal sulcus regions are the source of top-down signals interacting with motion-sensitive regions with incoming sensory stimulation including multisensory control [72]. According to [73, 57], FEF and LPFC flows to LIP for top-down control of attention.

MT/V5: middle temporal area (temporal lobe). Many studies have shown that MT is responsible for motion processing [74, 75, 76]. MT responds selectively when subjects are required to attend to moving stimuli [77, 78, 79, 80]. LIP flows to MT [81]. V5 together with MT are shown to be motion-sensitive [82, 83].

Prefrontal Cortex: Attentional control may be seen as a process of using task context to guide bias competition by appropriate weight setting. Frontal system is important in visual weight setting [84]. Prefrontal neurons can combine both inhibitory and facilitatory influences in contra-lateral space [85]. Recent works explored the particular cognitive processes supported by the frontal lobe: plan formation [86], error management [86], working memory [87] and goal selection [88].

FEF: frontal eye field (frontal lobe). It is located within prefrontal cortex. FEF contributes to transforming visual signals into saccade commands [89]. FEF is connected with extrastriate visual areas in both the dorsal stream and the ventral stream [90], and the projections between extrastriate visual cortex and FEF are topographically organized [91, 92]. The anatomical evidence also reveals a large degree of convergence of afferent from multiple extrastriate visual areas in FEF, such as signals representing color, form, depth, and direction of motion of objects in the image. The activation of FEF visual neurons represents a salience map in which stimulus locations are selected on the basis of visual conspicuousness, prior knowledge, and internal random variablity [93]. FEF requires saccade planning and execution. It represents the process of selecting conspicuous targets.

V4: visual area V4 (occipital lobe). The central field representation of retinotopically

organized areas such as V4, TEO and MT, as well as areas that over-represent the central field, project to the ventrolateral portion of FEF [94]. It produces short amplitude saccade [95]. Evidence has been found that attention modulation happens at intermediate cortical levels such as V4 [96]. Invariance transformations are also found in V4 [97, 98, 99]. Top-down attention signal flows from V4 down to V2 and subsequently to V1 [100].

IT: inferior temporal gyrus (temporal lobe). IT processes object features of moderate complexity such as certain shape primitives [101]. The receptive field of IT neurons are large and cover several degrees of visual angles [97]. Invariance transformations occur in IT [97, 98, 99].

STS: superior temporal sulcus, the sulcus separating the superior temporal gyrus from the middle temporal gyrus (temporal lobe). Diverse type-level representations are computed within specialized modules such as shape primitives, colour constancy and so on. And then this information flows back to V1[97, 98, 99].

Po and MSTd: dorsal aspect of the medial superior temporal area (temporal lobe). The peripheral field areas such as Po and MSTd, project to the dorsal medial part of FEF. This part of FEF produces larger amplitude saccades [92].

SC: superior colliculus (midbrain). Several lines of evidence suggest a similar function of FEF happens in SC as well, i.e. it can form visual salience in the brain [102, 103]. Disruption to this region disables animals to select a salient stimuli [104]. Exogenous control has two pathways, one of them is generated in SC, for instance, a bright object among dim objects [105]. However, there is another pathway originating from V1 which generates brightness differences or abrupt onsets such as pop-out displays or vertical line among horizontal lines. In the saccade system, FEF sends a signal to the low-level motor structure like brain stem either directly or via SC. This results in an overt saccade [106].

LIP: lateral intraparietal cortex (parietal lobe). It seems to contain saliency maps sensitive to strong sensory inputs [107]. It has a center-surround structure [108, 93]. It has shorter latency to a salient stimuli than the frontal cortex [109, 110]. From the signal synchronization hypothesis, FEF induces high-frequency oscillations in LIP when a target is in the neuron's receptive field [111]. According to [73, 57], LIP flows to LPFC

and then to FEF for bottom-up control of attention.

PPC: posterior parietal cortex (parietal lobe). It belongs to the type-level "where" pathway. PPC contains representations of locations that can be considered to be stable across eye movements [112]. However, where based control can presumably not be handled exclusively by the higher level location modules of PPC but needs the frontal lobe where areas that are connected with PPC [113]. With lesion in PPC, results show patients have disturbance in shifting and disengaging attention [114, 115]. Posterior parietal cortex (PPC) is vital for performing transformations between these different coordinate systems. Here, we review evidence for multiple pathways in the human brain, from PPC to motor, premotor, and supplementary motor areas, as well as to structures in the medial temporal lobe. These connections are important for transformations between egocentric reference frames to facilitate sensory-guided action, or from egocentric to allocentric reference frames to facilitate spatial navigation.[116]

Pulvinar in thalamus: It is supporting attentional control for achieving global segmentation. Neurobiological experimental evidence that suggests a role of the pulvinar for visual attention has indeed been collected [117, 105, 118]. Moreover, the pulvinar has the required neuronal connections. Two of its parts, the lateral and inferior pulvinar, are connected to V1 [119] and higher levels such as IT and PPC [120, 119]. Research also finds it has connections to SC for initiation and compensation of saccades [121, 122] and regulation of visual attention [117, 123].

2.1.2 Connection of Visual Attention with Memory

For decades, neuroscientists have focused on the studies of hippocampus and associated MT in order to investigate into memory in primates by neurophysiological methods. To assess memory, viewing behaviors is an effective approach. The growing trend in studying biological systems in more natural settings [124] enables us to observe the interplay between memory and looking behaviors under less controlled conditions.

On one hand, our memory gets influenced by human vision because the latent scene representation in the memory is encoded via a discrete sequence of eye fixations. In particular, the fovea on the retina of human eyes processes only parts of the scene within two degrees of viewing angles in high fidelity with the rest blurred. By shifting eye

fixation locations across time, visual information can be encoded and integrated in the memory. Studies [125, 126] have found that fixation count has high correlations with the picture memorability. Eye fixations, as indicator of visual attention, determine what we are aware of in the image. Having more fixations on regions of the picture is highly associated with stronger memory [127, 128, 129].

On the other hand, memory guides looking behaviors. In physiological experiments, novelty preference in looking behaviors is often used as a means of assessing memory [130, 131, 132]. The duration and the spatial distribution of eye fixations may imply how well information is encoded in memory. For example, people tend to have shorting viewing timings on a repeated image compared to novel ones since the information of a repeated image has already been encoded in the memory [133, 134, 135, 136, 137]. It is also reported that the spatial distribution of eye fixations is more concentrated when observers have more confidence in identifying repeated images compared with those ones that subjects feel uncertain about whether the image has been visited before [125]. There have also been evidences in neural activity in hippocampus and MT supporting that memory guides and changes viewing behaviors. For example, theta band oscillatory activity in hippocampus is closely linked with saccadic eye movements while monkeys are performing memory tests [138]. More examples have shown that neural responses in hippocampus and entorhinal cortex often get attenuated or enhanced when observers are presented with repeated or novel images [139, 140, 141].

2.2 Saliency Prediction on Static Images

Feature-integration theory describes a process where combined low-level features, such as color, contrast and intensity attract human's attention. Most computational visual attention models originate from this idea, including the most early work introduced by Koch *et al*. [7] and Itti *et al*. [142]. Subsequent works have emerged which improve the performance on predicting saliency maps [143, 144, 145, 146, 147, 148]. Harel *et al*. [144] proposes a multi-scale structure based on low-level features. Later on, Gabor features learnt from independent component analysis are introduced in Bruce [149] which are useful for saliency prediction. Recently, randomly threshold feature maps in Boolean Map Saliency (BMS) are introduced in Zhang [145]. However, all these meth-

ods ignore the semantic information such as objects which turn out to be essential for predicting saliency maps. Even though subsequent works have incorporated a few object detectors in visual attention models, such as face and text detectors, these methods are still restricted to a limited set of object categories [147, 148]. With the successful demonstration of deep learning in object recognition, extracted semantic regions or objects from the scene by deep convolutional neural networks have significantly boosted the saliency prediction accuracy [4, 5]. However, the temporal information from a sequence of eye fixations has been discarded which turns out to be valuable in studying decision making in eye movements according to [150].

Although there are also early works studying the temporal dynamics of scanpath characteristics [151], they only focus on shape silhouettes and it is unclear whether these models could be generalized in natural images. Similar as [151], Renninger *et al*. [152] exploits the visual cues in scanpath prediction. Subsequently, Sun *et al*. introduced super-Gaussian component (SGC) based approach [153]. Liu *et al*. further improved scanpath prediction accuracy by utilizing the semantic information and transition between fixations in the model [154] but all these semantic features are hand-crafted and transitions between fixations are pre-defined. There have been recent works where computational models learn to predict scanpaths from training examples of human fixation sequences, such as [155] and [156]. In particular, Ming *et al*. train the model to learn a visual exploration policy and assign a set of different weights for related semantic cues at each stage of the visual scanpath [157]; however, memory is not incorporated in the decision making process during eye movements.

Different from all these previous works, we show for the first time that deep neural networks can make predictions beyond saliency maps, as these networks can also estimate the sequence of eye-fixations across time by integrating the information from all the past eye fixations to predict the next fixation location.

2.3 Gaze Prediction on Egocentric Videos

The previous section focuses on saliency prediction on static images and the motion information across video frames has been discarded. There are a few works exploiting the relations among gaze motions, head motions and foreground object motions on the

video frames.

Ba *et al*. [158] proposed to predict gaze locations by analyzing the relations between head orientation and gaze direction. Similarly, Yamada *et al*. [8] models motion correlations with the assistance of motion sensor measurements. Borji *et al*. [159] integrates motor actions and low-level features to predict fixation locations in a driving simulation scenario. Other than egocentric videos, all these works require additional information from external motion sensors. For those works solely relying on visual features extracted from egocentric videos, their models are only suitable for particular egocentric activities and may not generalize well. For example, the most recent model proposed by Yin *et al*. [9] has demonstrated high gaze prediction accuracy; however, hand detection and pose recogntiion provide primary egocentric cues which are not always present in all egocentric activities.

Computer vision researchers have also devoted efforts in studying gaze prediction on third-person videos, such as [160, 161, 162, 163]. To model temporal dynamics, most of these works propose different approaches, *e.g*. space-time whitening [161], salient candidate selection across time [162], or video compression [160]. The recent Long Short-Term Memory (LSTM) based work [163] learns the essential spatial-temporal features via end-to-end training. However, all these works have not shown whether the same models could be directly applied on gaze prediction problem in egocentric videos. Moreover, different from all these methods which require multiple frames, our model requires one single current frame.

To go beyond gaze prediction problem, we introduce a new and important gaze anticipation problem and propose a novel GAN-based model. To deal with the complex motion dynamics, we untangle foreground and background motions in egocentric videos with two-stream architecture. According to the experimental results, our model is capable of capturing useful egocentric visual cues and modeling the temporal dynamics for anticipated gaze locations after the training phase. Even without fine-tuning on normal videos, our model can still be directly applied in gaze prediction problem and surpass state-of-the-art algorithms.

2.4 Target-driven Search Models

We summarize related computational models of visual search and we also compare the invariant visual search task with other computer vision problems.

Human search behavior is often guided by the characteristics of the target [164, 165, 166, 20, 21, 167, 164, 22]. Cognitive Science work has mostly focused on low-level properties of the target that can guide search (e.g., [54]). Several computational models have been developed to describe modulation of responses during feature-based attention or visual search [168, 169, 40, 39, 41, 42]. A recent model proposed by Miconi *et al*. [19] incorporates ideas from Neuroscience in a proof-of-principle demonstration of visual search for exact matches of an object. Another recent visual search model utilizes both the search image and target classification labels to back-propagate and infer the maximum of hidden unit activations and localize objects [170]. The ability of these previous models to generalize and invariantly search for novel objects is limited.

Object detection and localization are "search" processes with pre-defined object classes of interests, typically focusing on performance accuracy irrespective of computational efficiency. To localize objects, most approaches require a large amount of supervised data, such as bounding boxes or object segmentations. Multiple recent approaches use deep neural networks for locating class-specific bounding boxes [28, 171, 172, 173]. Typically, sliding windows [171] or region proposals at uniformly sampled grids on images [28, 173, 172] are used. The purely feed-forward approach often involves proposing regions at the uniformly sampled grids, performing feed-forward classification for each proposed region and making decisions. These heuristic methods are computationally inefficient (in terms of the number of "fixations" or proposed regions and require extensive training). An analogous strategy is used in image retrieval tasks where a similarity score is computed between a query and each candidate image [174, 175, 30, 31]. In contrast with these methods, humans can invariantly and efficiently locate target objects in a few fixations instead of scanning the whole search image grid by grid, even with no prior training with the targets or search images. We propose a biologically inspired model which performs invariant and efficient visual search tasks with zero training examples and mimics human performance.

2.5 Decoding Targets from Fixations

Although information about a target is available in the fixation behavior during visual search, this does not imply that subjects are able to extract this information and use it to infer a search target [52, 53, 54]. Whether humans can infer the target information from other people's fixation behavior or not remains controversial. Some researchers have reported that it is possible to decode task information from eye movements [49, 176, 177, 178, 179, 180] while others have argued against otherwise [47, 181].

The focus in our study is on designing a computational model capable of inferring what the subject's target is. There are a few studies on decoding target information in the context of visual search [182, 183, 180], but current methods are limited in using elementary search statistics [180] and handcrafted features [182, 183]. Moreover, existing approaches have only been tested with pre-defined object classes with constrained object set sizes. These computational models do not generalize to infer any target from arbitrary classes. In contrast, our model is capable of inferring any target on complex natural images.

Chapter 3

Scanpath Network: Predicting Sequences of Human Fixations on Images

This chapter is based on the paper named "Deep Scanpath: Predicting Human Sequences of Eye-Fixations using Recurrent Neural Networks"[1].

We define the fixation stages as the order in the sequence of fixations. The fixation stage of a scanpath discards the duration of the fixation and the saccade, and only takes into account the location and the order of the fixations. We use fixation stages rather than time to describe the scanpath. The aim of this work is to analyse the temporal sequence of fixations rather than the duration of the fixation. We use $t \in \mathbb{N}$ to index the different fixation stages, and we define T as the total number of fixation stages in which the scanpath is divided.

We formulate the scanpath prediction problem as an iterative process learnable by our Deep Scanpath Neural Network (DSNN). At the fixation stage t, DSNN predicts the fixation (x_t, y_t) given the image I and the fixation location (x_{t-1}, y_{t-1}) that has been predicted at the fixation stage $t-1$.

Note that humans subjects may have different visual scanpaths while looking at one static image. In order to handle the inconsistency among human visual scanpaths, we

[1]Paper download link: https://docs.wixstatic.com/ugd/d2b381_0ffd2ca5c2ef47cfb4705fec968d3644.pdf

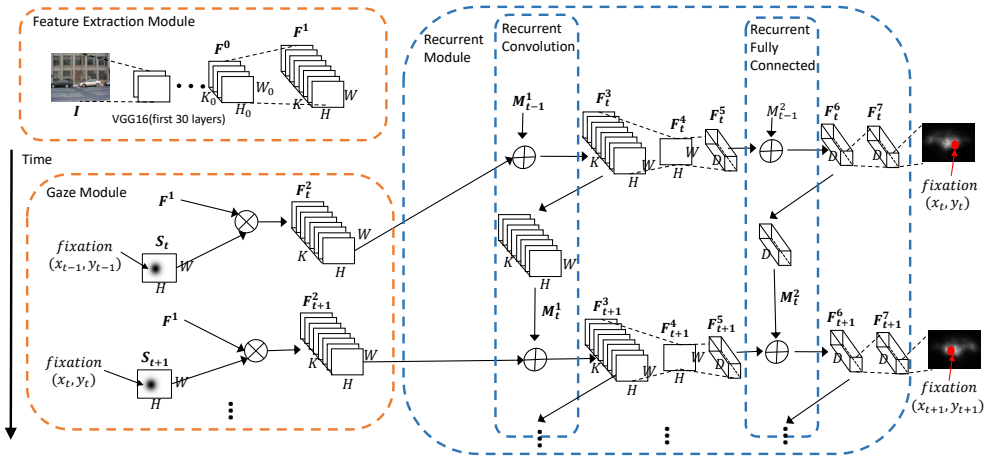


Figure 3.1: Architecture for Deep Scanpath Neural Net (DSNN)

align the fixations of the human subjects using the fixation stages. In other words, we align the fixation stage of the visual scanpath of two human subjects using the order of the fixation even though the fixations may happen at a different time.

Rather than directly predicting the fixation location, DSNN predicts a temporal saliency map that captures the scanpath variability among humans, *i.e*. the probabilistic map of the fixation locations across time. This temporal saliency map is used to predict a representative fixation location at each stage by using the spatial coordinate corresponding to the maximum of the temporal saliency map at stage t.

We define the standard 2D convolution operation as *conv*, the fully connected operation as *fc* and the linear rectifier function as $relu$ in deep learning. We use 0-padding in all convolutions in order to maintain the spatial resolution. Refer to [184] for the details of these operations in deep learning.

3.1 Recurrent Neural Network Model

DSNN is built on the recurrent neural network (RNN) as shown in Figure 7.3. DSNN comprises of two parts: *GazeModule*, and *RecurrentModule*. *GazeModule* is based on the deep Convolutional Neural Network (CNN) architectures for object recognition as it has been shown to be effective to predict saliency maps. The *GazeModule* also uses a mechanism to mimic inhibition of return that discourages DSNN to explore the already visited locations. The *RecurrentModule* is attached after *GazeModule* for adapting to

the changes of dynamics during scanpath prediction. It is composed by a series of convolutional and fully connected layers, and two recurrent layers that encode the memory. In the following, we introduce these two modules.

3.1.1 Gaze Module

As shown in previous works, *e.g.* [4], the semantic information in high layers of CNNs trained for object recognition can effectively be used to predict saliency maps. Thus, we use a pre-trained model on ImageNet for object recognition to generate multiple feature maps with semantic content. In particular, we use the first 30 layers of VGG16 [185] as an example. The output of the first 30 layers of VGG16 is denoted as F^0, and is composed by K_0 feature maps of size $H_0 \times W_0$ which correspond to different regions of the image. F^0 is rich in describing the semantic information across the image, which is essential for scanpath prediction.

F^1 is obtained by mixing the pool of feature maps F^0 using a *conv* operation. Moreover, in order to maintain a proper spatial resolution, we attach one up-sampling layer before this *conv* to scale up the size of each feature map in F^0. Thus, after one up-sampling layer and a *conv* layer, F^1 has K feature maps with each feature map of size $H \times W$. The spatial dimensions of this feature map, $H \times W$, are much smaller than the size of the original image. Yet, this is not a problem as one can still predict saliency even on a low resolution image [4].

Inhibition of return is used to discourage DSNN to explore the already visited locations [186]. We implement inhibition of return by multiplying each feature map in F^1 with a spatial map that encodes the previous fixation location. Let S_t be the spatial prior map for the inhibition of return at time t, which is of size $H \times W$. S_t is defined as function g dependent on the previous predicted fixation (x_{t-1}, y_{t-1}). Specifically, we choose g to be an inverted gaussian mask centered at (x_{t-1}, y_{t-1}) with variance σ and normalized to $[0, 1]$. The low intensity values on S_t near to (x_{t-1}, y_{t-1}) indicate the low probability for DSNN to explore in that location. Let F_t^2 be the feature maps after applying the inhibition of return on F^1. F_t^2 is defined as

$$F_t^2 = F^1 \otimes g(x_{t-1}, y_{t-1}) \tag{3.1}$$

where $\otimes$ represents element-wise product over each feature map of F^1, and hence, F_t^2 has the same dimensions as F^1. We can see by analyzing Eq. (3.1), that the $H \times W$ feature maps in F^1, which encode the semantic content among the image, are multiplied by values close to 0 in the location of the previous fixation. As a result, the feature maps of F_t^2 encode less salient content in the previous fixation location.

3.1.2 Recurrent Module

The *RecurrentModule* is attached after *GazeModule* for modeling the dynamics of the scanpath. *RecurrentModule* has two recurrent layers, *recurrent convolution layer (RC)* and *recurrent fully connect layer (RF)*. Let M_t^1 be the hidden state of *RC* at fixation stage t, and let M_t^2 be the hidden state in *RF* at fixation stage t. Between the two recurrent layers, there are other *conv* and *fc* layers that we introduce in the following.

First, F_t^2 is the same size as F^1. The recurrent layer *RC* then integrates the past feature maps stored in the memory, M_{t-1}^1, with the feature maps F_t^2 at fixation stage t. Let F_t^3 be the output of *RC*. M_{t-1}^1 and F_t^3 are of the same size as F_t^2. *RC* integrates the feature map F_t^2 with the memory M_{t-1}^1 using the element-wise addition $\oplus$. Thus, we define F_t^3 as

$$F_t^3 = \text{relu}(F_t^2 \oplus M_{t-1}^1). \tag{3.2}$$

Instead of memorizing all the output feature maps F_t^3, we use *conv* to learn how to selectively store these features in the memory M_t^1. A small weight in the convolution filter indicates that the corresponding feature map in F_t^3 is easier to forget. Thus, the memory M_t^1 at fixation stage t is updated as

$$M_t^1 = \text{conv}(F_t^3). \tag{3.3}$$

A *conv* operation is applied to F_t^3 to obtain the feature map F_t^4, which is of size $H \times W$ with the number of feature maps to be 1. We then use *fc* to transform the feature information F_t^4 into the latent representation of the following layer denoted as F_t^5. This latent representation in F_t^5 is a vector of length D where $D = H \cdot W$[2].

Similar to *RC*, *RF* uses an element-wise addition to integrate M_{t-1}^2 with F_t^5, and

[2] $(\cdot)$ is the scalar multiplication

obtains the output of *RF*, denoted as F_t^6. M_{t-1}^2 and F_t^6 have the same dimension as F_t^5, *i.e.* $D = H \cdot W$. Thus, F_t^6 is defined as

$$F_t^6 = \text{relu}(F_t^5 \oplus M_{t-1}^2). \tag{3.4}$$

Also, instead of storing F_t^6 directly in the memory, we use *fc* to tune its latent representation, *i.e.*

$$M_t^2 = \text{fc}(F_t^6). \tag{3.5}$$

In the next section, we show that this *fc* in *RF*, in fact learns a changing spatial bias across fixation stages.

Finally, the integrated latent representation F_t^6 is decoded into F_t^7 using *fc*. F_t^7 is again of dimension $D = H \cdot W$. Since this is equal to the spatial domain, F_t^7 can be used as the temporal saliency map before normalization. The spatial coordinate with the maximum probability from the temporal saliency map, *i.e.* F_t^7, is taken as the predicted fixation location (x_t, y_t) at fixation stage t. In the next iteration, *i.e.* fixation stage $t+1$, DSNN feeds back (x_t, y_t) as input together with image I to predict the subsequent fixation location (x_{t+1}, y_{t+1}). It is a sequential process. Hence, DSNN predicts the scanpath by generating a sequence of fixations across time.

3.1.3 Training

We train our network using end-to-end back-propagation in a fully supervised manner, *i.e.* all the parameters in our network are trained jointly.

We generate the ground truth data to learn the model by aligning all the fixations from all human subjects with the fixation stages. The aligned eye fixations from all human subjects produce a sparse fixation map, we put gaussian mask with variance σ over these maps to generate temporal fixation maps.

Let P_t be the temporal fixation map (the ground truth) and Q_t be the estimated temporal saliency map by DSNN at fixation stage t. The goal of the learning is to minimize a loss function between these two probability distributions across time. We use Kullback-Leibler divergence (KLD) loss function which has been shown to be one of the most effective loss functions for achieving the best saliency prediction [4]. In our

case, we average it across the fixation stages, *i.e*. the loss function is

$$\text{KLD}(P, Q) = \sum_t \sum_i P_t(i) \log \left[\frac{P_t(i)}{Q_t(i)} \right] \tag{3.6}$$

where i refers to the ith pixel on the maps P_t and Q_t.

We use stochastic gradient descent with learning rate fixed at 0.001 and batch size 1 to avoid being trapped in the local optimum. We stop the training at the turning point where we achieve the best scanpath prediction performance in the validation set, before there is over-fitting. Within each epoch, we randomize the sequence of inputs to the network. We train the network in a single NVIDIA Titan GPU with 12 GB memory.

3.1.4 Parameters of the Model

DSNN can predict sequences of eye-fixations for any number of fixation stages T. In our implementation and all the following experiments, we fix the number of stages to be $T = 6$. This choice produces an approximate correspondence between a fixation stage and the mean duration of an eye fixation (300ms), as the eye fixation recordings in the datasets used in the experiments are of duration between 1.5 to 2.5 seconds per image (*i.e*. $2\text{s}/6\text{stages} = 333\text{ms}/\text{stage}$). This choice is also made in accordance with previous works, *e.g*. [154, 157].

The parameters of the first 30 layers in *GazeModule* are preloaded from the first 30 layers in VGG16 [185]. These parameters are fine-tuned to scanpath prediction during learning.

The input image size is 300×400 with RGB channels. All the input images are normalized into $[0, 1]$. F^0 is denoted as the output from the 30th layer of VGG16, thus, F^0 has 512 feature maps with each feature map of size 19×25, *i.e*. we set $H_0 \times W_0 = 19 \times 25$ and the number of features $K_0 = 512$. After one upsampling layer to scale up the size of each feature map from F^0 and one *conv* layer to increase the pool of feature maps F^0, F^1 has 1024 feature maps with each feature map of size 38×50. That is, we set $H \times W = 38 \times 50$ and the number of features $K = 1024$ to maximize the rich representations of features extracted from VGG16 while maintaining a proper spatial resolution.

In the spatial prior maps for inhibition of return, and the temporal fixation maps,

we empirically fix σ to be 5 with respect to the size of the temporal saliency map 38×50, which is of the same size as the feature maps extracted from VGG16 after one upsampling layer and one *conv* layer. We set the width and the height of all the *conv* filters to be 1×1 in the last conv layer in *GazeModule* as well as all the conv layers in *RecurrentModule*. This is to assign a probability indicating how salient the response for each coordinate on the feature maps is.

3.2 Experiments - Scanpath Prediction

In this section, we evaluate DSNN for scanpath prediction.

3.2.1 Datasets

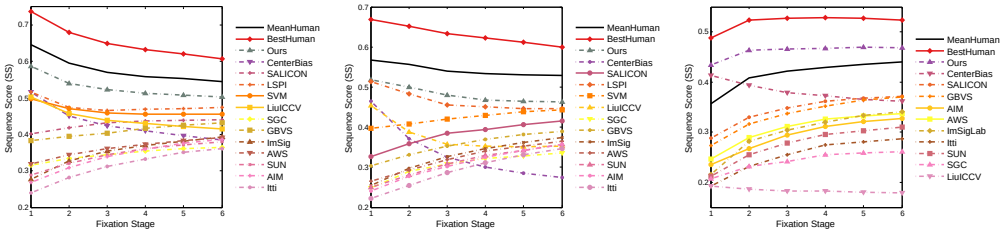
All the datasets are collected from [187], which include CROWD500 [188], MIT1003 [187], MIT2000 [187], FIGRIM2787 [189], KTH101 [190], LeMeur27 [191], VIU800 [192], OSIE700 [146], NUSEF760 [193] and Toronto120 [194]. In these datasets, the number of subjects per image vary from 7 to 104 depending on the datasets, and the subjects look at the images under the free-viewing conditions.

We use 3 different testing sets: 501 randomly chosen images from MIT1003, and 350 randomly chosen images from OSIE700 and all images from NUSEF760. For training, we use the training sets of all the aforementioned datasets, excluding all the testing images. We have about 9000 images in total for training.

We learn two different models of DSNN. In order to check for any dataset bias, the first model is tested in two testing sets, which are the ones from MIT1003 and NUSEF760. The second model is tested in OSIE700. For validation sets, the first model uses the test set of OSIE700, and the second model uses the 501 images randomly selected from MIT1003.

3.2.2 Evaluation Metric

Sequence score (SS) is proposed by Borji *et al*. [195], and it has been used to evaluate the accuracy of scanpath in the literature. We use the implementation by Jiang *et al*. [157]. SS computation is summarized: a mean-shift clustering for all human fix-



(a) MIT1003 (b) OSIE700 (c) NUSEF760

Figure 3.2: Quantitative Results on Scanpath Prediction on Static Images

ations is computed, and a unique character is assigned to each cluster center and corresponding fixations. The Needleman-Wunsch string match algorithm [196] is implemented to evaluate the similarity between human scanpath and the predicted scanpath.

3.2.3 Comparative Methods

For evaluation purposes, we provide a few comparative methods as below:

-*MeanHuman:* is the mean SS among pairs of sequences of human scanpath for all images.

-*BestHuman:* is obtained by taking the averaged SS of all the best subjects for all images. The best subject for each image is defined to have the maximum averaged SS across all fixation stages among all human pairs. This is the "the gold standard" for scanpath prediction.

-*Winner-take-all from Saliency Maps:* It generate scanpath from saliency maps with inhibition of return [142, 143]. During testing, we include the following saliency models: Graph-based Visual Saliency (GBVS) [144], Saliency Using Natural Statistics (SUN) [197], Adaptive Whitening Saliency (AWS) [198], Attention based on Information Maximization (AIM) [149], Itti's Model (Itti) [199], Image Signature Saliency (ImSig) [200], and SALICON [4].

-*Previous Scanpath Models:* We also compare our results with the previous methods for scanpath prediction: Least Squares Policy Iteration (LSPI) [157], Support Vector Machine (SVM) to combine the features at each fixation stage as in [157] [3], Hidden Markov Model from Liu (LiuICCV) [154] and Super Gaussian Component (SGC) [153]. These models have been reviewed in Chapter 2. We used the implementation of all these

[3] SS of LSPI, SVM are not provided on NUSEF760 dataset since we do not have annotated objects for this dataset.

models from [157].

-Previous Scanpath Models With Deep Features: For a fair comparision, we re-implement previous models and augment them with the deep learning features. We use LSPI (DeepLSPI) and SVM (DeepSVM) algorithms and extract the last convolution layer of SALICON as feature inputs to these algorithms. For DeepLSPI and DeepSVM, the number of superpixels is set to be 300. The rest of parameters remain the same as [157].

-Center Bias: To explore the effects of spatial biases, we create artificial fixation sequence with each fixation always in the center.

3.2.4 Results

We show the SS scores for the comparative analysis on MIT1003 in Figure 3.2(a), OSIE700 in Figure 3.2(b) and NUSEF760 in Figure 3.2(c). DSNN generalizes well across all three datasets. It outperforms state-of-the-art models, substantially reducing the gaps between machine and human. In particular, DSNN prediction surpasses *MeanHuman* on NUSEF760 (note that this is not surprising as NUSEF760 includes provocative and controversial images and the consistency of the eye fixations among subjects might be low).

To quantify how much DSNN has improved the state-of-the-art results, we report the mean difference between the SS score of DSNN and the second best algorithm, in percentage with respect to DSNN, *i.e*.

$$A(SS^r) = \frac{1}{T}\sum_{t}^{T}\frac{SS_t^{DSNN} - SS_t^r}{SS_t^{DSNN}}, \tag{3.7}$$

where SS^{DSNN} is the SS score of DSNN, and SS^r is the SS score for the second best algorithm (LSPI on MIT1003, LSPI on OSIE700 and Center Bias on NUSEF760 respectively). They are 10.5% on MIT1003, 3.6% on OSIE700, 21.4% on NUSEF760.

We observe that using deep learning features boosts the performance of all algorithms. This is because hand-crafted semantic features may not be sufficient to cover the wide range of salient objects. It is also shown that the algorithms which model the temporal information perform better than conventional saliency prediction methods with inhibition of return. For example, we observe that SALICON, which is based on

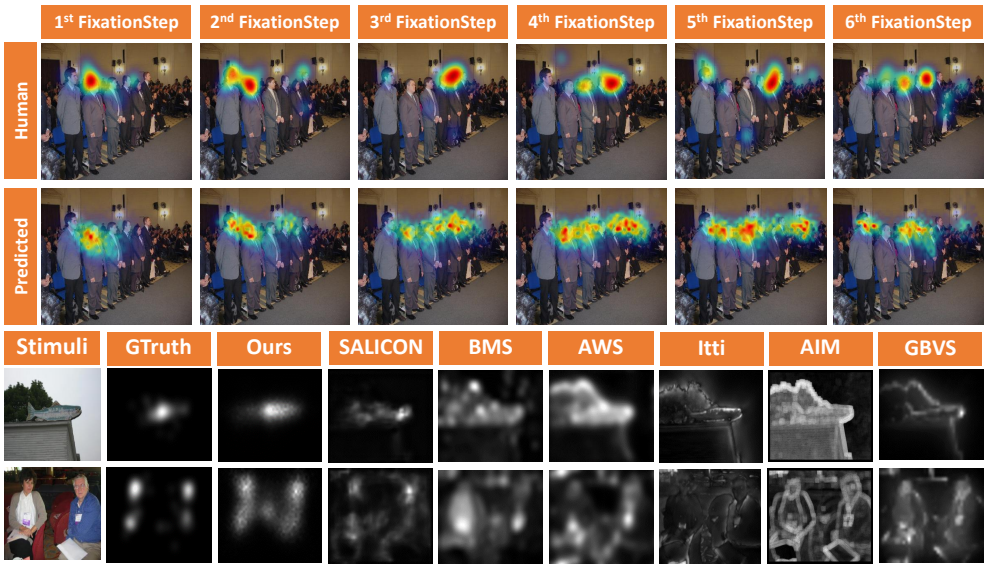


Figure 3.3: Example Scanpath and Example Saliency Maps predicted by our Deep Scanpath Neural Net (DSNN) and other Comparative Methods

deep neural networks, takes advantage of abundant semantic features, it cannot perform well for the first few fixations as it does not model the eye fixation temporal dynamics. Finally, one example of our predicted scanpath (first two rows) is shown in Figure 3.3.

3.3 Experiments - Saliency Prediction

We have shown that DSNN surpasses state-of-the-art algorithms on scanpath prediction. In order to show that DSNN in fact extends the predictive power of current state-of-the-art saliency predictors, we now show that DSNN can recover the accuracy of the most accurate saliency predictors in the literature. To convert the scanpath obtained with DSNN to a saliency map, we simply average the predicted temporal saliency maps.

3.3.1 Evaluation Metrics

We use several common evaluation metrics. The fixation map is based on all human subjects, and we evaluate across all images in the testing datasets.

-*Area Under the Curve (AUC) [195]:* It is the area under a curve of true positive versus false positive rates under various discrimination threshold values on saliency maps.

-*Shuffled-AUC (sAUC) [201]:* It compensates the center bias problem of AUC. Instead of uniformly sampling at random, it gets negative samples from other images.

-Normalized Scanpath Saliency (NSS) [201]: It computes the normalized saliency at fixation locations. It is sensitive to false positives and relative difference in saliency. Given a normalized saliency map P, and a binary fixation map Q with N fixations located at ith pixel (i is from 1 to N), NSS is defined as

$$NSS(P,Q) = \frac{1}{N}\sum_{i} P(i) \times Q(i). \tag{3.8}$$

-Correlation Coefficient (CC) [202]: treats saliency map P and a binary fixation map Q as random variables. It computes the linear relationship between them. This is useful in the context of scanpath analysis where relative saliency values at different image regions are concerned [203].

$$CC(P,Q) = \frac{cov(P,Q)}{cov(P,P) \times cov(Q,Q)} \tag{3.9}$$

where $cov(\cdot)$ is the covariance of two random variables

3.3.2 Comparative Methods

We compare the saliency map with the state-of-the-art saliency prediction algorithms as introduced in Section 3.2.3. We use the source codes from the authors, and follow the same parameter settings for saliency map prediction as [157]. Besides these saliency prediction algorithms, we also include the center bias.

3.3.3 Results

Table 3.1 shows that the accuracy of the saliency map by DSNN is comparable with the state-of-the-art models. Also, DSNN is best in AUC, NSS and CC, while it is almost as good as SALICON in sAUC over the three datasets. One possible reason for lower sAUC is that DSNN predicts temporal saliency maps which have strong center bias for the first few fixations, but sAUC gives more credit to off-center information [203]. Two example saliency maps predicted by our model and other comparative methods (last two rows) are shown in Figure 3.3.

	MIT1003 Dataset				OSIE700 Dataset				NUSEF700 Dataset			
Metrics	AUC	NSS	CC	sAUC	AUC	NSS	CC	sAUC	AUC	NSS	CC	sAUC
DSNN(ours)	**0.82**	**1.90**	**0.46**	0.69	**0.83**	**1.92**	**0.40**	0.74	**0.73**	**1.42**	**0.45**	0.58
SALICON [4]	0.80	1.51	0.36	**0.72**	0.83	1.90	**0.40**	**0.78**	0.71	1.0	0.31	**0.6**
CenterBias	0.78	1.05	0.27	0.55	0.72	0.80	0.17	0.52	0.73	0.88	0.29	0.49
GBVS [144]	0.77	1.21	0.29	0.64	0.80	1.34	0.28	0.70	0.70	0.83	0.27	0.55
AWS [198]	0.74	1.07	0.26	0.69	0.80	1.45	0.31	0.76	0.66	0.71	0.22	0.59
AIM [149]	0.72	0.86	0.21	0.65	0.79	1.08	0.23	0.72	0.65	0.56	0.18	0.56
SUN [197]	0.69	0.80	0.19	0.65	0.76	1.13	0.24	0.73	0.63	0.50	0.15	0.56
Itti [199]	0.63	0.55	0.13	0.59	0.67	0.74	0.15	0.63	0.58	0.30	0.09	0.53
ImSigLab [200]	0.55	0.55	0.13	0.54	0.62	0.7	0.16	0.60	0.56	0.36	0.11	0.53

Table 3.1: Quantitative Results in Saliency Prediction on Static Images by our Deep Scanpath Neural Net (DSNN)

3.4 Analysis of Temporal Dependencies across Fixations

In this section, we provide an analysis of DSNN via the ablation tests and the visualization of the hidden states in the recurrent module.

3.4.1 Ablation study

We report the performance in Table 5.5 after removing various components in DSNN to study their effects in SS on scanpath prediction. DSNN is relearnt for each case in the ablation study. These ablated models are: 1) *R(ReFC):* DSNN with *the recurrent fully connected layer (RF)* removed; 2) *R(ReConv):* DSNN with *the recurrent convolution layer (RC)* removed; 3) *R(ReConv):* DSNN with two recurrent layers replaced with the convolution layer and the fully connected layer respectively. Also, we include other variants based on the saliency map of SALICON with different spatial bias: 4) *SAL(1stCB):* the predicted visual scanpath of SALICON with the first fixation in the center; 5) *SAL(SP):* the spatial prior map pre-computed from human scanpaths at each fixation stage. The predicted saliency map from SALICON multiplied with the spatial prior map at each fixation stage, and then, inhibition of return is applied on these maps.

In Table 5.5, we show results for the test set from MIT1003. We also report the relative performance compared to DSNN (Row 1 in Table 5.5) as defined in Eq. (5.6). We find that removing any of the recurrent connections (*R(ReFC)*, *R(ReConv)*, and *R(AllRe)* in Table 5.5) reduces the accuracy of DSNN. This demonstrates that recurrent connections in DSNN are essential in learning the dynamics of the eye fixations.

The experiments with SALICON with different spatial bias (*SAL(1stCB)* and *SAL(SP)*

FixationStep	1	2	3	4	5	6	*A*
DSNN(ours)	**0.59**	**0.54**	**0.52**	0.51	0.51	0.50	–
R(ReFC)	0.52	0.52	0.52	**0.52**	**0.51**	**0.51**	3%
R(ReConv)	0.53	0.51	0.51	0.50	0.49	0.48	5%
R(AllRe)	0.55	0.51	0.48	0.46	0.44	0.42	10%
SAL(1stCB)	0.52	0.49	0.48	0.47	0.47	0.47	9%
SAL(SP)	0.52	0.46	0.45	0.45	0.45	0.45	12%

Table 3.2: Ablation Study of our Deep Scanpath Neural Network (DSNN)

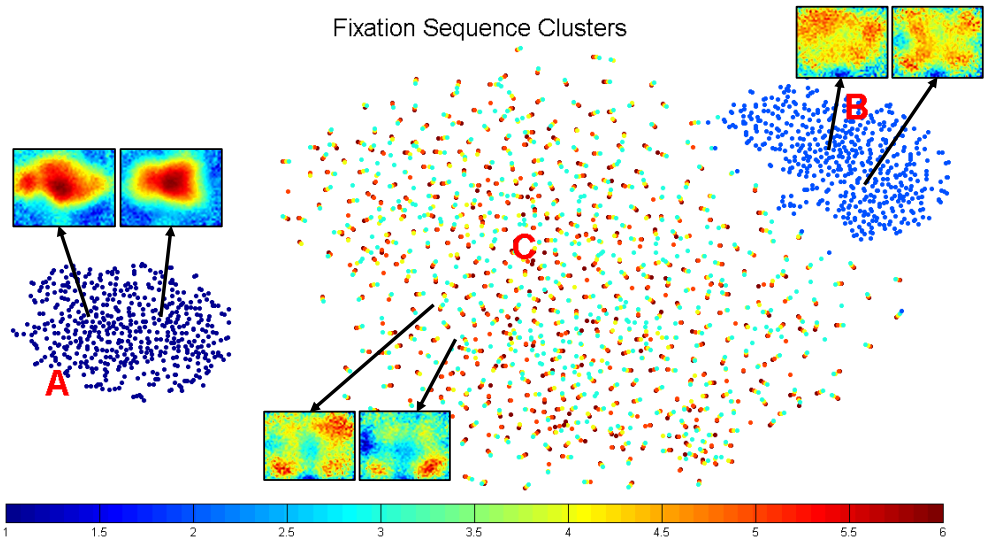


Figure 3.4: Visualization of the clustering of the latent representations in the hidden state of the recurrent fully connected layer across fixation stages.

in Table 5.5), obtain lower accuracy than DSNN. This shows that DSNN learns representations that are more useful than a spatial bias for deep learning features.

3.4.2 Visualization of Hidden States

In order to better understand the role of recurrent modules in DSNN, we provide a visualization method of the hidden state in *RF* by converting it to the spatial domain. T-Distributed Stochastic Neighbor Embedding (t-SNE) [204] is used for dimension reduction and clustering. We visualize the latent representations of the hidden state in *RF* over the first 6 fixation stages ($t = 1, ..., 6$) from 501 images in MIT1003, *i.e.* 3006 latent representations of the hidden state.

In Figure 3.4, we show the visualization result of the hidden states in *RF*. We use a different color to denote the different stages t, *i.e.* dots with the same color are from

different images at the same fixation stage. We observe that the hidden states at the first and second fixation stages form cluster A and B respectively, while cluster C contains the hidden states at the remaining 4 fixation stages. By analysing the pattern of these hidden states, we find that there exists a strong center bias for the first fixation. The latent representations in the hidden states shows higher activation to the surroundings as the fixation stages increase. At the 6th fixation stage, the spatial prior becomes more spread-out. This suggests that DSNN can emulate human visual scanpath behaviors by focusing attention on the salient objects nearest to the center at initial stages, and moves on to surrounding salient objects at later stages.

Chapter 4

Foveated Network: Predicting Human Gaze on Egocentric Videos

This chapter is based on the paper named "Foveated Neural Network: Gaze Prediction on Egocentric Videos"[1].

We first introduce an overview of our model, named as Foveated Neural Network (FNN), followed by a detailed analysis of each module in FNN. We provide training and implementation details in the end.

We formulate the gaze prediction problem on the current frame of egocentric videos as: given the previous frame and the current frame, FNN outputs the saliency map for the current frame. Hence, the spatial coordinate with the maximum probability on the saliency map is the predicted gaze location.

We define an egocentric frame I of low resolution and high resolution using superscript l and h respectively. The subscript denotes time t. A saliency map is defined as a probability distribution of gaze locations; thus, the spatial coordinate of the maximum probability in the saliency map is the predicted gaze location f^r. Similarly, we use the estimated saliency map obtained from the low-resolution frame to propose ROI centered at f^c. We use superscript r as the refined gaze location (the output of FNN) and superscript c as the center of the proposed region of interest (ROI).

[1]Paper download link: https://docs.wixstatic.com/ugd/d2b381_4609966b34ba417e825db191d3059838.pdf

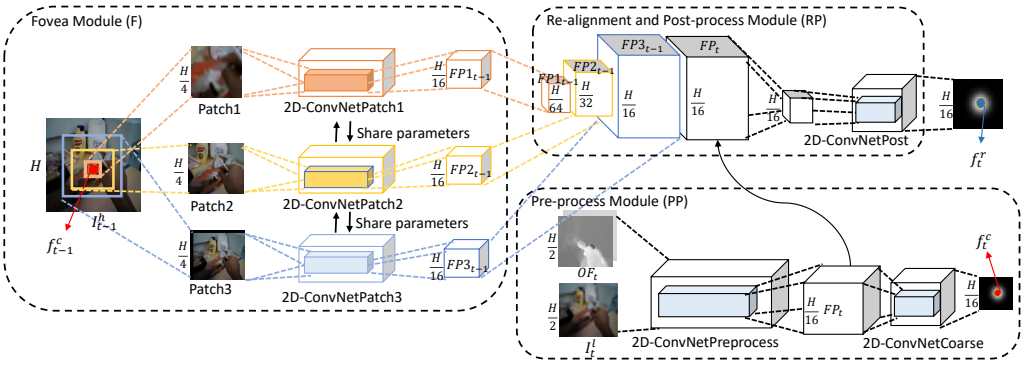


Figure 4.1: Architecture of our model for Gaze Prediction on Current Frame.

4.1 Foveated Neural Network

The overview of FNN is presented in Figure 7.3. FNN divides into three modules: *Pre-process Module (PP)*, *Fovea Module (F)*, and *Re-alignment and Post-process Module (RP)*.

In *PP*, based on the current frame I_t^l of low resolution and the optical flow OF_t in horizontal and vertical axis, FNN extracts the feature maps FP_t useful for gaze prediction and estimates the region of interest (ROI) on the current frame. The center of ROI f_t^c (red dot) will be used in the next iteration (time $t+1$). In *F*, given the high-resolution frame I_{t-1}^h and ROI on the previous frame centered at f_{t-1}^c, *F* simulates the human fovea and outputs the feature maps extracted from the retina-like image patches centered over ROI. They are of different resolution and cover different sizes of the receptive field. The patch covering the large receptive field is of low resolution while the one covering the small receptive field is of high resolution. In *RP*, the extracted feature maps from the patches $FP1_{t-1}$, $FP2_{t-1}$, and $FP3_{t-1}$ are re-aligned based on the center of ROI and concatenated with the feature maps FP_t extracted from the current frame. The combined feature maps are post-processed and output the refined saliency map and hence, the predicted gaze location f_t^r on the current frame at time t (blue dot).

4.1.1 Fovea Module

Given an egocentric high-resolution frame I_{t-1}^h and the center of ROI f_{t-1}^c at time $t-1$, ROI is attended in a foveated manner. In order to simulate the attentional processing in the retina, we use the same approach as [150]. Instead of assessing the frame in high resolution across all pixels, F extracts the retina-like representation focused on f_{t-1}^c,

i.e. different image patches of limited bandwidths centered at f^c_{t-1}. In our case, we use three bandwidths: $H \times H$, $\frac{H}{2} \times \frac{H}{2}$ and $\frac{H}{4} \times \frac{H}{4}$; however, not limited to three, *F* can be generalized to more than three depending on the applications. When the receptive field centered at f^c_{t-1} exceeds the frame boundary, we use zero padding to fill in the empty areas. These multiple resolution patches are then scaled to the same size $\frac{H}{4} \times \frac{H}{4}$. This is to simulate the fovea where the patch covering small receptive field (Patch1) is of high resolution whereas the patch covering large receptive field (Patch3) is downsampled to be of low resolution. Thus, it enables *F* to allocate the small amount of processing power (the same number of parameters in 2D-ConvNetPatch) on the large area of the frame in low resolution (Patch3) and vice versa.

As shown in [4], convolution layers of high levels in 2D convolution neural network (2D-ConvNet) trained for object recognition are effective in predicting saliency. We use the pre-trained 2D-ConvNet on ImageNet for feature extraction. The feature maps from these multiple resolution patches are extracted using branches of 2D-ConvNetPatch. The branches have the same architecture and share the same network parameters. The outputs of *F* are feature maps denoted as $FP1_{t-1}$, $FP2_{t-1}$ and $FP3_{t-1}$ respectively. Each of their feature maps are of size $\frac{H}{16} \times \frac{H}{16}$.

4.1.2 Pre-process Module

Before assessing to ROI of the current frame in high resolution, I^l_t (size $\frac{H}{2} \times \frac{H}{2}$) is perceived in low resolution at time t. *PP* uses 2D-ConvNetPreprocess for encoding features of I^l_t and 2D-ConvNetCoarse for proposing the ROI. As egocentric videos involve head motions, we compute the dense optical flow OF_t between I^l_t and I^l_{t-1} from [205] and use it to implicitly represent motions between adjacent frames. 2D-ConvNetPreprocess takes five channels as inputs: RGB channels from I^l_t and OF_t in horizontal and vertical axis. We denote the output from 2D-ConvNetPreprocess as feature maps FP_t with each feature map of size $\frac{H}{16} \times \frac{H}{16}$. FP_t and $FP1_{t-1}$, $FP2_{t-1}$, $FP3_{t-1}$ from *F* are of the same size and they will be used for predicting gaze location on the current frame. Based on FP_t extracted from I^l_t, 2D-ConvNetCoarse proposes one ROI where the model may be interested in focusing attention on. The ROI is represented using the center of ROI denoted as f^c_t. f^c_t is obtained by taking the spatial coordinate of the maximum

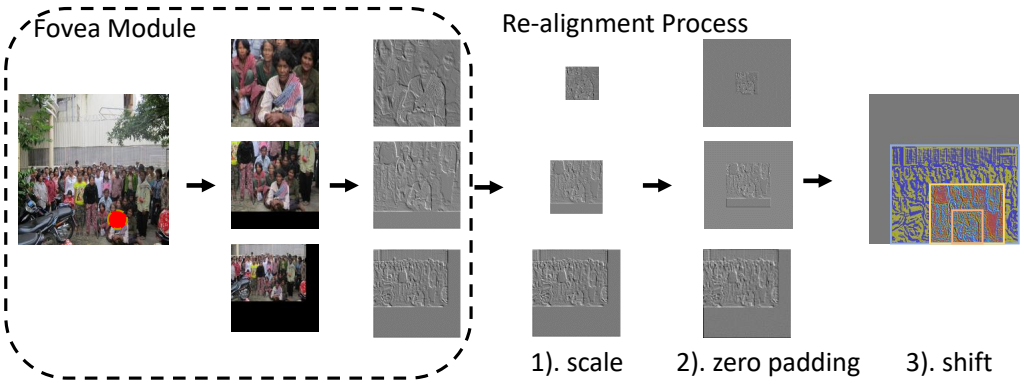


Figure 4.2: Illustration of the realignment process in *Re-alignment and Post-process Module*.

on the saliency map estimated from I_t^l in low resolution. It will be used in F in the next iteration (time $t+1$) where FNN predicts the next gaze location on frame I_{t+1}^l.

4.1.3 Re-alignment and Post-process Module

After obtaining $FP1_{t-1}$, $FP2_{t-1}$ and $FP3_{t-1}$, *RP* realigns these feature maps based on f_{t-1}^c. The realignment process includes the following steps as shown in Figure 4.2: 1). scale $FP1_{t-1}$, $FP2_{t-1}$ and $FP3_{t-1}$ to $\frac{H}{64} \times \frac{H}{64}$, $\frac{H}{32} \times \frac{H}{32}$ and $\frac{H}{16} \times \frac{H}{16}$ respectively; 2). add in zero paddings to each of the four sides of each feature map by $\frac{3H}{128}$ in $FP1_{t-1}$ and $\frac{H}{64}$ in $FP2_{t-1}$; 3). shift the concatenated feature maps back to f_{t-1}^c with respect to I_{t-1}^h. The realignment process is used for consolidating all the feature maps across multiple resolution patches to the same spatial location with respect to I_{t-1}^h.

In 2D-ConvNetPost, we use one 2D convolution layer to fuse the consolidated information on the previous frame together with FP_t from the current frame. The fused information is post-processed by another two fully connected layers before generating the final predicted saliency map of size $\frac{H}{16} \times \frac{H}{16}$. The coordinate with the maximum probability in the saliency map is the predicted gaze location f_t^r on I_t^l.

4.1.4 Training and Implementation Details

We train FNN in stochastic gradient descent with learning rate 0.01 and batch size 1. The fixation map (the ground truth) is defined as the binary map with human gaze locations. As a common practice, we put an isotropic gaussian mask over the binary map

and normalize it to be $[0, 1]$. Same as [4], we minimize Kullback-Leibler divergence (KLD) loss between the predicted saliency map and the fixation map. All the weights from 2D-ConvNet in FNN are pre-loaded using VGG-16 trained on ImageNet [185]. The parameter H is set to be 1200. All the numbers of feature channels for $FP1_{t-1}$, $FP2_{t-1}$, $FP3_{t-1}$ and FP_t are 512. The input frames to FNN are normalized to $[0, 1]$ with mean and standard deviation. We implement the proposed algorithm in Torch.

4.2 Experiments - Gaze Prediction

We compare FNN with the state-of-the-art using standard evaluation metrics on one publicly available dataset. In the following subsections, we introduce the evaluation metrics and comparative methods. In the end of the section, we present the results and the detailed analysis.

4.2.1 Datasets

We evaluate FNN using the publicly available egocentric dataset, GTEA [206]. It contains 17 video sequences in total with each video lasting for 4 minutes on average. 14 human subjects are asked to prepare for meals in a kitchen at their own wishes while wearing the eye-tracking devices. For fair comparison, we choose videos 1, 4, 6-22 as training and validation sets while the rest are used for testing same as [9].

4.2.2 Evaluation Metrics

We used two standard evaluation metrics to measure the performance of gaze prediction: Area Under the Curve (AUC) [195] and Average Angular Error (AAE) [9]. AUC is commonly used in the saliency prediction literature. It measures the consistency between a predicted saliency map and a fixation map of human gazes.

AAE is used in the gaze tracking literature and measures the error between the predicted and the human gaze locations in an angular distance. The smaller, the better.

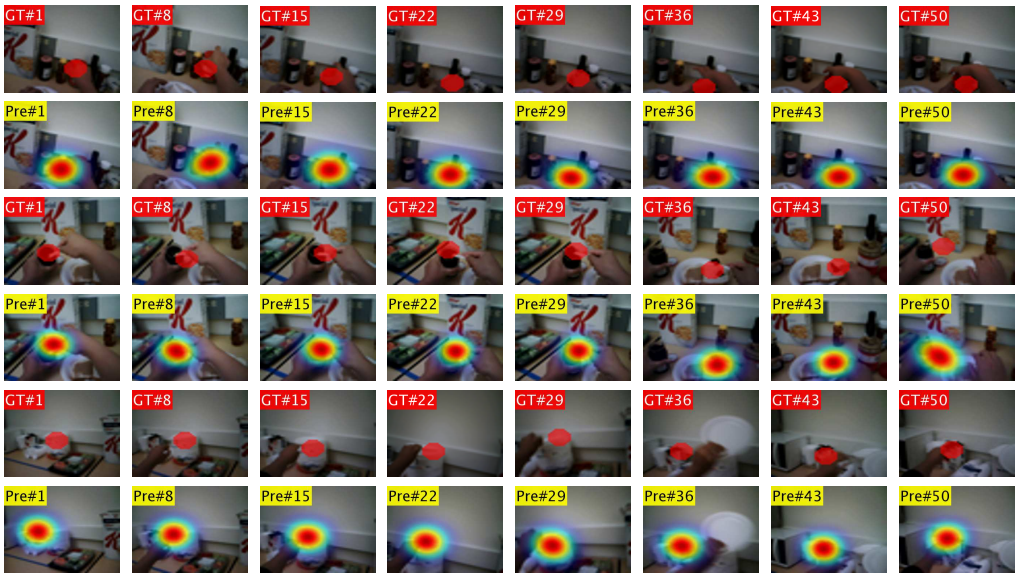


Figure 4.3: Exemplar results of gaze prediction on GTEA Dataset.

4.2.3 Comparative Methods

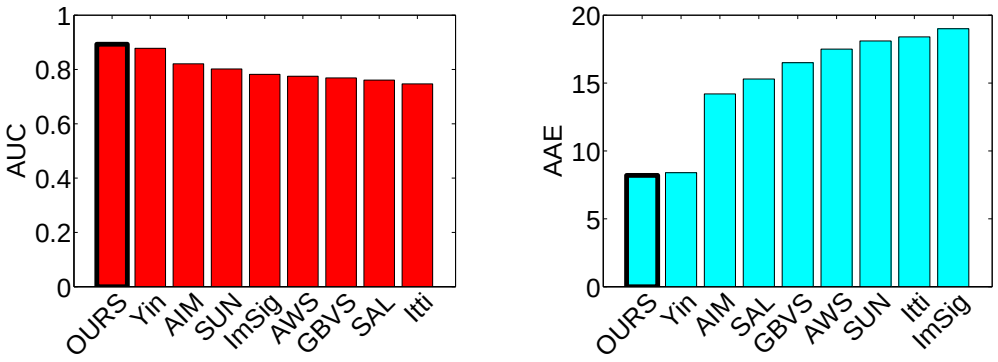
We compare our method with the state-of-the-art saliency prediction algorithms: Graph-based Visual Saliency (GBVS) [144], Saliency Using Natural Statistics (SUN) [197], Adaptive Whitening Saliency (AWS) [198], Attention based on Information Maximization (AIM) [149], Itti's Model (Itti) [199], Image Signature Saliency (ImSigLab) [200] and SALICON [4]. In particular, SALICON is a 2D-ConvNet with the current frame as the only input. We fine-tune SALICON on the training set and evaluate its predicted saliency maps in the test set.

In addition, we include [9] as it directly addresses the gaze prediction problem on egocentric videos by using Hidden-Markov model for the temporal dynamics.

4.2.4 Results

The results in AUC and AAE are presented in Figure 4.4. FNN outperforms the state-of-the-art algorithms on gaze prediction on current frames in egocentric videos in both AAE and AUC.

Compared with saliency prediction algorithms, FNN yields a significant boost in gaze prediction performance. Though SALICON learns the semantic features useful for gaze prediction, it fails to take temporal information into account. See the ablation



(a) AUC Evaluation

(b) AAE Evaluation

Figure 4.4: Results on GTEA Dataset using Area Under the Curve (AUC) in (a) and using Average Angular Error (AAE) in (b).

	AAE	AUC
SALICON (SAL)	16.5	0.76
SAL + 2 Fully Connected Layers (FC)	10.6	0.80
SAL + FC + OpticalFlow (OF)	8.33	0.88
SAL+ FC + OF + FoveaOnPreviousFrame	**8.15**	**0.89**

Table 4.1: Evaluation of Ablated Models and our Fovea Neural Network model on Gaze Prediction.

study in Table 5.7 (Row 3) for more details.

Though Li's work [9] uses the hidden markov model for temporal dynamics, FNN performs better with an improvement of 2.4% ($(8.33-8.18)/8.33 = 2.4\%$) in AAE due to the enriched pool of semantic feature representations in the network and the fovea module on the previous frame.

Some qualitative results are shown in Figure 4.3. Three exemplar egocentric video segments are presented with one out of every seven frames. Row 1, 3, 5 show the human gaze locations denoted by red dots (ground truth (GT)). Row 2, 4, 6 show the corresponding predicted saliency maps (Predicted (Pre)).

To further explore the effect of individual components introduced in FNN, we conduct an ablation study and report the results in Table 5.7. We build up FNN based on SALICON and we add in one component at a time. SALICON is a feedforward 2D-ConvNet with the last few fully connected layers removed. We added in 2 fully connected layers in the end which boosts up the performance to a significant extent in terms of AAE (Row 2). Compared with SALICON containing only convolution and pooling operations within a local receptive field, we hypothesize that the added 2 ful-

ly connected layers fuse all the information across space and increase the capacity of saliency representations.

To study the effect of the foreground and background motions, we add in the dense optical flow between the current frame and the previous frame as inputs to the network (Row 3). The first convolution layer has two additional input channels. The results improve by 2 in AAE and 0.08 in AUC. It suggests that the motion estimation between adjacent frames is an important egocentric cue for gaze prediction.

We present the result of FNN (Row 4). Compared with the one in Row 3, we add in the fovea module and fuse its feature maps with the one-stream network. Result shows an improvement of 0.18 in AAE and 0.01 in AUC. It explains that the integration of the foveated information on the previous frame is useful for predicting gaze on the current frame.

According to [9], there exists a strong center bias for gaze distributions on current frames in egocentric videos since the large gaze shift often gets compensated by the head motions. Hence, we use sAUC to evaluate FNN and compare it with the center bias. We create the artificial center as the predicted gaze location and we put an isotopical gaussian mask over the center for sAUC evaluation. We report sAUC results in GTEA: FNN (0.65) and center bias (0.5). It confirms that FNN predicts gaze locations more than center bias.

Chapter 5

Future Gaze Network: Anticipating Where People Will Look Next

This chapter is based on the paper named "Anticipating Where People Will Look Using Adversarial Networks"[1].

Our work presents the new and important problem of *gaze anticipation*: the prediction of gaze in future frames of egocentric videos within a few seconds. Figure 5.1 illustrates the gaze anticipation problem: given the current frame, the task is to predict the future gaze locations. Our proposed method solves this problem through synthesizing future frames (transparent ones) and predicting corresponding future gaze locations (red circles).

Gaze, as a perceptual variable, cues attention. Attention can be categorized into two distinct functions: the bottom-up attentional guidance driven by external stimuli due to their inherent features relative to the backgrounds, such as the visual contrast; and the top down attention mechanism according to the current goals and purposeful plans, such as the navigation task towards the driver's desired destination location. Inspired by these attention mechanisms, we tackle gaze anticipation problem in two streams. Given the current frame, our proposed model, Deep Future Gaze (DFG), generates future frames using generative adversarial network (GAN) through a competition between a generator and a discriminator, and then predicts the gaze locations on these frames as bottom-up approach (**DFG-G**). Meanwhile, DFG anticipates the gaze prior maps as task influences

[1]Paper download link: `https://docs.wixstatic.com/ugd/d2b381_86633109b089467e87abbf4fafaa14f3.pdf`

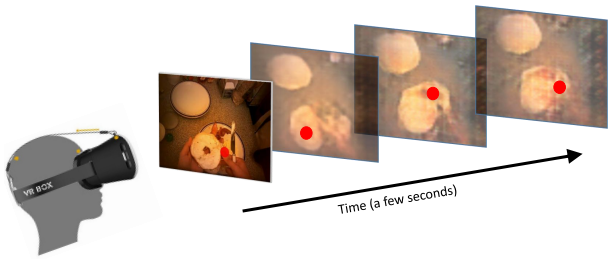


Figure 5.1: Problem illustration: gaze anticipation on future frames within a few seconds on egocentric videos.

(**DFG-P**) that mediates the bottom-up temporal saliency maps from the generator in **DFG-G**. Based on the latent representation extracted from the input frame before the generator, we use another 3D-CNN to predict spatial priors for gaze locations. This is the direct approach where **DFG-P** makes reasonings about the episodic steps in the task according to the semantic information extracted from the current frame without the intermediate future frame generation step. These goal-driven spatial priors bias the bottom-up saliency prediction leading to higher anticipation accuracy.

Evaluations of DFG on public egocentric datasets show that DFG boosts the performance of gaze anticipation to a considerable extent surpassing all the competitive baselines. In addition to egocentric videos in the cooking tasks, DFG demonstrates its capacity of generalizing to the object search task on Object Search Task Dataset (OST) [207]. Although DFG is not specifically trained for conventional gaze prediction problem on current frames, our GAN-based framework also significantly advances the state-of-the-arts for this problem. Moreover, we extend beyond egocentric videos and introduce the novel gaze anticipation problem on third person videos where the background is often static. In this case, DFG also achieves the best performance among all the baselines. Our rigorous analysis in the experiment section validates that our architecture can be generalized to diverse foreground and background motions. We add experimental investigations about our architecture design by exploring the potential factors influencing gaze anticipation performance and comparing the ablation results on both egocentric and third person videos. At last, we integrated our anticipated gaze locations with the existing activity recognition network. The reported results verify that anticipated gaze helps egocentric activity recognition.

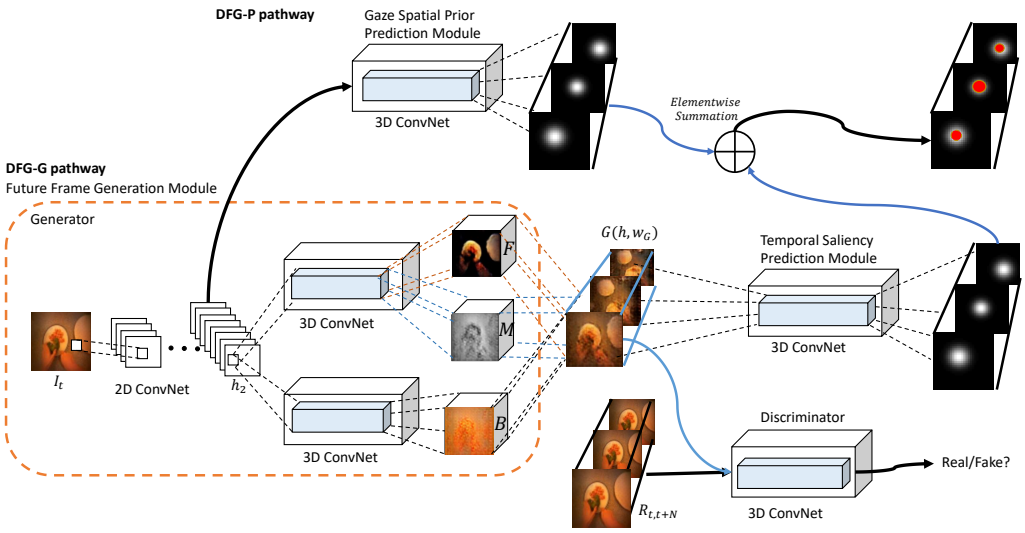


Figure 5.2: Architecture of our proposed Deep Future Gaze (DFG) model.

5.1 Generative Adversarial Network Model

In this section, we first introduce an overview of our proposed model, Deep Future Gaze (DFG), and then give the detailed analysis of its architecture as shown in Figure 7.3. We provide the training and implementation details in the end.

Given the current frame as the input, we aim to output a sequence of anticipated gaze locations in the next few seconds. To address this challenging problem, we propose an integrated framework consisting of two pathways: task-specific pathway **DFG-P** and bottom-up pathway **DFG-G**. In **DFG-G**, it consists of two modules: generative adversarial networks (GAN)-based **Future Frame Generation** and **Temporal Saliency Prediction**. In **Future Frame Generation**, it has two networks: **Generator** and **Discriminator**.

Generator generates future frames. In **Generator** in **Future Frame Generation Module** in **DFG-G**, latent representation of the current frame I_t is extracted by 2D ConvNet. To explicitly untangle foreground and background, it then branches into two streams: one for learning the representation for the foreground and the mask; one for learning the representation of the background. These 3 streams are combined to generate future frames (blue boundaries). As a competitor to **Generator**, **Discriminator** uses a 3D ConvNet to distinguish the generated frames from real frames $R_{t,t+N}$ (black boundaries) by classifying its inputs to real or fake. **Temporal Saliency Prediction**

predicts their corresponding temporal saliency maps, *i.e.*, spatial probabilistic maps of gaze locations across time.

DFG-G is regarded as the bottom-up pathway where the attention is driven by external stimuli (the generated future frames). Complementary to **DFG-G**, we add in **DFG-P** to estimate the priors of gaze locations without the intermediate future frame generation step. It makes inference about the gaze distribution in the task at hand based on the latent representation of the input frame I_t. In the end, the task-specific attention mechanism from **DFG-P** mediates the bottom-up attention in **DFG-G**. The temporal saliency maps predicted from **DFG-G** get biased by the gaze spatial priors via element-wise summation. The spatial coordinates with the maximum probability (red dots) are output as the anticipated gaze locations.

5.1.1 The Generator Network

In **Future Frame Generation**, the goal of **Generator** is to produce a sequence of N subsequent frames $I_{t+1,t+N}$ from a latent representation $h(I_t)$ of the current frame I_t. Hence, $I_{t+1,t+N}$ can be used for predicting N temporal saliency maps $S_{t+1,t+N}$ in **Temporal Saliency Prediction**. Here the latent representation $h(I_t)$ is learned from a 2D-CNN. In order to identify the foreground motions (hands and objects) out of the complex background motion due to the head movements, we propose a two-stream generator architecture. To avoid the error in the frame generation accumulating from one frame to another, **Generator** is designed to generate a sequence of N future frames at once instead of a system where the generated frame I_{t+1} is fed back as the input to generate the subsequent frame I_{t+2}. The number of predicted frames N is application dependent. We select 32 frames or about 2.5 seconds as we believe such duration is adequate for practical applications. The complete analysis regarding the performance of our model versus number of output frames is presented in Section 5.10.

We use 3D-CNN in two streams for learning motion representations. Meanwhile, fractionally strided convolution layers (upsampling layers) are added after the convolution to preserve proper spatial and temporal resolution for the output frame sequence.

The equation for generating the sequence of N predicted frames $I_{t+1,t+N}$ is

$$\begin{aligned} I_{t+1,t+N} =& F(h(I_t)) \odot M(h(I_t)) \\ &+ (1 - M(h(I_t))) \odot B(h(I_t)), \end{aligned} \tag{5.1}$$

where $\odot$ is the elementwise-multiplication operation, $F(\cdot)$ represents the foreground generation model and $B(\cdot)$ represents the background generation model. $M(\cdot)$ is a spatial-temporal mask untangling foreground and background motion where its pixel value ranges from $[0, 1]$. In particular, 1 indicates foreground and 0 indicates background. Both $F(\cdot)$ and $B(\cdot)$ generate a sequence of N predicted RGB-colored frames, each frame with dimension $3 \times W \times H$ where W and H are the width and the height of the predicted frame respectively. Foregrounds and backgrounds of predicted frames get merged by masks $M(\cdot)$ of dimension $N \times 1 \times W \times H$ replicated across 3 color channels to produce $I_{t+1,t+N}$. The foreground, background and mask models are parameterized by 3D-CNN. The foreground model and the mask model share the same weights until the last layer which has two branches, one for foreground generation for N frames with 3 color channels and one for the mask generation for N frames with single channels. The background generation model employs another separate 3D-CNN.

We note that, in egocentric videos, there often exists a clear distinction between foreground and background motions. While foreground objects tend to move together more coherently among themselves, they tend to distinguish from background objects due to motion relativity. For example, when the subject is transferring the food in hands from one place to another, foreground objects, such as arms and manipulated objects, tend to be always in the center of the egocentric frames while the background objects are moving in the opposite direction of head movements in the egocentric frames. The coherence within foreground and background motions themselves and the clear boundary between these motions make DFG learn to distangle the foreground objects from the background automatically during frame generation even though there is no specific training loss to explicitly supervise the network to distinguish these two.

As the rich information including the learnt egocentric motion dynamics on the generated future frames is useful for visual attention in egocentric videos, we adopt these features for gaze anticipation. Thus, **Generator** is followed by **Temporal Saliency**

Prediction to generate temporal saliency maps of dimension $N \times 1 \times W \times H$.

5.1.2 The Discriminator Network

Generating N frames implies the need of a large number of pixels. This is an extremely difficult task when only a single frame is given. To enhance the quality of generated frames, DFG employs **Discriminator** as a competitor to **Generator**, by providing the additional feedbacks to **Generator** [208].

Discriminator aims to distinguish the synthetic examples from the real ones. There are two criteria for the synthetic frames to be "real": first, the semantics from the scene are coherent across space (e.g. no table surface inside the refrigerator); second, the motions from both the foreground and the background are consistent across time (e.g. hand movements have to be smooth). Thus, **Discriminator** follows the same architecture as the foreground generation model other than replacing all the upsampling layers with the convolution layers and this architecture has also been shown to be effective in [208]. The output is a binary label indicating whether the input frame is fake or real.

5.1.3 DFG Gaze Spatial Prior Pathway (DFG-P)

As a complementary of **DFG-G** pathway, **DFG-P** estimates the gaze spatial priors based on the latent representation $h(I_t)$ of the current frame I_t in **Generator**. The semantic information in $h(I_t)$ underlying the task information contributes to the inference about the distribution of gaze locations in the next few seconds. To ensure the gaze movements to be coherent across spatial and temporal domains, we use a 3D-CNN in **DFG-P** to estimate the prior maps for gaze locations of dimension $N \times 1 \times W \times H$. At the training stage, the 3D-CNN encodes the spatial distributions of gaze locations and their motion trajectories corresponding to the episodic steps in the task at hand.

In the end, the gaze prior maps from **DFG-P** mediate the temporal saliency maps from **Temporal Saliency Prediction** module. The bias from the task information is fused with the stimuli-driven bottom-up attention mechanism via an element-wise summation operation. We normalize the spatial prior maps and the temporal saliency maps to be within range $[0, 1]$ before element-wise summation. Concerned with the large variance of gradient changes in element-wise multiplication, we use element-wise sum-

mation instead to adaptively tune the effect of the task-specific bias on the bottom-up saliency. The results after element-wise summation are normalized again and the highest activation points on these probabilistic maps are the most probable anticipated gaze locations.

We should be cautious that, there is no top-down modulation in DFG. **DFG-P**, which carries task-specific information, is still a feed-forward 3D-CNN. Complementary to realistic visual features that guides gaze anticipation in **DFG-G**, **DFG-P** relaxes constraints on visual features and learns task-specific gaze priors or any abstract representations of the task useful for gaze anticipation. For example, in "spreading jam on bread" task, given the current frame showing the human subject puts the bread on the plate which is probably in the lower half of the egocentric view, **DFG-P** predicts high attention values to the upper half of the egocentric view (the table where all bottles are located) in the next few seconds due to the "jam bottle grabbing" task while **DFG-G** estimates the visual saliency of all bottles on the table and selects the jam-bottle like visual features.

5.1.4 Training

We train DFG end-to-end by stochastic gradient descent with learning rate 0.00005 and momentum 0.5. Adam Optimizer [209] is used. **Generator** and **Discriminator** play against each other. **Generator** is designed to predict future frames as "real" as possible to fool **Discriminator**, while **Discriminator** strives to tell real frames from the generated ones. These two networks try to minimize the maximum payoff of its opponent with respect to their network parameters w_D and w_G respectively. In addition, we add another *L1* loss term to ensure that the first generated video frame is visually consistent with the input frame without the over-smoothing artifacts. A hyper-parameter λ is used for tuning the weight of losses between the min-max game and the consistency term. Both networks are trained alternatively. The objective function for **Discriminator** is:

$$\begin{aligned}\min_{w_D} f_D(R_{t:t+N}, h) \triangleq L_{ce}(D(R_{t:t+N}; w_D), 1) \\ + L_{ce}(D(G(h; w_G)), 0),\end{aligned} \tag{5.2}$$

where h denotes the hidden representation $h(I_t)$ of input frame I_t, $R_{t:t+N}$ represents the real frames and the binary cross entropy loss L_{ce} is defined as

$$L_{ce}(\hat{Y}, Y) = Y \log(\hat{Y}) + (1 - Y) \log(1 - \hat{Y}), \tag{5.3}$$

where $Y \in \{0, 1\}$ denotes real or fake and $\hat{Y} \in [0, 1]$ denotes the output from **Discriminator**.

As the opponent of **Discriminator**, **Generator** needs to satisfy two requirements: 1) the generated outputs should be real enough to fool **Discriminator**; 2) the initial output of the generated frames should be visually consistent with the current frame. The objective function for training **Generator** is thus formulated as

$$\begin{aligned} \min_{w_G} f_G(I_t) \triangleq\ & L_{ce}(D(G(h; w_G)), 1) \\ & + \lambda \| I_t - G(I_t; w_G) \|_1, \end{aligned} \tag{5.4}$$

where λ is set as 0.1 which shows to achieve the best performance in our case. $\| \cdot \|_1$ denoting L1 distance is preferred over the mean square error which results in over-smoothing in the frame generation [210].

Temporal Saliency Prediction takes $I_{t+1,t+N}$ as input to generate temporal saliency maps. **Temporal Saliency Prediction** is trained in a supervised approach using Kullback-Leibler divergence (KLD) loss function:

$$KLD(P_i, Q_i) = \sum_x \sum_y P_i(x, y) \log \left[\frac{P_i(x, y)}{Q_i(x, y)} \right], \tag{5.5}$$

where P_i is the temporal fixation map and Q_i is the temporal saliency map for the $(t + i)$th frame. The fixation map refers to the binary map where we use 1 to indicate the human gaze location. To avoid sparseness of fixation maps, we convolve each binary fixation map with a gaussian mask and then we normalize it to be within range $[0, 1]$.

Similarly, **DFG-P** takes the latent representation $h(I_t)$ of the current frame I_t as the input to generate gaze spatial prior maps. We train **DFG-P** in a supervised manner using the same KLD loss function in Equation 5.5 where P_i is the temporal fixation map and Q_i is the gaze spatial prior map for the $(t + i)$th frame.

5.1.5 Implementation Details

DFG is developed based on [208] in Torch. The source code is available at `https://github.com/Mengmi/deepfuturegaze_gan`. We train everything from scratch with the input frame size being $3 \times 64 \times 64$. The batch size is 32. The latent representation $h(I_t)$ is of dimension $1024 \times 4 \times 4$ after 5 layers of 2D convolution layers for encoding image representation. We normalize all videos to be within the range $[-1, 1]$. The gaze spatial prior maps and the temporal saliency maps are of the same dimensions where $N = 32$, $W = 64$, and $H = 64$.

Gaze prediction on current frame DFG can also be used for gaze prediction on the current frame. Since **Generator** outputs a sequence of generated frames where the first frame must be consistent with the input frame due to *L1* distance loss in Equation(5.4), we take the spatial coordinate with the maximum probability in the first predicted temporal saliency map as the predicted gaze location on the current frame.

5.2 Experiments on Third-person and Egocentric Videos

We test DFG on gaze anticipation as well as gaze prediction over current frames on all public datasets using standard evaluation metrics. We also provide detailed analysis of DFG through ablation study and visualization of the learnt convolution filters. In the end, we demonstrate our anticipated gazes are useful in egocentric activity recognition.

5.2.1 Datasets

GTEA Dataset [206] This dataset contains 17 sequences on meal preparation tasks performed by 14 subjects. Each video clip lasts for about 4 minutes with the frame rate 15 fps and frame resolution 480×640. The subjects are asked to prepare meals freely. Same as Yin *et al*. [9], we use videos 1, 4, 6-22 as training set and the rest as test set.

GTEAplus Dataset [9] This dataset consists of 7 meal preparation activities. There are 5 subjects, each performing these 7 activities. Each video clip takes 10 to 15 minutes on average with frame rate 12 fps and frame resolution 960×1280. We do 5-fold cross validation across all 5 subjects and take their average for evaluation as [9].

Object Search Tasks (OST) To explore whether DFG can be generalized well for oth-

er tasks in egocentric contexts, we include the public egocentric video dataset in object search [207]. This dataset consists of 57 sequences on search and retrieval tasks performed by 55 subjects in a fully furnished and functional model home. Each video clip lasts for around 15 minutes with the frame rate 10 fps and frame resolution 480×640. Each subject is asked to search for a list of 22 items and move them to the packing location (dining table). Compared with GTEA and GTEAplus, this dataset involves larger head motions and the human subjects have to walk around and look for objects in the search list with hands appearing less frequently.

Hollywood2 Dataset [211] This is a public third person video dataset with 12 classes of human actions. [212] provides the gaze data for this dataset to study gaze dynamics. We include a subset of this dataset to evaluate DFG on gaze anticipation in the context of third person videos. In particular, video clips with these four actions related to social interactions are included in our experiment: handshaking, person hugging, kissing and person fighting. Among 3669 video clips in total, there are 365 video clips for training and 127 for testing and validation.

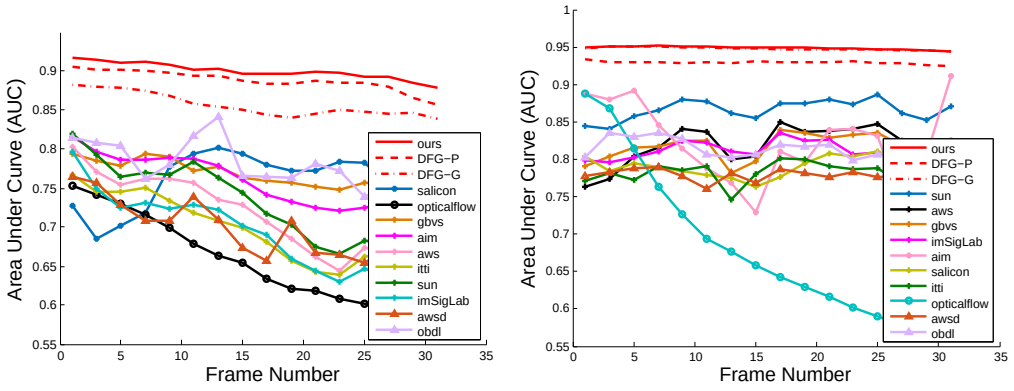
5.2.2 Evaluation Metrics

We use four standard evaluation metrics on gaze anticipation: Area Under the Curve (AUC) [195], Average Angular Error (AAE) [213], Normalized Scanpath Saliency (NSS) [203] and Precision-Recall Curve (PR) [214] as below.

Area Under the Curve (AUC) is the most commonly used saliency evaluation metric. It measures the area under a curve of true positive versus false positive rates under various threshold values on saliency maps.

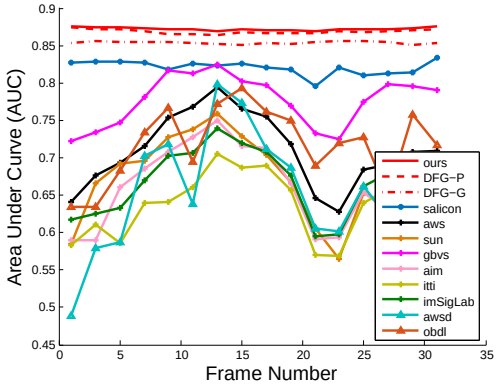
Average Angular Error (AAE) is the angular distance between the predicted gaze location and the ground truth.

Normalized Scanpath Saliency (NSS) computes the average normalized saliency at the fixated locations.

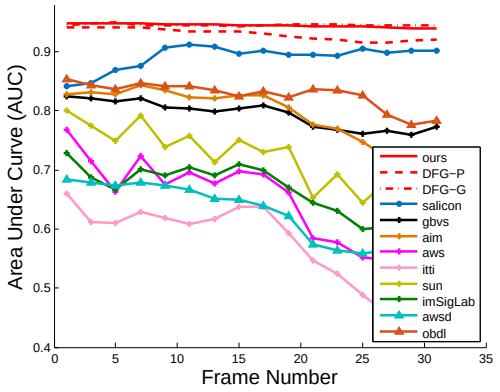


(a) GTEA Dataset (2.1 sec ahead)

(b) GTEAplus Dataset (2.7 sec ahead)



(c) OST Dataset (3.2 sec ahead)



(d) Hollywood2 Dataset (1.3 sec ahead)

Figure 5.3: Evaluation of Gaze Anticipation using Area Under the Curve (AUC) on the current frame as well as 31 future frames in GTEA, GTEAplus, OST and Hollywood2 Dataset.

Precision-Recall Curve (PR) represents results for binary decision in machine learning [214]. We report the area under the precision-recall curve at the ith future frame.

There are four datasets with four evaluation metrics resulting in 16 combinations. We report the gaze anticipation evaluation results in full using *all* evaluation metrics across *all* four datasets. For simplicity, in ablation study and architecture analysis, we opt to focus on reporting the analysis results on GTEA in egocentric videos and Hollywood2 in third person videos as representatives only using AUC and AAE. For consistency, except for Figure 5.3 and 5.4 where we show the metrics scores for all future 31 frames, we report the *mean* gaze anticipation accuracy by *averaging* the metrics scores over the current frame as well as the next 31 future frames.

5.2.3 Baselines

We create several competitive baselines as follows.

First, to show the effectiveness of end-to-end learning where all the parameters are trained jointly, we use **Generator** to generate future frames after the training phase and compare DFG with state-of-the-art saliency prediction algorithms on these frames including Graph-based Visual Saliency (GBVS) [144], Natural Statistics Saliency (SUN) [197], Adaptive Whitening Saliency (AWS) [198], Attention-based Information Maximization (AIM) [149], Itti's Model (Itti) [199], and Image Signature Saliency (ImSig) [200]. Moreover, we also include gaze prediction methods on videos [161] (AWSD) and [160] (OBDL).

Second, SALICON [4] is a deep learning architecture for saliency prediction on static images. We train SALICON from scratch on the egocentric datasets by using real frames and their corresponding fixation maps. After that, the pre-trained SALICON model is tested on our generated frames for gaze anticipation.

Third, we create another baseline (OpticalShift) to study the effect of temporal dynamics. We use our model to predict gaze on the current frame and compute the dense optical flow between the previous frame and the current frame using [205]. The predicted gaze is then warped to the future frames by shifting it based on the flow at that position as the future gaze locations.

Fourth, we include the graph-based method to model gaze transition dynamics as proposed by [9] for gaze prediction on current frames in GTEA and GTEAplus. We exclude this method on OST since the required hand annotations by [9] are not available. We also cannot extend this method to gaze anticipation problem.

5.3 Results of Gaze Anticipation

In this section, we provide results for gaze anticipation on egocentric and normal videos.

5.3.1 Results on Egocentric Videos

DFG surpasses all the competitive baselines significantly in gaze anticipation in egocentric videos. We report the quantitative evaluation results in Figure 5.3 (AUC), Figure 5.4

Table 5.1: Averaged gaze anticipation performance over current frame as well as 31 future frames using Normalized Saliency Scanpath (NSS) and the area under the Precision-Recall Curve (PR).

	GTEA		GTEAplus		OST		Hollywood2	
Metrics	NSS	PR	NSS	PR	NSS	PR	NSS	PR
ours	**1.62**	**0.50**	**1.95**	**0.53**	1.45	**0.48**	**1.91**	**0.56**
SAL [4]	0.97	0.46	1.11	0.43	**1.91**	0.45	1.76	0.49
GBVS [144]	0.94	0.42	1.52	0.44	0.75	0.43	0.54	0.41
AWS [198]	0.73	0.39	0.74	0.42	0.13	0.39	-0.05	0.41
AIM [149]	0.91	0.39	0.85	0.39	0.55	0.42	0.73	0.41
SUN [197]	0.77	0.38	1.58	0.46	0.74	0.41	0.65	0.38
Itti [199]	0.67	0.40	1.01	0.40	0.18	0.43	-0.22	0.41
ImSig [200]	0.62	0.38	1.03	0.39	0.40	0.42	0.56	0.41
AWSD [161]	0.69	0.40	1.06	0.42	0.56	0.41	0.44	0.41
OBDL [160]	1.02	0.42	1.21	0.42	0.78	0.42	1.14	0.44

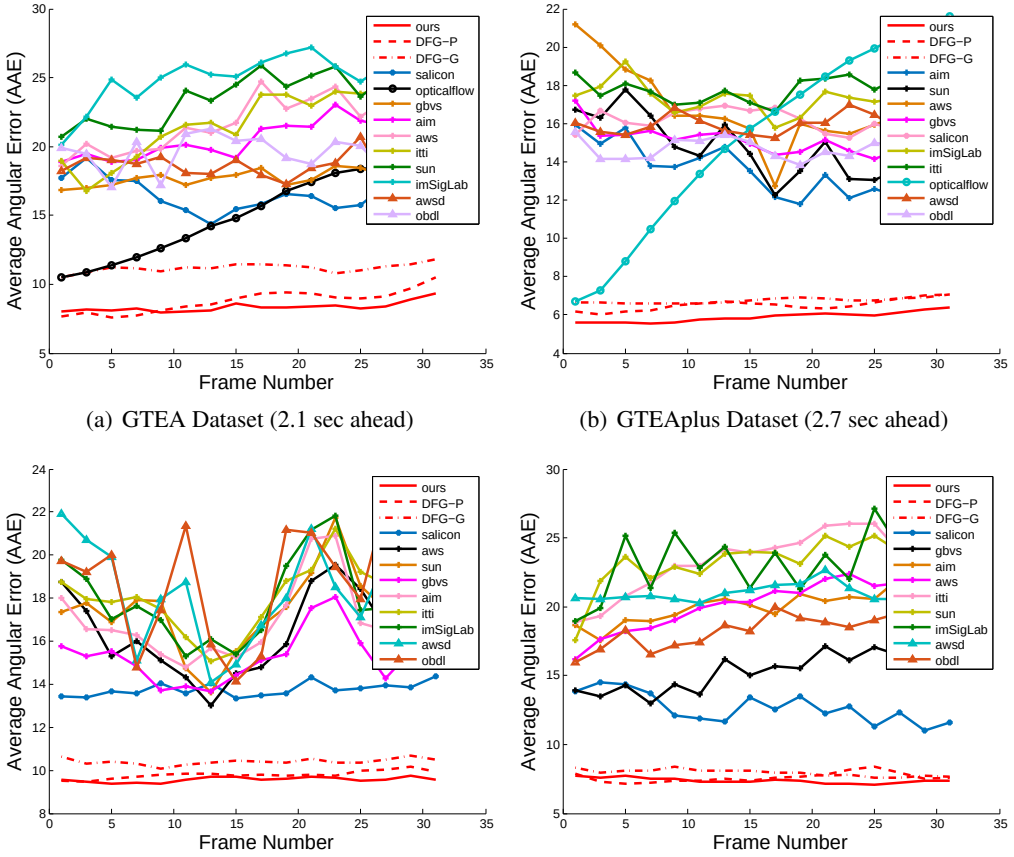
(AAE) and Table 5.1 (NSS and PR) on egocentric datasets.

Over all egocentric datasets (GTEA, GTEAplus, and OST), DFG outperforms all the competitive baselines. In particular, we observe a significant performance boost with respect to our previous method (**DFG-G**) [207] which is the second best as shown in Figure 5.3 and 5.4 by 26.2%, 12.0% and 8.8% in relative advance (RA) in AAE and 4.5%, 0.05% and 2.3% in RA in AUC. RA in percentage is computed as

$$\mathrm{RA}(\mathrm{OUR}, \mathrm{BB}) = \frac{\| \sum_{i=1}^{N} \mathrm{OUR}_i - \sum_{i=1}^{N} \mathrm{BB}_i \|}{\sum_{i=1}^{N} \mathrm{BB}_i}, \tag{5.6}$$

where N=32 is the number of generated future frames, OUR_i is the metric score of our model and BB_i is the metric score of **DFG-G** on the ith future frame. Complementary to **DFG-G**, fusion with **DFG-P** greatly improves the gaze anticipation performance which emphasizes the necessary role of **DFG-P** pathway which predicts gaze priors for the task at hand and biases the saliency maps predicted by **DFG-G**. See Section 5.7 for more analysis.

Qualitative results in Figure 5.5 demonstrate that DFG learns to untangle foreground and background motions. Our DFG model produces 31 future frames based on the current frame. From first to last rows, results on future frames #1, 5, 9, 17, 29 with respect to the current frame are shown. The leftmost column shows the ground truth (GT) with red circle denoting human gaze locations. Column 2, 3, 4 (FG, mask, BG) show the foreground $F(\cdot)$, the mask $M(\cdot)$, and the background $B(\cdot)$ learnt by **Generator** respec-



(a) GTEA Dataset (2.1 sec ahead)

(b) GTEAplus Dataset (2.7 sec ahead)

(c) OST Dataset (3.2 sec ahead)

(d) Hollywood2 Dataset (1.3 sec ahead)

Figure 5.4: Evaluation of Gaze Anticipation using Average Angular Error (AAE) on the current frame as well as 31 future frames in GTEA, GTEAplus, OST and Hollywood2 Dataset.

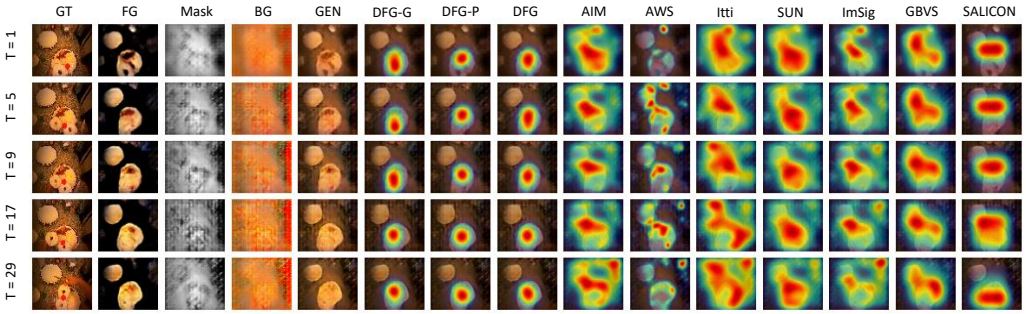


Figure 5.5: Example results of gaze anticipation on GTEAplus egocentric video dataset.

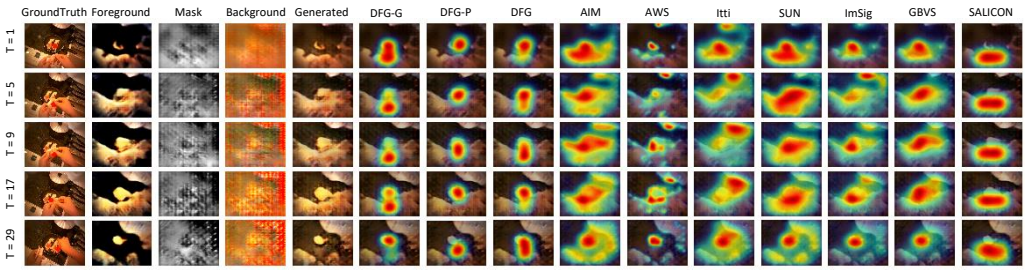


Figure 5.6: Example results of gaze anticipation on egocentric video datasets. The format and conventions follow those in Figure 5.5.

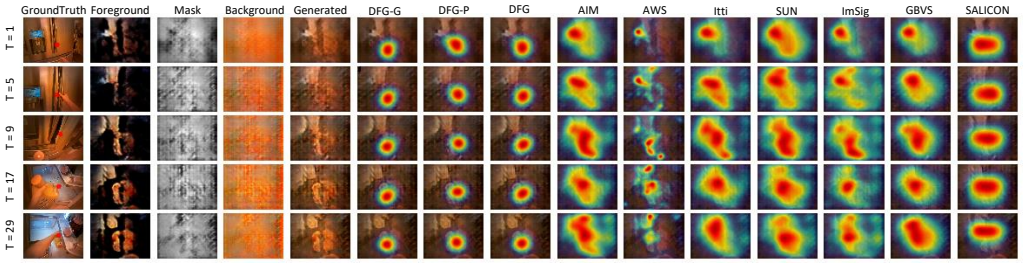


Figure 5.7: Example results of gaze anticipation on egocentric video datasets. The format and conventions follow those in Figure 5.5.

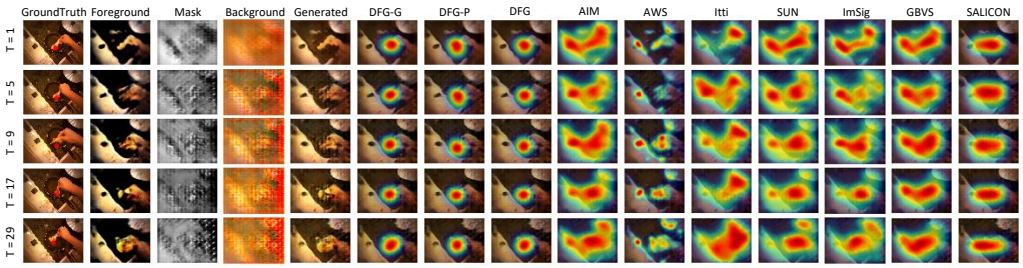


Figure 5.8: Example results of gaze anticipation on egocentric video datasets. The format and conventions follow those in Figure 5.5.

tively. Column 5 shows the generated future frames (GEN). Column 6 and 7 show the corresponding predicted temporal saliency maps from two pathways **DFG-G** and **DFG-P** in our model. Column 8 show the final integrated temporal saliency maps predicted by our model. Column 9 and onwards show the predicted temporal saliency maps by all baselines (See Section 7.2.2). For example, both the hand and the object (the bun) get highlighted in the foreground. As the high intensity value on the mask denotes the foreground, the manipulation point (the control point where the subject is manipulating the object with hands) shows the highest activation on the mask whereas the background (the table surface) is uniform over time as shown in the darker regions of the mask.

It is also observed that the temporal saliency maps anticipated by **DFG-P** and **DFG-G** are visually different. Though **DFG-P** assigns high attention values to the manipulation point (slightly below the center of the egocentric field of view across all future frames in general during the table-top food preparation process), it fails to capture the hand motion when the subject is rotating the bun within the local region; conversely, **DFG-G** anticipates the effect of local hand motion and hence, predicts slight attention shifts in the future frames. More qualitative results in Figure 5.6, 5.7, 5.8 demonstrate that **DFG-G** and **DFG-P** can be jointly adapted in different tasks which cover varieties of illumination conditions, head orientations, hand poses, and manipulated objects.

Though SALICON learns an abundance of semantic information, it excludes temporal dependencies which are crucial for gaze anticipation on egocentric videos. Although SALICON has performed better than conventional saliency prediction methods, its performance is inferior to DFG which learns spatial-temporal information.

For OpticalShift, we observe that its AUC and AAE curves drop monotonically. It confirms that the optical flow computed from the current state cannot adapt to the complexity of the temporal dynamics in longer time periods.

We provide comparisons with gaze prediction methods on videos [161, 160]. Although these methods take temporal information into account, these feature cues (space-time whitening and information from video compressors) on synthetic frames are still not sufficient compared with **DFG-G** [207]. Another missing element in these models is task-specific information which is also critical for gaze anticipation.

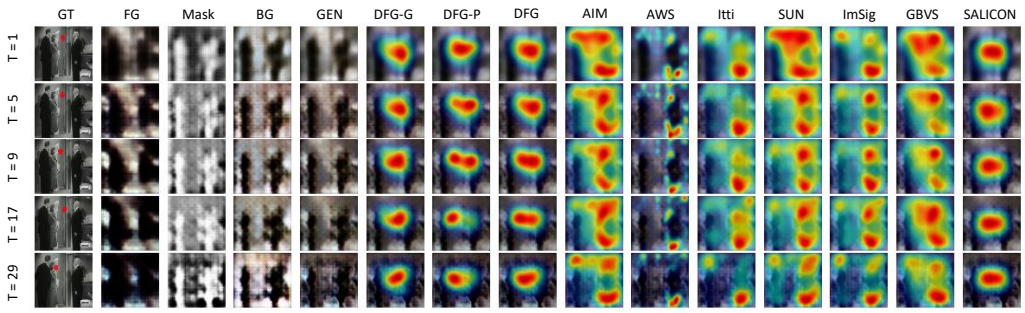


Figure 5.9: Example results of gaze anticipation on Hollywood2 third person video dataset. The format and conventions follow those in Figure 5.5.

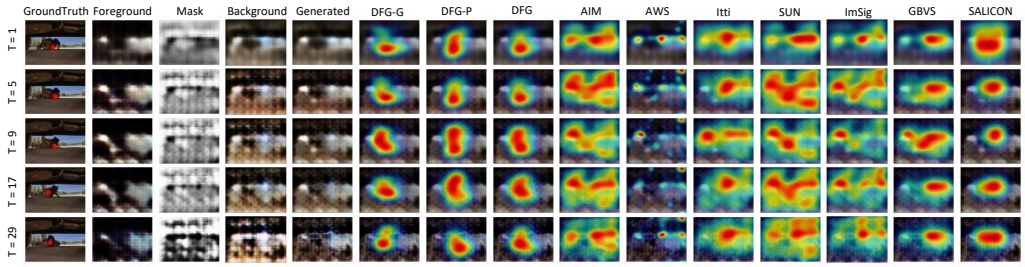


Figure 5.10: Example results of gaze anticipation on Hollywood2 third person video dataset. The format and conventions follow those in Figure 5.5.

5.3.2 Results on Normal Videos

Beyond egocentric videos, we test DFG on third person videos where the backgrounds are often static. From the quantitative evaluation results in Figure 5.3(d) (AUC), Figure 5.4(d) (AAE) and Table 5.1 (NSS and PR), DFG achieves the best performance in Hollywood2 dataset with four evaluation metrics. Using Equation 5.6, DFG outperforms our previous method (**DFG-G**) [207] by 7.1% in relative advance (RA) in AAE and 0.09% in RA in AUC.

We present a qualitative example in Figure 5.9 in hand shaking scenario in Hol-

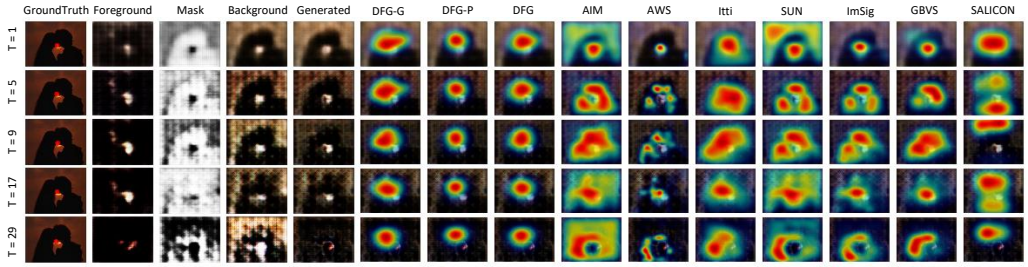


Figure 5.11: Example results of gaze anticipation on Hollywood2 third person video dataset. The format and conventions follow those in Figure 5.5.

lywood2. From the results, it demonstrates that DFG is also capable of segmenting foreground objects from static backgrounds in third person videos. For example, the three persons get highlighted in the mask. As the background is uniform over time, this is reflected in the darker regions of the mask as well as the bright regions in the background stream. Furthermore, we also observe that DFG can adaptively generate "realistic" future frames regardless of variant color conditions, such as the gray-scale video frames as shown in Figure 5.9.

We also note that **DFG-P** learns the general gaze anticipation patterns when it requires complex gaze shifts while human subjects are observing a video clip in a social interaction task. The qualitative example in Figure 5.9 shows an occasion where three persons are having a conversation. Though there is no significant visual change in this social interaction case and **DFG-G** predicts almost static future frames over time, **DFG-P** anticipates attention spread across the three persons where the highest activation points on the saliency maps shift from the center to the left across frames which is consistent with the ground truth gaze patterns.

Compared with the performance on egocentric videos, SALICON performs relatively better on third person videos. This is because the backgrounds in video clips in Hollywood2 are often static which alleviates the demands of temporal information. In addition, the semantic information such as faces appear often in social interaction tasks where SALICON is good at attending to these semantic objects on each frame. The performance of the rest of the baselines on Hollywood2 is consistent with those in egocentric videos.

5.4 Spatial Bias Analysis

In this section, we study the various spatial biases including center bias, gaze fixation distribution from the training data as well as head motion and how they may effect the gaze anticipation performance in egocentric and normal videos.

5.4.1 Center Bias

We often observe a strong center bias in egocentric videos. This is due to the fact that egocentric videos are captured from the first person view. Humans always move

Table 5.2: Evaluation of Center Bias Effect over the Next 31 Frames

sAUC	GTEA	GTEAplus	OST	Hollywood2
DFG(ours)	**0.62**	**0.57**	**0.57**	**0.52**
Center Bias	0.5	0.5	0.49	0.49

Table 5.3: Average Spatial Bias and Human Performance over the Next 31 Frames on GTEA and GTEAplus Datasets.

	GTEA		GTEAplus	
	AUC	AAE	AUC	AAE
Our Best	**0.90**	**8.3**	**0.94**	**5.9**
GazeDistriMap	0.86	9.3	0.93	7.4
GazeDistriMap + DFG-G	0.88	9.0	0.94	6.8
Human	0.66	9.5	0.77	6.8

their heads to attend to the regions of interest. In this case, gazes often align with head orientations. Thus, gaze shift in the large distance gets compensated by head movements with small gaze shifts. Similarly, center bias is also present in free-viewing tasks in static images and third person videos [215]. As AUC favors center bias, we use shuffled-AUC (sAUC) to compare our model with center bias and we report its sAUC score in Table 5.2. It confirms that our model learns to anticipate gaze by taking various semantic information and motion dynamics into account instead of predicting center bias on future frames over all datasets.

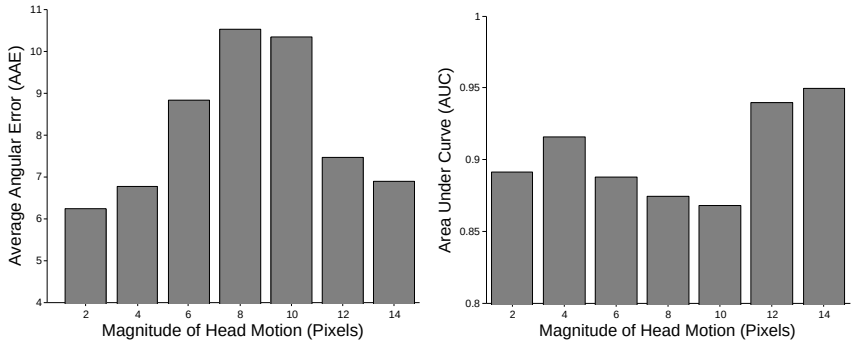
5.4.2 Gaze Distribution Map

We report the two variations of utilizing the 2D gaze distribution map computed from all human fixations in the training set: (1). the 2D gaze distribution map alone as the predicted temporal saliency map on all future frames; (2) we replace **DFG-P** in our DFG model with the gaze distribution map.

Table 5.3 shows the gaze distribution map alone (Row 2) is much worse than our DFG model (Row 1). Though **DFG-G** with gaze distribution map (Row 3) is better than gaze distribution alone, it is still inferior to DFG by 1 in GTEA and 1.5 in GTEAplus in terms of AAE. This suggests the gaze prior has complex dynamics and **DFG-P** which learns gaze prior variations depending on the task specifications is important for gaze anticipation.

Table 5.4: Statistics of Camera and Gaze Motions

	Gaze Motion			Camera Motion		
	Mean	Median	Variance	Mean	Median	Variance
GTEA	20.4	13.5	508	6.7	3.6	92
GTEAplus	7.1	5.0	89	9.9	5.8	135



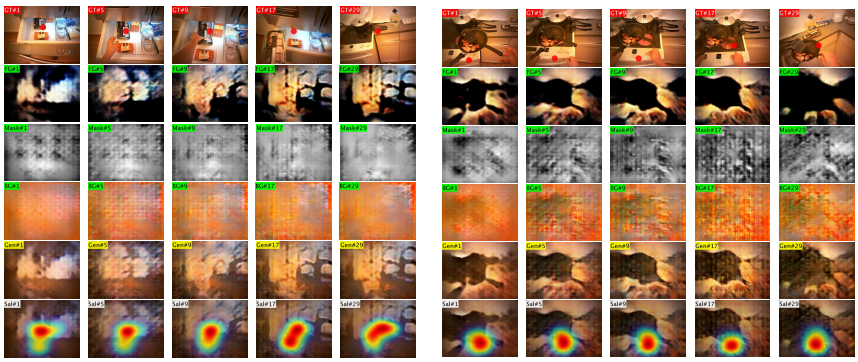
(a) Average Angular Error (b) Area Under the Curve

Figure 5.12: Evaluation of Average Gaze Anticipation Performance over 31 future frames versus magnitude of head motions in GTEA

5.4.3 Head Motion

We provide the statistics of head and gaze motion in pixels in our test data in GTEA and GTEAplus datasets. As there is no ground truth for head motion, we estimate it by averaging the dense optical flow in the boundary pixels between adjacent frames. With respect to a frame (480 by 640 in pixels), the statistics of amplitudes for these motion are reported in Table 5.4. To study the effect of head motion on gaze anticipation, we calculate the averaged magnitude of head motion across the next 31 ground truth frames and report the averaged gaze anticipation performance on these frames in Figure 5.12.

In general, the gaze anticipation performance of our DFG model drops when there is larger head motion. Here we show two examples where **Generator** fail to synthesize realistic future frames due to large head motion. We quantify the large head motion as the averaged magnitude of head motion vector to be larger than 6 pixels calculated based on optical flow on boundary pixels over the next 31 future frames. In Figure 5.13(a), the anticipated gaze location still matches the ground truth despite the large head motion but in Figure 5.13(b), it fails. In each example, frames #1, 5, 9, 17, 29 are shown (left to right columns). The topmost row shows the ground truth with red circle denoting human gaze locations. Row 2, 3, 4 show the foreground $F(\cdot)$, the mask $M(\cdot)$, and



(a) Example 1 (Head motion = 6.4) (b) Example 2 (Head motion = 27.8)

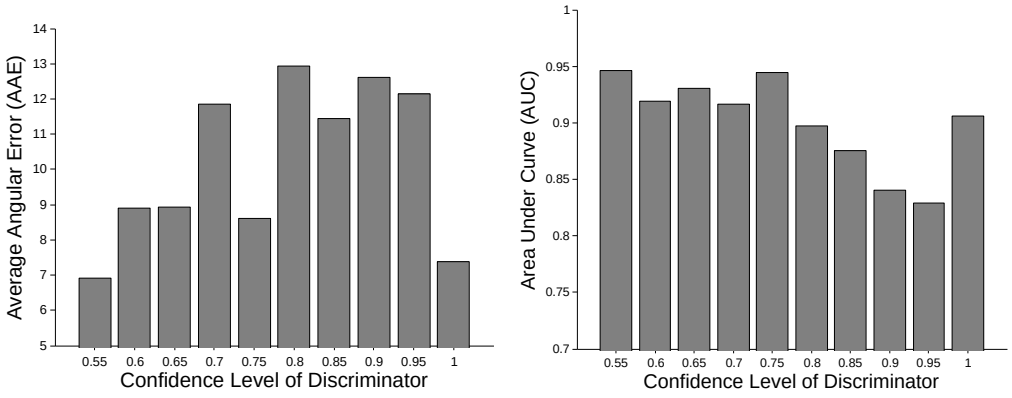
Figure 5.13: Example results of gaze anticipation when there is large head motion.

the background $B(\cdot)$ learnt by **Generator Network** respectively. Row 5 shows the generated future frames. Row 6 shows the corresponding predicted temporal saliency maps.

These two examples again validate the point that egocentric videos have characteristics of having small gaze shifts in space as they often get compensated by the head motion. However, this phenomenon does not imply that either center bias or the gaze distribution map from all human fixations in the training set is sufficient for gaze anticipation. Due to the complex nature and large variances between gaze and head motions, our analysis confirms that the two-stream **Generator** in our DFG model is critical for better gaze anticipation by estimating the these two motions separately.

5.5 Discrepancy of Future Frames from Real Scenes

We study how discrepancy of the future frames from the real scene will effect gaze anticipation performance. To quantitatively evaluate the quality of the generated future frames from **Generator**, we compute the confidence of **Discriminator** which acts as a competitor against **Generator** striving to distinguish whether the generated frames are real or synthetic. The more confident **Discriminator** is, the easier for **Discriminator** to tell real ones from the synthetic; hence, the more discrepancy there is between the generated future frames generated by **Generator** and the real scene. Ideally, if the synthetic frames are indistinguishable from real frames, the **Discriminator** confidence is 0.5. Figure 5.14 shows the average gaze anticipation performance over the next 31



(a) Average Angular Error (b) Area Under the Curve

Figure 5.14: Evaluation of Average Gaze Anticipation Performance over 31 future frames versus confidence of **Discriminator** in our model in GTEA

future frames versus the confidence of **Discriminator**. The gaze anticipation performance is positively correlated with the quality of the generated frames which validates that **Discriminator** is critical for providing feedbacks to **Generator** in order to generate more realistic future frames useful for improving gaze anticipation performance.

5.6 Human Performance on Gaze Anticipation

As human benchmark is a gold standard in many computer vision tasks and it is not clear how humans perform in our gaze anticipation task, we conduct human psychophysics experiments to test human performance in this task. For fair comparison with the computational models, we provide 4 human subjects (22-28 years old, 2 females, 2 males) with two training phases and test them on gaze anticipation tasks on 50 video clips per test set from GTEA and GTEAplus datasets. See Figure 5.15 for experiment schematics. Here we provide detailed description of the psychophysics experiment.

The experiment started with a briefing on the study's objectives and procedures. During briefing, all participants are instructed on the objectives of the study: comparison between algorithms and human performance on the gaze anticipation task. The gaze anticipation task is prediction of gaze point on future unseen frames from a single video frame. Participants were given unlimited time to complete the task. There are 2 sessions with each session containing 50 testing video clips either from GTEA or GTEAplus datasets.

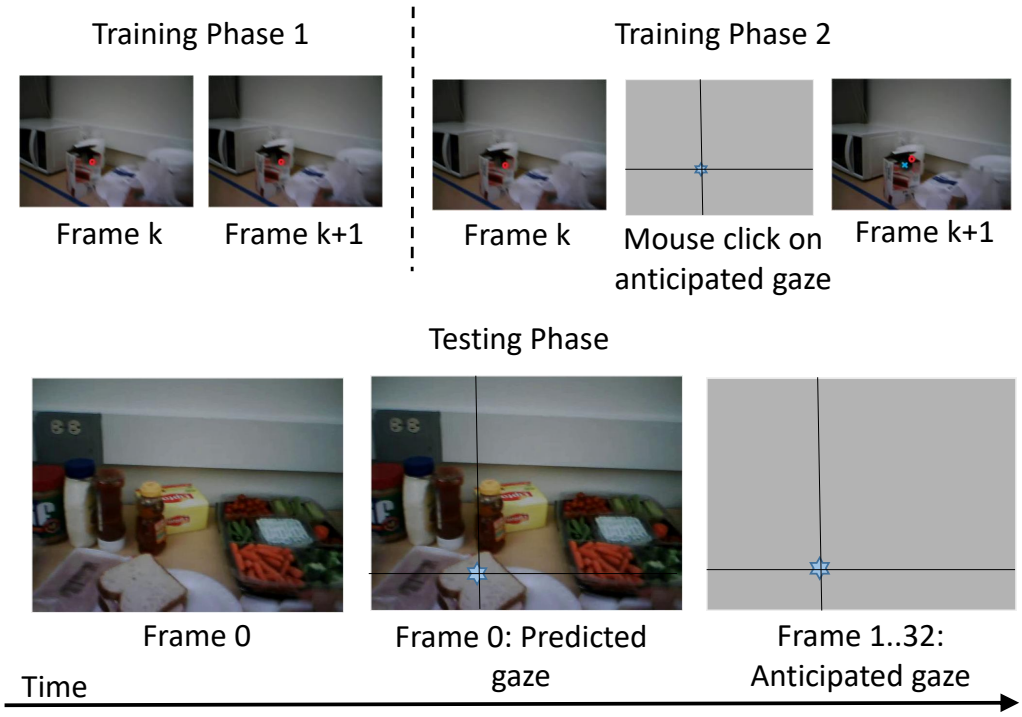


Figure 5.15: Schematic description of human psychophysics experiment on gaze anticipation.

In each session, there are 2 training phases and 1 testing phase. Training phase 1 is to familiarize the participants with the system. Training phase 2 is similar to the supervised learning of our model. Testing phase is the same setup as our machine experiments, that is given one frame, anticipate the gaze positions for some future frames.

In training phase 1, the participant was shown a video frame. It is the ego-centric view of the scene with the recorded gaze (red circle) overlay on it. This is repeated for all frames of each video clip. There are 5 video clips during this training phase.

In training phase 2, participant was shown a video frame followed by a blank gray screen. The participant was then instructed to imagine the next frame and click on their anticipated gaze location for the next frame. The participant was shown the recorded gaze (red circle) overlay on the next frame (i.e. ground truth). The user's mouse click position (blue cross) was also overlaid as the feedback to the participant. This was repeated for all frames of all video clips. There are also 5 video clips for this training phase.

In testing phase, participant was shown a video frame. The participant was then instructed to click on their location of the predicted gaze for this frame. A blank gray

screen was shown to the participant. The participant was then instructed to imagine the next frame and click on the anticipated gaze location for the subsequent frame. The blank gray screen was repeated for 32 frames of the video clip. For each blank screen, participant imagined the future frame and clicked on the anticipated gaze location. The test set includes a total of 100 testing video clips (50 clips per dataset in GTEA and GTEAplus).

We report the average human performance on gaze anticipation task over the next 31 future frames in Table 5.3, Row 4. Human performance is as good as gaze fixation maps with **DFG-G** but still inferior to our DFG model. However, this result cannot be over-interpreted as there are several differences between humans and the computational models: (1) number of training samples (humans are exposed to fewer training samples compared with DFG); and (2) knowledge of the tasks (humans do not have full knowledge about all the task information in each dataset while computational models are trained with more varieties of tasks). This is an interesting future research direction and it suggests promising real life applications where the computational models could assist humans in several domains involving gaze anticipation, such as health care and autonomous driving.

5.7 Ablation Study on Egocentric and Normal Videos

In order to study the effect of the individual component of DFG on both egocentric and third person videos, we do an ablation study and test on GTEA, OST and Hollywood2 datasets by removing *only* one component in DFG at one time while the rest of the architecture remains the same. There are five tests: (1) we remove **DFG-G** and evaluate the predicted temporal saliency maps from **DFG-P** only; (2) we remove **DFG-P** and this is the same as our previous algorithm with only **DFG-G** [207]. (3) we replace the two-stream 3D-CNN in **Generator** with the same structure as [208], *i.e.* the background stream is 2D-CNN which assumes the background is "static" while the foreground stream remains the same; (4) we train **Temporal Saliency Prediction** directly on real frames and test it on the generated frames from **Generator**; (5) we remove **Discriminator** and we only use L1 distance loss for future frame generation. Scores for gaze anticipation in AAE and AUC are averaged across future 31 frames as shown

Table 5.5: Ablation Study on GTEA, OST and Hollywood2 Datasets

	GTEA		OST		Hollywood2	
	AUC	AAE	AUC	AAE	AUC	AAE
Our Best (DFG)	**0.90**	**8.3**	**0.87**	**9.5**	**0.95**	**7.4**
DFG-P	0.88	8.9	0.87	9.8	0.93	7.5
DFG-G	0.86	11.3	0.85	10.3	0.94	7.9
One-stream	0.85	12.0	0.86	10.5	0.95	7.7
Replace(GT)	0.82	13.5	0.80	13.0	0.86	12.6
Remove(D)	0.83	12.0	0.85	10.6	0.88	14.3

in Table 5.5.

Compared with our previous method **DFG-G** [207], we proposed a complementary task-specific **DFG-P** and integrated it with **DFG-G**. To study its effectiveness, we test each of these two pathways individually. **DFG-P** alone performs better than **DFG-G** by 2.4 in GTEA, 0.5 in OST and 0.4 in Hollywood2 in terms of AAE but both pathways are worse than our integrated framework (DFG). We also duplicate the results of **DFG-P** (Row 2) and **DFG-G** (Row 3) in Figure 5.3 and Figure 5.4. We observe that both individual pathways outperform all the baselines significantly. It suggests that both the bottom-up attention mechanism **DFG-G** and the gaze prior maps predicted from task-specific information by **DFG-P** have essential contributions to gaze anticipation in egocentric and third-person videos.

5.7.1 Ablation Analysis on Egocentric Videos

The third ablation study (Row 4) on changing the background stream to a static one leads to an increase of 3.7 in GTEA and 1 in OST in terms of AAE. This implies the two-stream 3D-CNN in **Generator** is essential for learning foreground and background motions which can further improve gaze anticipation accuracy.

Compared with DFG, the fourth ablated model (Row 5) with **Temporal Saliency Prediction** trained on real frames performs worse with an increase of 5.2 in GTEA and 3.5 in OST in terms of AAE. In DFG, **Temporal Saliency Prediction** is attached after **Generator** for temporal saliency map prediction using end-to-end training. However, **Temporal Saliency Prediction** in the third ablated model, which are trained only on real frames, cannot perform well since it cannot learn the essential features on the generated frames. It demonstrates that the features on the generated frames are different

Table 5.6: Results of Gaze Prediction on the Current Frame

	GTEAplus		GTEA		Our OST		Hollywood	
Metrics	AUC	AAE	AUC	AAE	AUC	AAE	AUC	AAE
DFG(ours)	**0.95**	**5.6**	**0.92**	**8.1**	**0.88**	**9.6**	**0.95**	**7.75**
DFG-P	0.93	6.2	0.9	7.69	0.88	9.5	0.94	7.9
DFG-G [207]	0.95	6.6	0.88	10.5	0.85	10.6	0.95	8.3
Yin [9]	0.87	7.9	0.88	8.4	-	-	-	-
SAL [4]	0.82	15.6	0.76	16.5	0.85	13.3	0.84	14.0
GBVS [144]	0.80	14.7	0.77	15.3	0.71	18.8	0.75	10.5
AWS [198]	0.82	14.8	0.78	17.5	0.56	22.8	0.5	17.5
AIM [149]	0.76	15.0	0.82	14.2	0.77	17.0	0.75	14.4
SUN [197]	0.84	14.7	0.80	18.1	0.53	25.0	0.66	17.7
Itti [199]	0.75	19.9	0.75	18.4	0.62	19.0	0.67	26.7
ImSig [200]	0.79	16.5	0.78	19.0	0.56	24.2	0.60	20.9
AWSD [161]	0.78	16.0	0.77	18.2	0.49	21.9	0.68	20.6
OBDL [160]	0.82	19.9	0.80	15.6	0.63	19.7	0.85	16.0

from those on real frames and hence, end-to-end training is necessary for **Temporal Saliency Prediction** to learn these essential features on the generated future frames.

The fifth ablation study with **Discriminator** removed (Row 6) shows an increases of 3.7 in GTEA and 1.1 in OST in terms of AAE. This demonstrates that **Discriminator** is important as the feedback to **Temporal Saliency Prediction** which provides the additional constraints such that **Generator** can generate more "realistic" future frames in longer time duration. These "realistic" future frames are critical for gaze anticipation.

5.7.2 Ablation Analysis on Normal Videos

Results in Hollywood2 dataset show DFG outperforms **DFG-G** by 0.5 and **DFG-P** by 0.1 in Hollywood2 in terms of AAE. Compared with GTEA, we observe that the task-specific influences from **DFG-P** have less impacts in Hollywood2 which is a third person video dataset. As gaze information reflects human intention and behaviors, this implies that the gazes in egocentric videos are often guided by willful plans or current goals as task-specific attentional effect. This has also been verified in the literature [216, 217].

The third ablated model (Row 4) has shown marginal effect in Hollywood2 with an increase of 0.3 in terms of AAE while there is an increase of 3.7 in GTEA dataset. As the backgrounds in Hollywood2 are often static in most cases, the 2D-CNN stream in **Generator** in the ablated model could still model the semantics on the background

in normal videos. However, in GTEA, the second ablated model cannot learn complex motion dynamics in the backgrounds which leads to a significant performance drop. This further verifies the necessity of splitting **Generator** into two 3D-CNN streams in order to model the foreground and background motions in egocentric videos.

Compared with DFG, the fourth ablated model (Row 5) with **Temporal Saliency Prediction** trained on real frames performs worse with an increase of 5.2 in Hollywood2 in terms of AAE. It implies that the end-to-end training on the generated frames is equivalently important in both egocentric videos and third person videos such that **Temporal Saliency Prediction** can learn essential features on the synthesized frames.

The fifth ablation study (Row 6) with **Discriminator** removed shows an increases of 6.9 in Hollywood2 in terms of AAE. This again validates the point that **Discriminator** plays a critical role in generating more realistic future frames. Moreover, we note that the performance drops more in Hollywood2 compared with GTEA. This implies that **Discriminator** is more important in the case of third person videos as the supervision from **Didscriminator** prevents over-fitting problems of **Temporal Saliency Prediction** in a more simplified task where there is less motion involved.

5.8 Results on Current Frame Gaze Prediction

We compare DFG with state-of-the-art saliency prediction algorithms in Section 7.2.2 on real frames in the testsets of all egocentric and third person video datasets and we report both AAE and AUC scores of gaze prediction on current frames in Table 5.6. Number denoted in bold is the best. Results show that DFG performs better than the-state-of-the-arts even without explicitly specifying useful visual cues, such as hands, objects of interest and faces. Moreover, different from the traditional methods, our model takes the current frame as the only input without any past information. Compared with **DFG-G**, we observe that AAE scores decrease significantly and even surpass Yin *et al*. [9] on GTEA. It implies that the integration of task-specific information from **DFG-P** with **DFG-G** contributes to gaze prediction on current frames.

Table 5.7: Evaluation of Gaze Anticipation on Frames at Time $t + 16$ and $t + 32$

Average Angular Error (AAE)				
	GTEAplus		GTEA	
Models	Ours(DFG)	SALICON	Ours(DFG)	SALICON
time $t + 16$	**6.0**	11.4	**8.4**	18.4
time $t + 32$	**6.5**	19.5	**9.0**	16.6
Area Under Curve (AUC)				
	GTEAplus		GTEA	
Models	Ours(DFG)	SALICON	Ours(DFG)	SALICON
time $t + 16$	**0.939**	0.916	**0.891**	0.710
time $t + 32$	**0.937**	0.722	**0.873**	0.767

5.9 Analysis on Temporal Dependency of Gaze States

It is observed that the gaze movement on individual frames is dependent on their previous states; *e.g.* to anticipate gaze on the frame $t + 32$, we need to consider gaze transitions across frames by also anticipating gaze on frames t to $t + 31$. For verification, we created one baseline: train SALICON model, a 2D-ConvNet, directly for gaze anticipation at time $t + 16$ and $t + 32$ using their respective ground truth at time $t + 16$ and $t + 32$. See Table 5.7 for results in terms of AUC and AAE on GTEA and GTEAplus. Number denoted in bold is the best. DFG performs much better than SALICON. This suggests the temporal dependence across frames plays fundamental roles in gaze anticipation in egocentric videos and future frame generation using GANs is useful.

Table 5.8: Correlation Between Number of Frames and Corresponding Performance of Our Model

Angular Average Error (AAE)					
	# 1−2	# 3−4	# 5−8	# 9−16	# 17−32
#2	10.4	–	–	–	–
#4	10.7	10.9	–	–	–
#8	10.4	10.4	10.3	–	–
#16	10.2	10.0	10.3	10.8	–
#32	8.0	8.0	8.0	8.2	8.5
Area Under the Curve (AUC)					
	# 1−2	# 3−4	# 5−8	# 9−16	# 17−32
#2	0.87	–	–	–	–
#4	0.86	0.86	–	–	–
#8	0.87	0.87	0.86	–	–
#16	0.88	0.88	0.87	0.86	–
#32	0.91	0.91	0.91	0.90	0.89

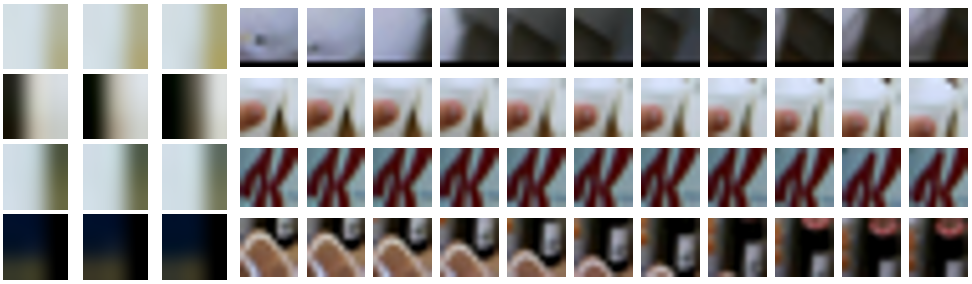
5.10 Analysis on Frame Numbers

In video analysis, the number of consecutive frames is a key parameter in practice. To study the effect of the number of frames on which we anticipate gaze, we assign the scalar weights to tune the losses in both **Generator** and **Temporal Saliency Prediction** for the next 32 frames while maintaining the same architecture. For example, we design the weight matrix to be $[1, 1, 1, 1, 0, ..., 0]$ for gaze anticipation in the next 4 frames while ignoring the subsequent frames. In Table 5.8, we present the averaged metric scores of our model for gaze anticipation in the next 2, 4, 8, 16, 32 frames starting from the current frame #1. Scores for gaze anticipation in both AAE and AUC are computed every # frames indicated in columns in the testset in GTEA Dataset.

From the results, we observe that given an input frame, in order to anticipate gazes on subsequent L frames, models trained with $L + K$ frames will perform better as K increases. This is because **Temporal Saliency Prediction** can learn the temporal dynamics with more information flowing back from the future K frames.

5.11 Visualization of Convolution Filters

As **Temporal Saliency Prediction** estimates temporal saliency maps based on the generated frames, we analyze the learnt convolution filters in **Temporal Saliency Prediction** and align the observations with human bottom-up visual attention mechanism.



(a) GP1 (b) GP4

Figure 5.16: Visualization of the convolution filters in the first (GP1) and the second last (GP4) 3D convolution layers of **Temporal Saliency Prediction Module** in our DFG model.

[218] proposed a top 4 patch visualization approach in 2D-CNN. We extend their work to visualization of 3D-CNN. As a simplified version of their method, we parse all video frames from the test set in GTEA and record the regions with the highest filter activation in both spatial and temporal dimensions for the first and the second last convolution layer in **Temporal Saliency Prediction** in our model. Those regions are then projected back into their input video frames based on their corresponding receptive fields across both space and time dimensions where the input frames are the current frame and its subsequent 31 frames. Due to the consistency of egocentric videos between adjacent frames, we increase the diversity of the visualization by sorting the filter activation from highest to lowest and selecting these top filters where their receptive fields do not overlap with their neighboring frames by a pre-defined threshold.

We observe that the filters in the first convolution layer of **Temporal Saliency Prediction** learn the low level features, such as edges and regions of high contrast. This observation aligns well bottom-up visual attention which is driven by low level features at the initial stage according to [6]. More interestingly, we also find the learnt features change across time, e.g. the black region increases from left to right across time (row 2 in Figure 5.16(a)) and the brightness in the bottom regions decay across time (row 4 in Figure 5.16(a)). This demonstrates DFG learns motion dynamics such as translation and the gradient change of surfaces. As the level of convolution layers increases, we can see more complex patterns. In the second last layer, the regions containing semantic information get activated with some examples shown in Figure 5.16(b). This includes salient objects, such as the white bowl, the tip of the milk box, the fonts on the oatmeal

box and the bread with butter. Overall, we infer that **DFG-G** not only learns egocentric cues in the spatial domain but also motion dynamics in the temporal domain.

5.12 Application in Gaze-aided Egocentric Activity Recognition

Recent papers have shown that visual attention could help in egocentric activity recognition [219, 9]. To verify our proposed future gaze model is also useful for egocentric activity recognition, we integrate gaze information into the feedforward 3D-CNN for egocentric activity recognition. As [220] shows that 3D-CNN can be used for activity recognition, we adapt the down-scaled framework from [220] (C3D) and integrate the anticipated gaze into the network. A Gaussian mask at the gaze location for each frame, as an additional channel, is concatenated with the input frames of RGB color channels. Cross entropy loss is used for training. Since GTEAplus dataset contains rich instances per activity class as recommended by [219], we follow their evaluation settings and select the top 44 activity classes which have the most instances per class in our recognition task. Confusion matrix of the model with our anticipated gaze is shown in Figure 5.17. In comparison, we also use the same architecture, discard the gaze information and train the network from scratch. In addition, we provide the baseline that the same architecture with the ground truth gaze information as the upper bound. Since center bias is also effective in gaze prediction, we create an artificial baseline where the network with the center gaze is also evaluated. Activity recognition rates are reported in Table 5.9.

From the results, one can observe that our gaze-aided model surpasses C3D network

Table 5.9: Accuracy of Gaze-aided Egocentric Activity Recognition

Models	Activity Recognition Rate
Guess At Random	2.3%
STIP	14.9%
Cuboids	22.7%
C3D	26.9%
C3D + center gaze	13.6%
C3D + DFG-G gaze	28.5%
C3D + our pred gaze	29.3%
C3D + ground truth gaze	33.5%

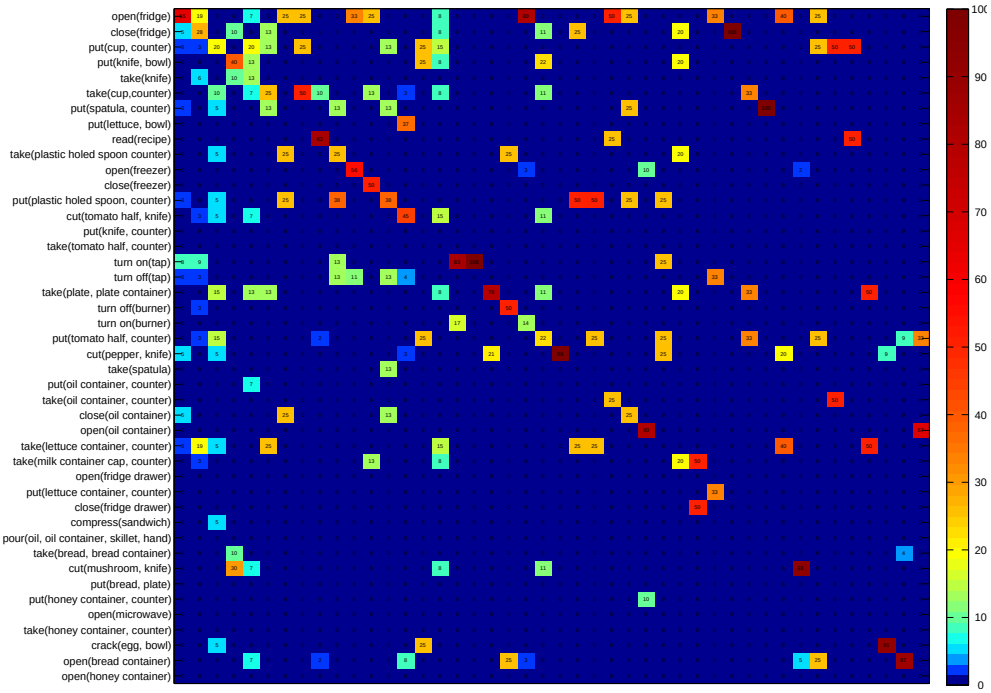


Figure 5.17: Confusion matrix of 44 egocentric activity classes from GTEAplus Dataset.

[220] and several traditional methods [221, 222] and the guess-at-random basline significantly. By comparing the model with our predicted gaze and the one with the center gaze, it can be found that more accurate gaze prediction could result in better egocentric activity recognition. However, the wrong gaze information may be misleading for the network, which may result in poor performances as the baseline uses the center bias.

Chapter 6

Search Network: Modeling Human Visual Search by Top-down Attention

This chapter is based on the paper named "Finding any Waldo with zero-shot invariant and efficient visual search"[1]. Please refer to the same weblink for supplementary materials mentioned in this chapter.

We provide a high-level intuitive outline of our invariant visual search network (IVSN) model. IVSN posits an attention map, M_f, which determines the fixation location by conjugating local visual inputs with target information (Figure 7.3). Both the target image (I_t) and the search image (I_s) are processed through the same deep convolutional neural network, which aims to mimic the transformation of pixel-like inputs through the ventral visual cortex [15, 16, 17, 223]. Feature information from the top level of the visual hierarchy is stored in a module which we refer to as pre-frontal cortex, based on the neurophysiological role of this area during visual search (e.g., [23]). Activity from the pre-frontal cortex module provides top-down modulation, based on the target high-level features, on the responses to the search image, generating the attention map M_f. A winner-take-all mechanism selects the maximum local activity in the attention map M_f for the next fixation. If the fixation location contains the target, the search stops. Otherwise, an inhibition-of-return mechanism leads the model to select

[1]Paper download link: `https://www.nature.com/articles/s41467-018-06217-x`

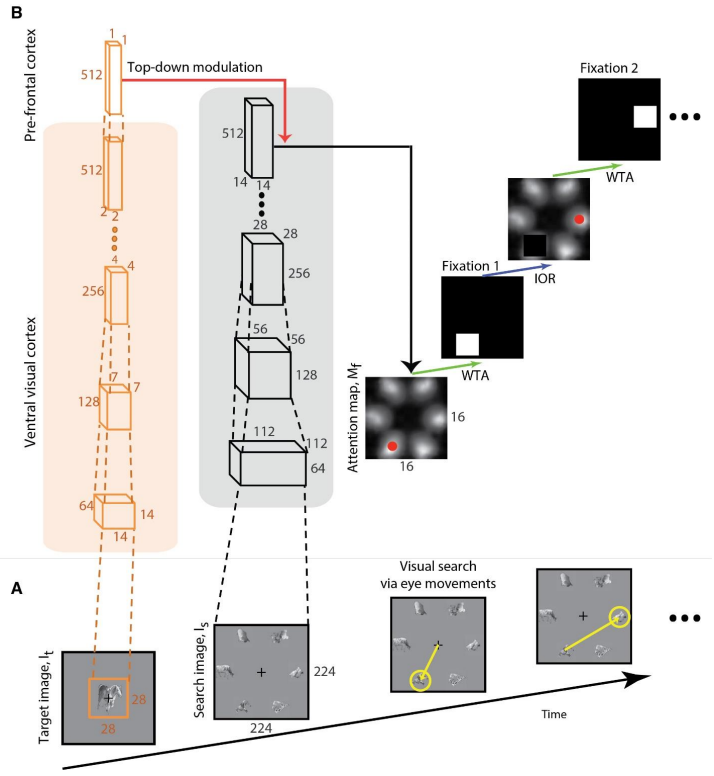


Figure 6.1: Architecture of our proposed invariant visual search network (IVSN) model.

the next maximum in the attention map and the process thus continues until the target object is found. The model was always presented with the exact same images that were shown to the subjects in the psychophysics experiments described in the later section.

6.1 Zero-shot Visual Search Model

In this section, we provide detailed description of each functional module in our IVSN model.

6.1.1 Ventral Visual Cortex

The deep feed-forward network builds upon the basic bottom-up architecture for visual recognition described in previous studies (*e.g.* [15, 16, 17, 18, 223]). We used a state-of-the-art deep feed-forward network, implemented in VGG16 [17], pre-trained for image classification on the 2012 version of the ImageNet dataset [224]. The network weights W learnt from image classification extract feature maps for an input image of size 224×224 pixels. The same set of weights, that is, the same network, is used to process

the target image and the search image. Only a subset of the multiple layers is illustrated in Figure 7.3 for simplicity (see [17] for full details of the VGG16 architecture). The images from the ImageNet dataset used to train the ventral visual cortex network for object classification are different from all the images used in the experiments. The weights W do not depend on any of the target images I_t or the search images I_s (hence the model constitutes a zero-shot training architecture for visual search). The output of the ventral visual cortex module is given by the activations at the top-level (Layer 31 in VGG16), $\phi_{31}(I_t, W)$, and the layer before that (Layer 30 in VGG16), $\phi_{30}(I_s, W)$, in response to the target image and search image respectively. As noted above, it is the same exact network, with the same weights W that processes the target and search images, and we use the activations in layer 31 in response to the target image to provide top-down modulation to layer 30's response to the search image (Figure 7.3).

6.1.2 Pre-frontal Cortex

The top-level of the VGG-16 architecture conveys the target image information to the pre-frontal cortex module, consisting of a vector of size 512. To search for the target object, IVSN uses the ventral visual cortex responses to that target image stored in the pre-frontal cortex to modulate the ventral visual cortex responses to the search image. This modulation is achieved by convolving the representation of the target with the representation of the search image before max-pooling:

$$M_f = m(\phi(I_t, W), \phi(I_s, W)) = m(\phi_{31}(I_t, w), \phi_{30}(I_s, W)) \tag{6.1}$$

where $m(\cdot)$ is the target modulation function defined as a 2D convolution operation with kernel $\phi_{31}(I_t, W)$ on the search feature map $\phi_{30}(I_s, W)$. M_f denotes the attention map.

6.1.3 Fixation Sequence Generation

At any point, the maximum in the attention map determines the location of the next fixation. In the figures, we normalize the attention map to $[0, 1]$ for visualization purposes.

A winner-take-all mechanism selects the fixation location. The model needs to decide whether the target is present at the selected location or not (see below). If the target

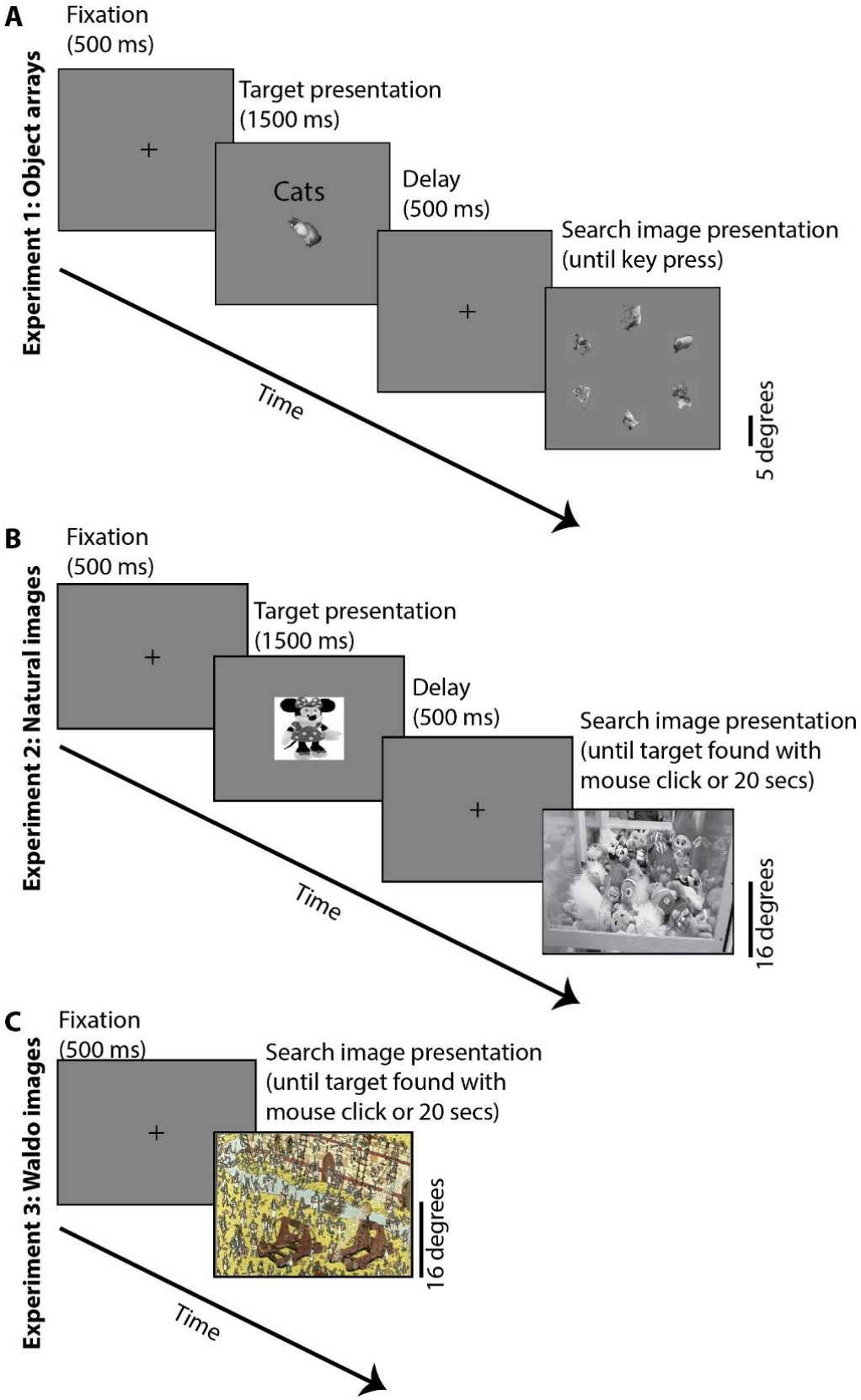


Figure 6.2: Schematic description of the three visual search tasks.

is located, search ends. Otherwise, inhibition-of-return [186] is applied to M_f by reducing the activation to zero in an area of pre-defined size, centered on the current fixation location. This reduction is permanent, in other words, infinite memory is assumed for inhibition of return here. These window size choices were based on the average object sizes in each experiment. Similar to other attention models (*e.g.* [3]), the winner-take-all mechanism then selects the next fixation location and this procedure is iterated until the target is found. In the psychophysics experiments, we limited the duration of each trial to 20 seconds. When we compared the number of fixations at the image-by-image level, we restricted the analyses to those images when the target was found and excluded those images where the target was not found in 20 seconds. Otherwise, all images were included in the analyses.

6.1.4 Target Presence Decision

Given a fixation location, the model needs to perform visual recognition to decide whether the target is present or not (in a similar way that humans need to decide whether they found the target after moving their eyes to a new location). There has been extensive work on visual recognition models (*e.g.* [15, 17, 18, 223]). In this study, we focus on the attention selection mechanism. To isolate the search process from the verification process, in the default IVSN model we bypass the recognition question by using an "oracle" system that decides whether the target is present or not. The oracle checks whether the selected fixation falls within the ground truth location, defined as the bounding box of the target object. The bounding box is defined as the smallest square encompassing all pixels of the object. For fair comparison between models and humans, we implemented the same oracle system for the human psychophysics data, by considering the target to be found the first time a subject fixated on it.

6.2 Experiments on Visual Search

We designed four sets of psychophysics experiments and tested our IVSN model as well as humans in these experiments.

6.2.1 Experiment 1 - Object Arrays

We selected segmented objects without occlusion from 6 categories in the MSCOCO dataset of natural images [225]: sheep, cattle, cats, horses, teddy bears and kites (*e.g.* Figure 6.3A). Due to the uncontrolled and diverse nature of stimuli in the MSCOCO dataset, the images may differ in low-level properties that could contribute to visual search performance. To minimize such contributions, we took the following steps: (1) resized the object areas such that a bounding box of 156×156 pixels encompassed the outermost contour of the object while maintaining their aspect ratios; (2) converted the images to grayscale; (3) equalized their luminance histograms, and (4) randomly rotated the objects in 2D. We conducted a verification test to make sure that the low-level features of all the objects were minimally discriminative: we considered the feature maps from the first convolution blocks of four pre-trained image classification networks (ResNet [226], AlexNet [227], VGG16 and VGG19 [185]), and performed cross-validated category classification tests on these features maps as well as on the image pixels using a Support Vector Machine (SVM) classifier 51. The total of 2000 object images were split into 5 groups for training, validation and testing. The classification performance obtained with these low-level features was consistent across the different computational models and was slightly above chance levels (Supplementary Table 1).

A schematic of the sequence of events during the task is shown in Figure 6.2A. After fixation for 500 ms, a random exemplar from the target category was shown in the fixation location, subtending 5.5 degrees of visual angle, for 1500 ms. The object was shown at a random rotation (0-360 degrees) along with the category name. After another 500 ms of fixation, the search image was presented. Subjects searched for the target in a search image containing an array of 6 objects (Figure 6.3A). In the search images, the 6 objects, each 156×156 pixels and subtending 5 degrees of visual angle, were uniformly distributed on a circle with a radius of 10.5 degrees eccentricity. All the objects could be readily recognized by humans at this size and eccentricity. The target was always present only once within these 6 objects and was placed randomly in one of the 6 possible positions. There was one distractor from each category, randomly chosen.

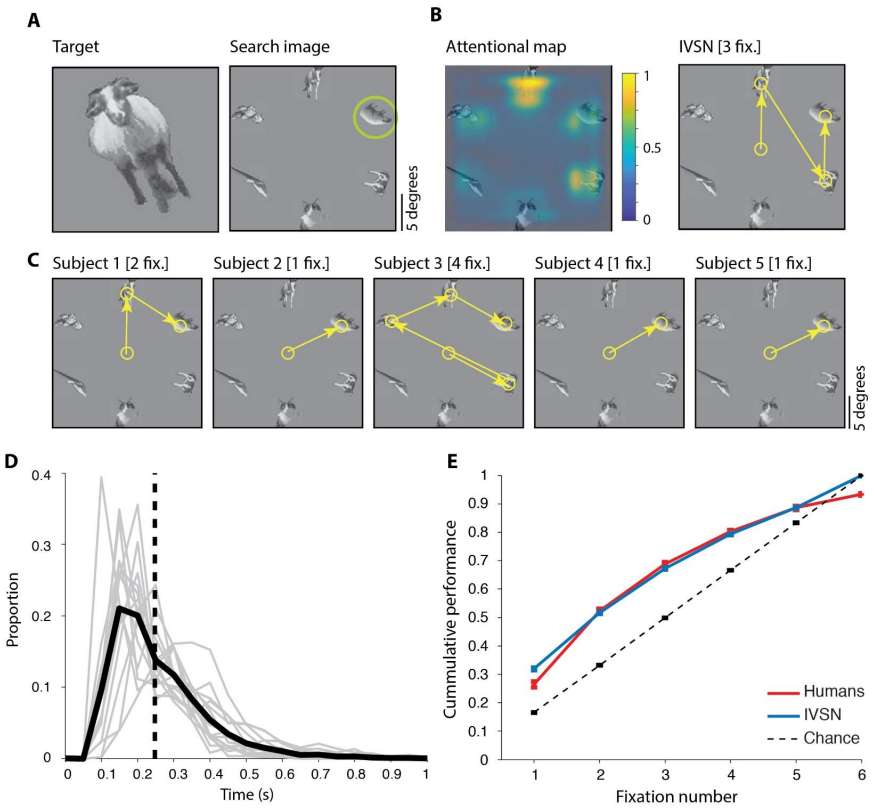


Figure 6.3: Experiment 1 (Object arrays)

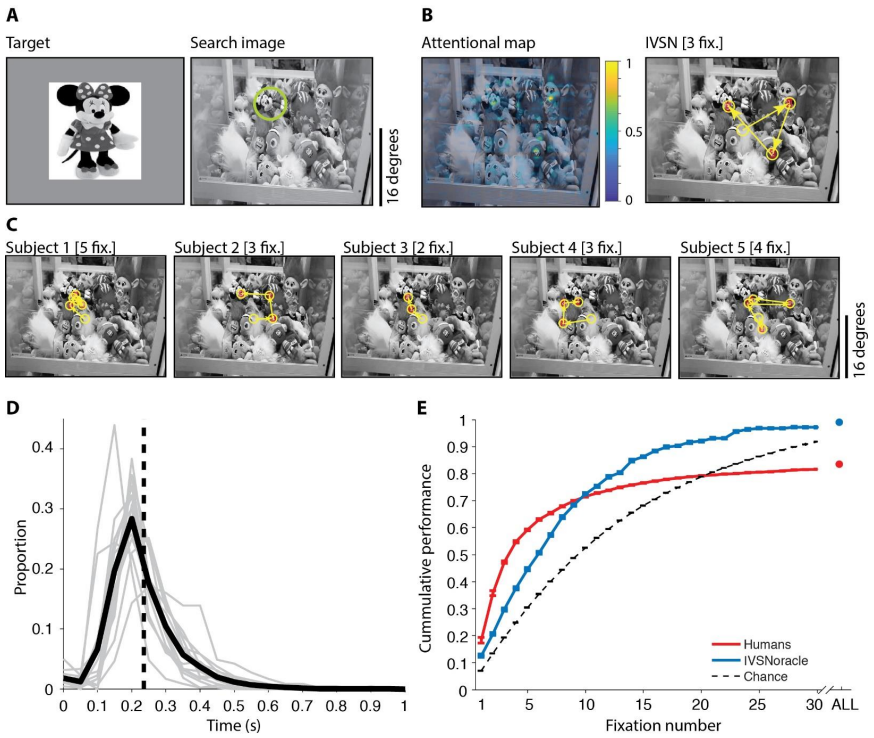


Figure 6.4: Experiment 2 (Natural images)

Subjects were instructed to find the target as soon as possible by moving their eyes and pressed a key to go to the next trial. To evaluate within-subject consistency, and unbeknown to the subjects, each trial was shown twice (the exact same target image and search image was repeated). The order of all trials was randomized. There were $300 \times 2 = 600$ trials in total, divided into 10 blocks of 60 trials each. We split the 300 unique trials into 180 target-different trials and 120 target-identical trials (Supplementary Figure 9A). In the target-identical trials, the appearance of the target object within the search image was identical to that in the target image. In the target-different trials, the target object was a random exemplar from the same category as the one shown in the target image, and was presented at a random rotation (0-360 degrees). Target-different and target-identical trials were randomly interleaved, except in the additional experiment discussed in Supplementary Figure 9D (see below). To evaluate between-subject consistency, the same target and search images were shown to different subjects.

We initially hypothesized that performance would be higher in target-identical trials compared to target-different trials. Upon examining the results, this hypothesis was found to be correct but the difference in performance between target-identical and target-different trials was small (Supplementary Figure 9C). In addition, performance in the target-identical trials was lower than what we reported previously in a different experiment consisting exclusively of target-identical trials and using different objects10. We conjectured that the task instructions and structure including the presence of target-different trials influenced performance in the target-identical trials. To further investigate this possibility, we conducted an additional variation of Experiment 1 in which target-identical and target-different trials were blocked (Supplementary Figure 9D). In this task variation, subjects were told whether the next block would include target-identical or target-different trials. To counter-balance any presentation order biases, we tested 2 subjects on target-identical trials first followed by target-different trials and 3 subjects on the reversed order. This experiment confirmed our intuitions and showed that performance was higher in target-identical trials when they were blocked, compared to when they were interleaved, while performance in target-different trials did not depend on the task structure and instructions. Throughout the text (and except for Supplementary Figure 9D), we focus all the analyses on the original and more nat-

ural version of the task where target-identical and target-different trials were randomly interleaved.

6.2.2 Experiment 2 - Natural Images

We considered 240 objects from common object categories, such as animals (*e.g.* clownfish) and daily objects (*e.g.* alarm clock). The object sizes were 106.5 ± 71.9 pixels high x 114.4 ± 74.8 pixels wide. The 240 objects were not restricted to the 6 categories in Experiment 1 but could involve any object. To test whether IVSN can generalize to searching for novel objects (zero-shot training), we also included objects that are not part of the 2012 ImageNet data set45 (the database of images used to train the model, see Model section below). Examples of such objects include SpongeBob toys, Eve robot, Ironman figures, QuickTime app icon, deformed flags or clothes, weapons, tamarind fruits, fried chicken wings, special hand gesture, Lego blocks, push toys, chopsticks, and ribbons on gifts, among others. There were 140 images out of the selected 240 images containing target objects that were not included in ImageNet. All target objects were manually selected such that each search image contained only one target object. The object shown in the target image was not segmented from the search image, but rather was a similar object: for example, Figure 6.4A shows a vertically and rotated version of Minnie with a dress and bow displaying white circles (left) whereas the target as rendered in the search image shows Minnie at a different scale, with a different attire, partially occluded and under different rotation (right). The search images were 1028×1280 pixel natural images that contained the target amidst multiple distractors and clutter (*e.g.* Figure 6.4A). Both the search images and the target images were presented in grayscale. As illustrated in Figure 6.4A, the target objects were picked such that they were visually different from the ones rendered on the search images; these changes included changes in scale, 2D and 3D rotation, changes in attire, partial occlusion, etc.

The sequence of steps in Experiment 2 followed the one described for Experiment 1 (Figure 6.2B), with three differences described next. The presentation of the target image did not include any text. The search image was a grayscale natural image, always containing the target, and occupied the full monitor screen (subtending 32×40 degrees

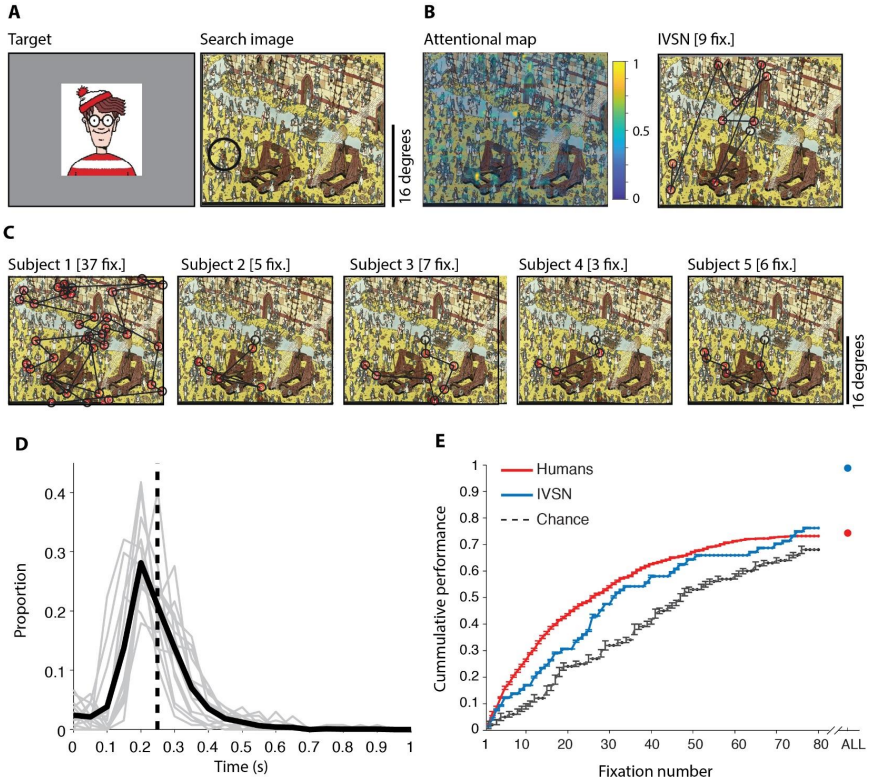


Figure 6.5: Experiment 3 (Waldo images)

of visual angle). The distribution of target object sizes and locations within the search image, which were approximately uniformly distributed. The appearance of the target object within the search array was always different from that in the target image, that is, there were no target-identical trials. Subjects were instructed to find the target as soon as possible by moving their eyes. Experiment 2 was harder than Experiment 1 because objects in the search image were not segmented and were shown embedded in complex natural clutter, and because the appearance of the target object was more different from the target object than in Experiment 1. As the search task became more difficult, subjects could fixate on the target object, yet fail to realize that they had landed on the target (Supplementary Figure 12). Hence, to ensure that subjects had consciously found the target, they had to use the computer mouse to click on the target location. If the clicked location fell within the ground truth, subjects went on to the next trial; otherwise, subjects stayed on the same search image until the target was found. If the subjects could not find the target within 20 seconds, the trial was aborted, and the next trial was presented. Subjects were unable to find the target within 20 seconds in

16.4% of the trials. To evaluate between-subject consistency, different subjects were presented with the same images. To evaluate within-subject consistency, every trial was repeated once, in random order (same target image and same search image). To avoid any potential memory effect (whereby subjects could remember the location of the target), we restricted the analyses to the first presentation, except in the within-subject consistency metrics reported in Figure 6.6, Supplementary Figure 7 and S8. The results were very similar for the first instance of each image versus the second instance of each image and any memory effects across trials were minimal, but we still implemented these precautions focusing the results on the first instance of each image in all the experiments.

6.2.3 Experiment 3 - Waldo Images

"Where's Waldo" is a well-known search task [228] with crowded scene drawings containing hundreds of individuals that look similar to Waldo undertaking various activities. Exactly one of these individuals is the character known as Waldo (*e.g.* Figure 6.5A). We tested 67 Waldo images from [228]. The target object sizes were 24.7 ± 4.5 pixels wide and 40.3×7.4 pixels high. Given the large size of the Waldo search images and the limited precision of our eye tracker in terms of individual characters on these images, we cropped each Waldo image into four quadrants and only showed the human subjects the quadrant containing Waldo. There were 13 out of 67 images that had an instruction panel in the upper left corner that could contain additional renderings of Waldo. Subjects were explicitly instructed not to look at the instruction panel. At the model evaluation stage, these areas were also discarded. The locations of these panels can be approximately glimpsed from less dense fixation patches in Supplementary Figure 1H. Because all subjects were familiar with the Waldo task, we changed the overall structure such that there was no target image presentation in each trial (Figure 6.2C). The target (Waldo) in color was presented at the beginning of the experiment. After fixation, the search image, always containing Waldo, was presented occupying the full monitor screen (subtending 32×40 degrees of visual angle). Subjects were instructed to find Waldo as soon as possible by moving their eyes. Similar to Experiment 2, once the target was found, subjects had to click on the target location. If the clicked location fell

on the ground truth, subjects proceeded to the next trial; otherwise, subjects stayed on the same search image until the target was found. If subjects could not find the target in 20 seconds, the trial was aborted. The limit of 20 seconds was based on pilot tests and was dictated by a compromise between allowing enough time to find the target in as many trials as possible while at the same time maximizing the number of search trials. Subjects were unable to find the target within 20 seconds in 27% of the trials. There were 67 trials in total and the trial order was randomized. Within- and between-subject consistency was evaluated as described above for Experiments 1 and 2. In addition to searching for Waldo, we conducted a separate set of trials where subjects searched for the "Wizard", another character in the Waldo series. The results for the Wizard search were similar to those for the Waldo search. We restrict this report to the Waldo search task for simplicity.

6.2.4 Experiment 4 - Novel Objects

We conducted an additional experiment to evaluate whether human subjects are able to search for novel objects that they have never encountered before (other than the single exposure to the target image). We collected a total of 1860 novel objects belonging to 98 categories. These objects were composed from well-designed novel object parts and we also included novel objects used in previous studies (Supplementary Figure 10) [229, 230]. We used the same pre-processing steps to normalize the novel objects' low-level features as in Experiment 1. Supplementary Figure 10A shows 6 example novel objects. The task structure followed the one in Experiment 1, except that here there was no text indicating the object category during the target presentation (Supplementary Figure 10B). The number of trials for target identical and target different trials was balanced (80 target-identical vs. 80 target-different trials in novel objects). To directly compare the results for novel objects versus those obtained with known objects, the objects from Experiment 1 (known objects) were also presented in this experiment, randomly intermixed with the novel object trials.

In visual search experiments, the similarity between the target object and the distractor objects plays a critical role in the difficulty of the task. As a proxy for task difficulty, we computed the similarity between the target object and the distractors by computing

the Euclidian distance between all possible target-distractor object pairs in each image (x-axis in Supplementary Figure 10C). The target and distractor novel objects were chosen so as to match the distribution of similarities for known objects (Supplementary Figure 10C) to avoid scenarios where one set of stimuli could be easier to discriminate than in the other set. The results for the novel object visual search experiment are shown in Supplementary Figures 10D-E.

6.2.5 Human Participants

We conducted four psychophysics experiments with 60 naive observers (19-37 years old, 35 females, 15 subjects per experiment). The sample size was chosen based on the results in one of our previous experiments [19]. In Experiment 1, we used a sample size that was effective in a previous study with a similar structure [19]. For Experiments 2 and 3, we used the same sample size to facilitate comparisons across experiments. We focus on the first 3 experiments in the main text and report the results of the fourth experiment in Supplementary Figure 10. All participants had normal or corrected-to-normal vision. Participants provided written informed consent and received 15 USD per hour for participation in the experiments, which typically took an hour and a half to complete. All the psychophysics experiments were conducted with the subjects' informed consent and according to the protocols approved by the Institutional Review Board at Children's Hospital.

6.2.6 Experimental Protocol

The general structure for all three experiments was similar (Figure 6.2). Subjects had to fixate on a cross shown in the middle of the screen, a target object was presented followed by another fixation delay (Experiments 1 and 2), a search image was presented, and subjects had to move their eyes to find the target. In Experiments 2 and 3, subjects also had to indicate the target location via a mouse click. Stimulus presentation was controlled by custom code written in MATLAB using Version 3.0 of the Psychophysics Toolbox. Images were presented on a 19-inch CRT monitor (Sony Multiscan G520), at a 1024×1280 pixel resolution, subtending approximately 32×40 degrees of visual angle. Observers were seated at a viewing distance of approximately 52 cm. We recorded the

participants' eye movements using the EyeLink D1000 system (SR Research, Canada).

6.2.7 Psychophysics Fixation Analysis

We used the EDF2Mat function provided by the EyeLink software (SR Research, Canada) to automatically extract fixations. We clustered consecutive fixations that were within object bounding boxes of size 45×45 pixels for more than 50 ms. If fixation was not detected during the initial fixation window, the experimenter re-calibrated the eye tracker. The last trial before re-calibration and the first trial after calibration were excluded from analyses. In Experiment 1, we filtered out fixations falling outside the six object locations ($13.7 \pm 5.6\%$ of the trials). Upon presentation of the search image, we considered the first fixation away from the center. We considered that a fixation had landed on the target object if it was within a square window centered on the target object. The window sizes were 45x45 for Experiment 1, 200×200 pixels for Experiment 2 and 100×100 pixels for Experiment 3. These values correspond to the mean widths and heights of all the ground truth bounding boxes for each dataset (Supplementary Figure 1). In Experiments 2 and 3, subjects had to click the target location with the mouse. The mouse click location had to fall on the window defining the target object location for the trial to be deemed successful. In $15.9 \pm 4.9\%$ of trials in Experiment 2 and $10.1 \pm 7.0\%$ of trials in Experiment 3, the initial mouse clicks were incorrect. If the location indicated by the mouse click was incorrect, subjects had to continue searching; otherwise, the trial was terminated. It should be noted that in several cases, subjects could fixate on the target object but not click the mouse, most likely because they were not consciously aware of finding the target despite the correct fixation (Supplementary Figure 12, see Discussion). For fair comparison with the models, we used an oracle version such that the target was considered to be found upon the first fixation on the target, except in Supplementary Figure 12.

6.2.8 Comparisons of Fixation Patterns

We evaluated the degree of within-subject consistency by comparing the fixations that subjects made during the first versus second presentation of a given target image and search image. We evaluated the degree of between-subject consistency by performing

pairwise comparisons of the fixations that subjects made in response to the same target image and search image for all 15-choose-2 subject pairs. We compared the fixations of the IVSN model against each of the 15 subjects. We used the following metrics to compare fixations within subjects, between subjects and between subjects and the IVSN model: (1) we considered the cumulative accuracy as a function of the number of fixations to evaluate the overall search performance (Figures 6.3E, 6.4E, 6.5E); (2) we compared the number of fixations required to find the target on an image-by-image basis (Supplementary Figure 7); (3) we compared the spatiotemporal sequence of fixations on an image-by-image basis (Figure 6.6, Supplementary Figure 8).

Cumulative performance We compute the probability distribution $p(n)$ that the subject or model finds the target in n fixations. Figures 6.3E, 6.4E, 6.5E show the cumulative distribution of $p(n)$.

Number of fixations to find the target For each image, we plot the number of fixations required to find the target for S1 and S2 where S1 and S2 can be different repetitions of the same image (within-trial consistency), different subjects (between-trial consistency), or subject and model (model-subject consistency). This metric is reported in Supplementary Figure 7.

Spatiotemporal dynamics of fixations on an image-by-image basis We used the scanpath similarity score proposed by [33]. This measure takes into account both spatial and sequential order by aligning the scanpath between two sequences. We used the implementation described in [231]. Briefly, a mean-shift clustering for all human fixations was computed, and a unique character was assigned to each cluster center and corresponding fixations. The Needleman-Wunsch string match algorithm [196] was implemented to evaluate the similarity of a scanpath pair. In Supplementary Figure 8, we compare the entire sequences. In Figure 6.6, we compare the first x fixations as shown in the x-axis in the figure.

6.2.9 Comparison with Other Models

We performed several comparisons with other models (Supplementary Figures 4, 11, 13, 14). In all cases, the alternative models proposed a series of fixations. In all cases except for $IVSN_{recognition}$ (described below), we used the oracle method to decide whether to stop search or to move on to the next fixation. In all cases except for $IVSN_{fIOR}$ (described below), the models had infinite inhibition of return (IOR), as described above. We considered the following alternative models:

Chance We considered a model where the location of each fixation was chosen at random. In Experiment 1, we randomly chose one out of the six possible locations, while still respecting infinite IOR. In Experiments 2 and 3, a random location was selected in each fixation, while still respecting IOR; this random process was repeated 100 times. The selected location was the center of a window of the same size used for the recognition model described above. This window was used to determine the presence of the target and also to set IOR.

Sliding Window (SW) We considered a sliding window approach which takes the fixated area (a window of the same size used for the recognition model described above) as inputs, scans the search image from the top left corner with stride 28 pixels, and uses oracle verification to determine target presence. In Experiment 1, the sliding window sequentially moves through the 6 possible objects.

Template Matching To evaluate whether pixel-level features of the target were sufficient to direct attention, we introduced a pixel-level template-matching model where the attention map was generated by sliding the canonical target of size 28×28 pixels over the whole search image. Compared with the SW model, the Template Matching model can be thought of as an attention sliding window.

IttiKoch It is conceivable that in some cases, attention selection could be purely driven by bottom-up saliency effects rather than target-specific top-down attention modulation. We considered a pure bottom-up saliency model that has no information about the target [3].

RanWeight Instead of using VGG16, pre-trained for image classification, we randomly picked weights W from a Gaussian distribution with mean 0 and standard deviation 1000. The network was otherwise identical to IVSN. We ran 30 iterations of this model, each iteration with random selection of weights.

6.2.10 Extensions and Variations of Visual Search Model

We considered several possible extensions and variations of the IVSN model.

$IVSN_{AlexNet}$ **(Supplementary Figure 14)** The "ventral visual cortex" module in Figure 7.3 was replaced by the AlexNet architecture [227]. The "pre-frontal cortex" module corresponded to layer 8 and sent top-down signals to layer 7.

$IVSN_{ResNet}$ **(Supplementary Figure 14)** The "ventral visual cortex" module in Figure 7.3 was replaced by the ResNet200 architecture [226]. The "pre-frontal cortex" module corresponded to the output of residual block 8 in the target image and sent top-down signals to residual block 8 in the search image.

$IVSN_{FastRCNN}$ **(Supplementary Figure 14)** The "ventral visual cortex" module in Figure 7.3 was replaced by the FastRCNN architecture [28] pre-trained on ImageNet for region proposal and pre-trained on PASCAL VOC for object detection. The "pre-frontal cortex" module corresponded to layer 24 and sent top-down signals to layer 23.

$IVSN_{24\rightarrow23}, IVSN_{17\rightarrow16}, IVSN_{10\rightarrow9}, IVSN_{5\rightarrow4}$**(Supplementary Figure 13)** In the IVSN model as presented in Figure 7.3 (based on the VGG16 architecture3), the "pre-frontal cortex" module corresponded to layer 31 and sent top-down signals to layer 30. We considered several variations using top-down features from different levels of the VGG16 architecture as described by the model sub indices.

$IVSN_{recognition}$ **(Supplementary Figure 11A-C)** The IVSN model presented in the main text uses an oracle to determine whether the target was found at a given fixation or not. In the brain, of course, there is no oracle. Each fixation places the new location within the high-resolution fovea, and responses along the ventral visual stream

within this region are enhanced via attention modulation [23, 35, 37]. By emphasizing the selected areas, IVSN allows the ventral pathway to perform fine-grained object recognition. As a schematic proof-of-principle of a model that addresses whether the target was found or not, in Supplementary Figure 11A-C we implemented an additional step that included recognition after fixation. This recognition machinery involved an object classifier which determined whether the fixated area contained the target or not (IVSNrecognition). We implemented this step by cropping the search image centered at the fixation location using the same window sizes described for inhibition of return (45×45, 200×200, and 100×100, for Experiments 1, 2, and 3, respectively), and using the object recognition network, VGG16 [185], pre-trained on ImageNet [224], to extract the classification vector from the last layer, which emulates responses in inferior temporal cortex with high object selectivity and large receptive fields, for both the target image It and the cropped area. The Euclidean distance between activation of this top layer to It and the cropped area was computed. If this Euclidian distance was below a threshold of 0.9, the target was deemed to be found and search was stopped. Otherwise, the search continued after applying inhibition-of-return, as described above for the oracle. In this model including a recognition component, failure to locate the target could be due to fixating on the wrong location or fixating on the right location but not realizing that the target was there.

$IVSN_{fIOR}$ The IVSN model assumes infinite inhibition-of-return, that is the model never revisits a given fixation location. In contrast, humans do tend to revisit the same location even if the target is not there. An example of this behavior can be seen in multiple fixations from subject 1 in Supplementary Figure 5C and also in fixations 3 and 6 in Supplementary Figure 7B2 (the reader may have to zoom in on the figures to appreciate this phenomenon). The finite inhibition of return is a well known phenomenon in the psychophysics literature42,47,48. We implemented a variation of the IVSN model with finite inhibition-of-return ($IVSN_{fIOR}$). At each location in the image (x, y) and at time t, the feature attention map M_f was multiplied by a memory function M_m to generate a new attention map $A_f(x, y) = M_f(x, y) * M_m m(x, y, t)$. In the implementation with infinite IOR, $M_m(x, y, t)$ is 0 if the location (x, y) was visited previously

and 1 otherwise (independently of time t). In the $IVSN_{fIOR}$ model, $M_m(x, y, t)$ was fitted to the empirical probability of revisiting a location from the human psychophysics data. The inaccuracy in our eye movement measurements is on the order of 1 degree of visual angle. To be overly cautious, we defined a location as revisited if another fixation landed within 3 degrees of visual angle. None of the parameters in the default IVSN model were trained or fitted to human psychophysics data. In contrast, the function M_m was fitted to the human psychophysics data, separately for each experiment. To avoid overfitting, we randomly selected 7 out of the 15 subjects to fit M_m and all the comparisons between $IVSN_{fIOR}$ and human psychophysics was based on the remaining 8 subjects.

$IVSN_{size}$ The IVSN model has no constrain on the size of each saccade (*e.g.* one fixation could be in the upper left corner and the immediate next fixation could be in the lower right corner). In contrast, humans tend to make smaller saccades following a gamma-like distribution (Supplementary Figure 11G-I). We implemented a variation of the IVSN model where the saccade size was constrained by the empirical distribution of human saccade sizes ($IVSN_{size}$). We defined the attention map as a weighted sum of the feature attention map M_f and a size constraint function M_{sc} : $A_f(x, y) = wM_f(x, y) + (1 - w)M_{sc}(x, y)$. The weight factor w was set to 0.2346 across all the experiments, selected to optimize the fit between human and IVSNsize saccade sizes. In a similar fashion to $IVSN_{size}$ and to avoid overfitting with did cross-validation by fitting M_{sc} separately for each experiment, using only a random subset of 7 out of the 15 subjects.

6.3 Consistency between Human and Model Search Performance

We considered the problem of localizing a target object that could appear at any location in a cluttered scene under a variety of shapes, scales, rotations and other transformations. We conducted 3 increasingly more difficult visual search experiments where 45 subjects had to move their eyes to find the target (Figure 6.2). We propose a biologically inspired computational model to account for the fixations during visual search (Figure

7.3).

6.3.1 Searching for A Target within An Array of Objects

Many visual search studies have focused on images with isolated objects presented on a uniform background such as the ones in Experiment 1 (Figure 6.2A, 6.3A). We used segmented grayscale objects from 6 categories from the MSCOCO dataset [225] (Methods). After fixation, 15 subjects were presented with an image containing a word describing the object category and a target object cue at a random 2D rotation (Figure 6.2A). After an additional fixation delay, a search image was introduced, containing a different rendering of the target object, randomly located in one of 6 positions within a circle, along with 5 distractors from the other categories. The target was always present and appeared only once. The rendering of the target in the search image was different from the one in the target cue (*e.g.* Figure 6.3A): it was a different exemplar from the same category, and it was shown at a different random 2D rotation. Subjects were instructed to rapidly move their eyes to find the target. Example fixation sequences from 5 subjects are shown in Figure 6.3C: in these examples, subjects found the target in 1 to 4 fixations, despite the fact that the rendering of the target in the search image involved a different sheep, shown at a different 2D rotation. The target locations were uniformly distributed over the six possible positions (Supplementary Figure 1A) and subjects did not show any appreciable location biases (Supplementary Figure 1B). Subjects made their first fixation at 287 ± 152 ms ($mean \pm SD$, $n = 15$ subjects, Figure 6.3D). The interval between fixations was 338 ± 203 ms (Supplementary Figure 2A). The rapid deployment of eye movements is consistent with previous studies [19], and shows that subjects followed the instructions, without adopting alternative strategies such as holding fixation in the center and searching for the target purely via covert attention (Discussion).

Subjects located the target in 2.60 ± 0.22 fixations ($mean \pm SD$, Figure 6.3E), corresponding to 640 ± 498 ms ($mean \pm SD$, Supplementary Figure 2B). The number of fixations required to find the target was significantly below the number expected from random exploration of the 6 possible locations, which would require 3.5 fixations in this experiment (Figure 6.3E, $p < 10^{-15}$, two-tailed t-test, t=10, df=4473). Even

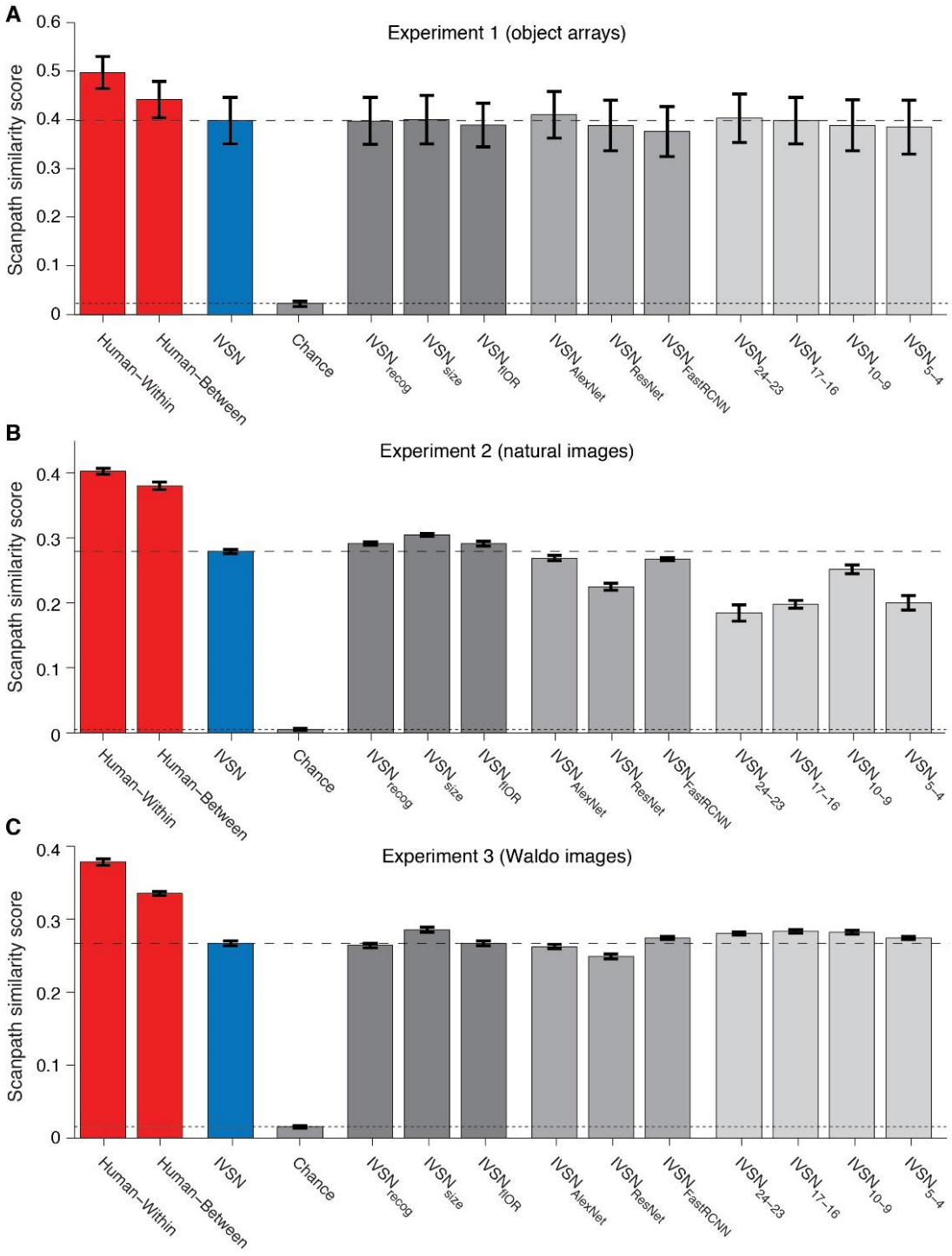


Figure 6.6: Image-by-image consistency in the spatiotemporal pattern of fixation sequences

in the first fixation, subjects were already better than expected by chance (performance = $26.4 \pm 4.1\%$ versus 16.7%). At 6 fixations, the cumulative performance was below 100% ($93.3 \pm 1.6\%$), since subjects revisited the same locations, even when they were wrong. The number of fixations required to find the target was lower when the target was identical in the target and search images (Supplementary Figure 3A-B), yet subjects were able to efficiently and robustly locate the target despite changes in 2D rotation (Supplementary Figure 3B) and despite the exemplar differences (Supplementary Figure 3A).

To better understand the guidance mechanisms that incorporate target shape information to dictate the sequence of fixations, we implemented a computational model inspired by neurophysiological recordings in macaque monkeys during visual search tasks. The Invariance Visual Search Network (IVSN) model consists of a deep feed-forward network that mimics processing of features along ventral visual cortex, a way of temporarily storing information about the target tentatively associated with pre-frontal cortex, modulation of visual features in a top-down fashion to generate an attention map, and sequential selection of fixation locations (Figure 7.3B, Methods). Of note, IVSN was neither trained with any of the images used in this study, nor was it trained in any way to match human performance. The same images used for the psychophysics experiments were presented to the model. The model builds an attention map (Figure 6.3B, left) in response to the target and search images from Figure 6.3A, and uses this map to generate a sequence of fixations, locating the target in 3 fixations (Figure 6.3B, right). Despite the lack of training with this image set, and the large degree of heterogeneity between the target cue and the targets appearance in the search image, IVSN was able to efficiently locate the targets in 2.80 ± 1.71 fixations across all the trials (Figure 6.3E, blue). IVSN performed well above the null chance model ($p < 10^{-11}$, two-tailed t-test, t=7.1, df=598), even in the first fixation (performance = $31.6 \pm 0.5\%$ compared to chance = 16.7%). The model had infinite inhibition-of-return and therefore never revisited the same location, by construction thus achieving 100% performance at 6 fixations (see Supplementary Figure 11 and Discussion). Although there were no free parameters tuned to match human behavior, IVSN's performance was similar to that of humans. The strong resemblance between IVSN and human performance shown in

Figure 6.3E should not be over interpreted: there was still a small difference between the two (p=0.03, two-tailed t-test t=2.2, df=4473); in addition, we will discuss below other differences between humans and the IVSN model. Similar to human behavior, the model required fewer fixations when the rotation of the target cue matched the one in the search image, but the model was also able to efficiently locate the target at all the rotations tested (Supplementary Figure 3A-B).

We considered several alternative null models to further understand the image features that guide visual search (Supplementary Figure 4A). In the sliding window model, commonly used in computer vision, a fixed-size window sequentially scans the image (here scanning was restricted to the 6 locations), which is equivalent to random search with infinite inhibition of return in this case, and fails to explain human behavior. Visual search was not driven by pure bottom-up saliency features as represented by the Itti and Koch model [3]. The weight features in the ventral visual cortex part of the model are important to generate the shape-invariant target-dependent visual attention map, as demonstrated by two observations: (i) randomizing those weights led to chance performance (RanWeight model); (ii) template matching algorithms based on pixels, using rotated templates or not, which are poor at invariant visual object recognition, were insufficient to explain human behavior (Template Matching model). In sum, both humans and IVSN significantly outperformed all the alternative null models.

6.3.2 Searching for A Target in Natural Scenes

The object array images used in Experiment 1 lack critical components of real world visual search. In natural scenes, there is no fixed type and number of distractors equidistantly arranged in a circle, the target object is not segmented nor is it generally present on a uniform background, and the appearance of the target object can vary along multiple dimensions that are not pre-specified. In Experiment 2, we directly tackled visual search in natural images (Figure 6.4). The structure of the task was essentially the same as that in Experiment 1 (Figure 6.2B) with the following differences: (i) search images involved natural images (*e.g*. Figure 6.4A), (ii) objects and distractors were not restricted to 6 categories, (iii) the appearance of the target object in the target and search images could vary along multiple dimensions, (iv) a trial was ended if the target was

not found within 20 seconds, and (v) to ensure that the target was correctly found, subjects had to use the computer mouse to indicate the target location (see Section 6.1). The target locations were randomly and uniformly distributed (Supplementary Figure D). Subjects made rapid fixation sequences throughout the entire search image, with certain biases such as a larger density of fixations in the center and a smaller density of fixations along the borders (Supplementary Figure 1E). Figure 6.4C shows example sequences where subjects were able to rapidly find the target in 2 to 5 fixations despite the changes in target appearance and despite the large amount of image clutter. The first fixation occurred at 285 ± 135 ms (Figure 6.4D), and the interval between fixations was 290 ± 197 ms (Supplementary Figure 2C). The last fixations became progressively closer to the target (Supplementary Figure 2H). Subjects found the target in 1867 ± 2551 ms (Supplementary Figure 2D), which was about three times as long as in Experiment 1 (Supplementary Figure 2B).

Subjects located the target in 6.2 ± 0.7 fixations (Figure 6.4E, red). Performance saturated at 15 fixations, well below 100%. In $16.4 \pm 5.9\%$ of the images, subjects were unable to find the target within 20 seconds, hence human performance was well below ceiling. Human performance was more efficient than the chance model ($p < 10^{-15}$, two-tailed t-test, t=14, df=3247). Subjects tended to revisit the same locations even though the target was not there. In part because of this behavior, the null chance model showed a higher cumulative performance after 20 fixations. The average number of fixations that humans required to find the target was below that expected from the null chance model. Even in the first fixation, subjects were better than expected by chance (performance = $18.3 \pm 3.8\%$ versus $7.0 \pm 0.2\%$). The target as rendered in the search image could be larger or smaller than the target cue. Intuitively, it could be expected that performance might monotonically increase with the target size in the search image. However, subjects performed slightly better when the size of the target in the search image was similar to the original size in the target cue. Subjects were still able to robustly find the target across large changes in size (Supplementary Figure 3D). In addition to size changes, the target's appearance in the search image was generally different in many other ways, which we quantified by computing the normalized Euclidian distance between the target cue and the target image in the search image. Subjects robustly found

the target despite large changes in its appearance (Supplementary Figure 3C).

Next, we investigated the performance of IVSN in natural images. Importantly, we used exactly the same model described for Experiment 1, with no additional tuning or any free parameters adjusted for Experiment 2. IVSN generated the attention map and scanpath in Figure 4B in response to the target and search images from Figure 6.4A: the model located the target in 3 fixations even though it had never encountered this target or any similar target before, despite the large amount of clutter, and despite the visual appearance changes in the target. IVSN efficiently located the target in natural scenes, requiring 8.3 ± 7.5 fixations on average (Figure 6.4E, blue). IVSN performed well above the null chance model ($p < 10^{-15}$, two-tailed t-test, t=8.5, df=478), even in the first fixation ($14 \pm 5\%$ versus $7.0 \pm 0.2\%$). IVSN had infinite inhibition-of-return, never revisiting the same location, and achieving 100% accuracy in about 45 fixations. Humans outperformed the model up to approximately fixation number 10, but the model performed better than humans thereafter. Consistent with human behavior, IVSN was also robust to large differences between the size of the target in the search image and target cue (Supplementary Figure 3D) and it was also robust to other changes in target object appearance (Supplementary Figure 3C).

As described in Experiment 1, we considered several alternative null models, all of which were found to show lower performance than humans and IVSN (Supplementary Figure 4B). A pure bottom-up saliency model was worse than chance levels, because it did not incorporate features relevant to the target and instead concentrated on regions of high contrast in the image that were not relevant to the task. Similarly, template matching models were also worse than chance because they generated attention maps that emphasized regions that showed high pixel-level similarity to the target without incorporating invariance and therefore failing to account for the transformations in the target object shape present in the search image.

6.3.3 Searching for Waldo

The IVSN model could find objects that it had never encountered before (see also Supplementary Discussion and Supplementary Figure 5). To further investigate invariant visual search for novel objects, we designed Experiment 3 to test IVSN with more ex-

treme images that bear no resemblance to those used in Experiments 1 and 2, or to the images in the ImageNet data set. We considered the traditional "Where is Waldo" task41 (Figure 6.5), comprising colorful cluttered drawings with scene statistics that are very different from those in natural images. The structure of Experiment 3 was similar to that of Experiment 2, except that a picture of Waldo was only presented at the beginning of the experiment and not in every trial (Figure 6.2C). The target locations were randomly and uniformly distributed (Supplementary Figure 1G). Subjects made fixations throughout the entire search image, with certain biases such as a higher density in the center and a smaller density of fixations along the borders (Supplementary Figure 1H). Subjects made rapid sequences of fixations (*e.g.*, Figure 6.5C), with the first fixation occurring at 264 ± 112 ms (Figure 6.5D), and an interval between fixations of 278 ± 214 ms (Supplementary Figure 2E). On average, subjects progressively became closer to the target in their last fixations (Supplementary Figure 2I).

Searching for Waldo constitutes a difficult challenge for humans, as confirmed by our results. On average, subjects found the target in 21.1 ± 3.1 fixations corresponding to 6051 ± 4962 ms (Figure 6.5E, Supplementary Figure 2F), about three times longer than in Experiment 2 and about nine times longer than in Experiment 1. Performance reached a plateau at about 60 fixations, well below 100%. In $26.9 \pm 9.6\%$ of the images, subjects were unable to find the target within the allocated 20 seconds. Despite the task difficulty and despite infinite inhibition of return in the null chance model, subjects were able to find Waldo more efficiently than by random exploration ($p < 10^{-15}$, two-tailed t-test, t=18, df=800). There were also differences between the rendering of the target object in the search image and target image. Subjects were able to find Waldo despite these changes in target appearance (Supplementary Figure 3E).

We evaluated IVSN responses on the images from Experiment 3, without fine-tuning any parameters. IVSN had no prior experience with Waldo images or drawings of any kind. In the example in Figure 6.5A-B, the model located Waldo in 9 fixations. IVSN efficiently located Waldo, requiring 29.0 ± 21.6 fixations on average (Figure 6.5E, blue). IVSN performed well above the null chance model ($p < 10^{-15}$, two-tailed t-test, t=10, df=116). Despite the task difficulty, humans were more efficient in finding Waldo than IVSN (p=0.001, two-tailed t-test, t=3.3, df=784). IVSN was robust to changes in

the appearance of the target (Supplementary Figure 3E). The alternative null models did not perform as well as humans or the IVSN model (Supplementary Figure 4C).

Waldo was completely novel to IVSN but not for humans. We conducted a separate experiment with objects that were completely novel for humans and showed that subjects were still able to find targets under situations where they had no prior exposure to the target objects (Supplementary Figure 10, Supplementary Discussion).

6.4 Comparisons between Human and Model Search at Image Levels

The results presented thus far compared average performance between humans and models considering all images. We next examined consistency in the responses at the image-by-image level within-subjects (identical trials presented to the same subject), between-subjects, and between IVSN and subjects. We compared the number of fixations required to find the target in each trial in Supplementary Figure 7. Subjects were slightly more consistent with themselves than with other subjects, and the between-subject consistency was slightly higher than the consistency with IVSN (Supplementary Discussion).

The number of fixations provides a summary of the efficacy of visual search but does not capture the detailed spatiotemporal sequence of eye movements (Supplementary Figures 6, 8). We used the scanpath similarity score [33], to compare two fixation sequences (Supplementary Discussion). This metric captures the spatial and temporal distance between two saccade sequences, ranging from 0 (maximally different) to 1 (identical). Within-subject comparisons yielded slightly more similar sequences than between-subject comparisons in all 3 experiments (Figure 6.6, $p < 10^{-9}$). The between-subject scanpath similarity scores, in turn, were higher than the IVSN-human similarity scores for all 3 experiments. The IVSN-human similarity scores were higher than the human-chance similarity scores for all 3 experiments. In sum, IVSN captured human eye movement behavior at the image-by-image level in terms of the number of fixations and the spatiotemporal pattern of fixations.

6.5 Variational IVSN Computational Model Performance

We next considered variations of the IVSN model architecture and revisited several simplifications and assumptions of the model. The results presented thus far assumed that the model can perfectly recognize whether the target is present or not at the fixated location. After each fixation, an "oracle" decides whether the target is present or not. Rapidly recognizing whether the target is present or not is not easy, particularly in Experiments 2 and 3. Subjects sometimes fixated on the target, yet failed to recognize it, and continued the search process (Supplementary Figure 12A-B). Examples of this behavior are illustrated for Subjects 1 and 5 in Figure 6.4C where the second fixations land on the target, yet the subjects make additional saccades and subsequently return to the target location. For fair comparison, all the psychophysics results presented thus far also used an oracle for recognition (search was deemed successful the first time that a fixation landed on the target). Without the oracle, human performance was lower but still well above chance (Experiment 2: $p < 10^{-15}$, t=14, df=3247, Supplementary Figure 12C; Experiment 3: $p < 10^{-15}$, t=18, df=800, Supplementary Figure 12D). We introduced a simple recognition component into the model to detect whether the target was present or not based on the features of the object at the fixated location ($IVSN_{recognition}$, Supplementary Figure 11A-C, Methods). $IVSN_{recognition}$ performed slightly but not significantly below IVSN, particularly in the more challenging case of Experiment 2. $IVSN_{recognition}$ was still able to find the target above chance levels (Experiment 1: $p < 10^{-11}$, t=7.3, df=594, Supplementary Figure 11A; Experiment 2: $p < 10^{-13}$, t=8, df=434, Supplementary Figure 11B; Experiment 3: $p < 10^{-15}$, t=12, df=112, Supplementary Figure 11C).

Another simplification involved endowing IVSN with infinite inhibition of return. In contrast, humans show a finite memory and tend to revisit the same locations not only for the target (Supplementary Figure 12C-D) but also for non-target locations (e.g. subject 1 in Figure 6.5C) [232]. We fitted an empirical function to describe the probability that subjects would revisit a location at fixation i given that they had visited the same location at fixation $j < i$ [232]. We incorporated this empirical function into the IVSN model so that previous fixated locations could be probabilistically revisited, thus creating a model with finite inhibition of return ($IVSN_{fIOR}$, Methods, Supple-

mentary Figure 11D-F). The $IVSN_{fIOR}$ model showed lower performance than the IVSN model but this difference was not significant or marginally significant (Experiment 1: p=0.11; Experiment 2: p=0.02; Experiment 3: p=0.07). Despite this drop in performance, $IVSN_{fIOR}$ was still able to find the target better than chance (Experiment 1: p=10^{-15}, t=9.7, df=864; Experiment 2: $p < 10^{-15}$, t=11, df=617; Experiment 3: p=10^{-15}, t=16, df=145, two-tailed t-tests). Furthermore, $IVSN_{fIOR}$'s performance was closer to humans for all 3 experiments (Supplementary Figure 11D-F, $IVSN_{fIOR}$ versus human performance: Experiment 1: p=0.87; Experiment 2: p=0.03; Experiment 3: p=0.29; two-tailed t-tests).

Another difference between humans and the model is the size of saccades (Supplementary Figure 11G-I). For example, in Experiment 2, the average saccade size was 7.6 ± 5.7 degrees for humans and 16.8 ± 8.4 degrees for IVSN (Experiment 2: Supplementary Figure 11H, $p < 10^{-15}$, two-tailed t-test, t=62, df=22960; Experiment 3: Supplementary Figure 11I, $p < 10^{-15}$, two-tailed t-test, t=100 df=29263). Humans typically made relatively small saccades (Supplementary Figure 11H-I). In contrast, the saccade sizes for the model were approximately uniformly distributed (Supplementary Figure 11H-I). We used the empirical distribution of saccade sizes to probabilistically constrain the saccade sizes for the model, creating a new variation of the model, $IVSN^{size}$ (Methods) The distribution of saccade sizes for the IVSNsize model resembled that of humans. IVSNsize showed similar performance to IVSN (Experiment 1: p=0.97; Experiment 2: p=0.52; Experiment 3: p=0.47; Supplementary Figure 11J-L), suggesting that the distribution of saccade sizes plays a lesser role in overall search efficiency.

Attentional modulation based on the target features is implemented in the IVSN model as a top-down signal from layer 31 to layer 30 in the VGG16 architecture (Figure 7.3, Methods). Connectivity in cortex is characterized by ubiquitous top-down signals at every level of the ventral visual stream. We considered variations of the model where attention modulation was implemented via top-down signaling at different levels: layer 31 to 30 (default, Figure 7.3), layer 24 to 23 ($IVSN_{24\to23}$), layer 17 to 16 ($IVSN_{17\to16}$), layer 10 to 9 ($IVSN_{10\to9}$), layer 5 to 4 ($IVSN_{5\to4}$) (Supplementary Figure 13). In general, these model variations were also able to find the target above

chance levels (all models were statistically different from chance except for $IVSN_{5\to4}$ in Experiment 1). The low-level features (layer 5 to layer 4) showed the lowest performance, probably because they lack the degree of transformation invariance built along the ventral stream hierarchy. Generally, model features at higher levels showed better performance but the trend was not monotonic. For example, $IVSN_{24\to23}$ showed slightly better performance than IVSN in Experiment 1 (Supplementary Figure 13A), but this difference was not statistically significant (p=0.045, two-tailed t-test, t=2, df=299).

We also considered the AlexNet [227], ResNet [226] and FastRCNN [28] architectures instead of the VGG16 architecture for the ventral visual cortex in Figure 7.3 (Supplementary Figure 14). All of these alternative models were above chance in all the experiments ($p < 0.006$, Supplementary Discussion).

6.6 Discussion

We examined 219,601 fixations to evaluate how humans search for a target object in a complex image under approximately realistic conditions and proposed a biologically plausible computational model that captures essential aspects of human visual search behavior. Subjects efficiently located the target in object arrays (Figure 6.3), natural images (Figure 6.4), and Waldo images (Figure 6.5) despite large changes in the appearance of the target object when rendered in the search image. Search behavior could be approximated by a neurophysiology-inspired computational network consisting of a bottom-up architecture resembling ventral visual cortex, a pre-frontal cortex-like mechanism to store the target information in working memory and provide top-down guidance for visual search, and a winner-take-all and inhibition-of-return mechanism to direct fixations. Both humans and the Invariant Visual Search Network (IVSN) model, demonstrated selectivity, efficiency and invariance, and did not require any training whatsoever with the sought targets.

Human visual search was efficient in that it required fewer fixations than alternative null models including random search, template matching, and sliding window models (Figures 6.3E, 6.4E, 6.5E). Humans actively sampled the image in a task-dependent manner, guiding search towards the target. Human visual search demonstrated invari-

ance in being able to locate objects that were transformed between the target image and the search image in size (Experiments 1, 2, 3), 2D rotation (Experiments 1 and 2), 3D rotation (Experiment 2), color (Experiment 3), different exemplars from the same category (Experiments 1 and 2), and other appearance changes including occlusion (Experiments 2 and 3). The large dissimilarity between how the targets were rendered in the search image and their appearance in the target image indicates that humans do not merely apply pixel-level template matching to find the target. These results suggest that the features guiding visual search must be invariant to target object transformations.

The problem of identifying objects invariantly to image transformations has been extensively discussed in the visual recognition literature (e.g.,[15, 185, 227, 226], among many others). Indeed, the ventral visual cortex module in IVSN is taken from a computational model that is successful in object recognition tasks, VGG16 [185]. The invariance properties in IVSN are thus inherited from VGG16. The current results show that the types of features learned upon training VGG16 in an independent object labeling task (ImageNet [224]), can be useful not only in a bottom-up fashion for visual recognition, but also in a top-down fashion to guide feature-based attention changes during visual search. The current results show that top-down features guiding visual search must show invariance to object transformations.

There has been extensive work characterizing the features that guide visual search [18]. IVSN incorporates those ideas into a quantitative image-computable framework to explain how the brain decides where to allocate attention in a task-dependent manner. Importantly, there is no additional training in IVSN to achieve invariance. The current model, as well as other models of feature-based attention [19, 20, 35, 39, 233], assume that such top-down influences provide feature-selective and transformation-tolerant information. The lack of any training or fine-tuning in IVSN distinguishes the proposed model from other work in the object detection literature that focuses on supervised learning from a large battery of similar examples to locate a target [28, 173]. The ability to perform a task without extensive supervised learning by extrapolating knowledge from one domain to a new domain is usually referred to as "zero-shot training". The specific exemplar objects in Experiments 1-2 were new to the subjects, even though subjects had extensive experience with those object categories. Subjects were also able

to efficiently search for novel objects from novel categories that they had never encountered before (Supplementary Figure 10). IVSN was able to find novel objects from known categories in Experiment 1. More strikingly, IVSN could find target objects in natural images even when those objects came from categories that it had never encountered before (Experiment 2, Supplementary Figure 5). Furthermore, IVSN could find Waldo in images that did not resemble any of the images used to train VGG-16 (Experiment 3). The ability to generalize and search for novel objects that have never been encountered before is consistent with the psychophysics literature showing that there are common feature attributes that guide visual search [18]. IVSN extends and formalizes the set of attributes from the low-level features that have been extensively studied in psychophysics experiments (*e.g.* color, orientation, etc.) to a richer and wider set of transformation-tolerant features relevant for visual recognition and for visual search under natural conditions.

Beyond exploring average overall performance, it is interesting to examine the spatiotemporal sequence of fixations for individual images. There is a large degree of variability when scrutinizing visual search at this high-resolution level. The same subject may follow a somewhat different eye movement trajectory when presented with the same exact target image and search image (Figure 6.6, Supplementary Figures 7-8), an effect that cannot be accounted for by memory for the target locations (Supplementary Figure 7). As expected, the degree of self-consistency was higher than the degree of between-subject consistency, which was in turn higher than the degree of subject model consistency at the image-by-image level both for the number of fixations (Supplementary Figure 7) and for the spatiotemporal sequences of fixations (Figure 6.6, Supplementary Figure 8).

Chapter 7

Target Inference Network: Inferring What A Person is Looking For

This chapter is based on the paper named "What am I searching for?"[1].

Figure 7.1 illustrates the target inference problem. Human subjects were instructed to move their eyes to search for a given target (A) in the search image (B) irrespective of changes in size, rotation angles, or other format changes. The visual search task resulted in a sequence of fixations (C, yellow circles with the arrows). The red bounding box refers to the ground truth target location in the search image (not shown in the actual experiment). In this example, the subject required 2 fixations to find the target. We defined the fixations falling on the *non-target* objects as "error fixations" before the target was found. In the target inference task, given the error fixations recorded from the psychophysics visual search task (D, yellow circle), the model is asked to infer what target object the subject was searching for out of the remaining possible objects (E, question marks in orange color, the question marks are not shown to the computational model). In this example, there is only 1 error fixation, in general, there could be anywhere from 1 to 4 error fixations in these experiments with arrays of 6 objects.

[1]Paper download link: `https://docs.wixstatic.com/ugd/d2b381_dc785deebb184aebb56fbc7522a70837.pdf`

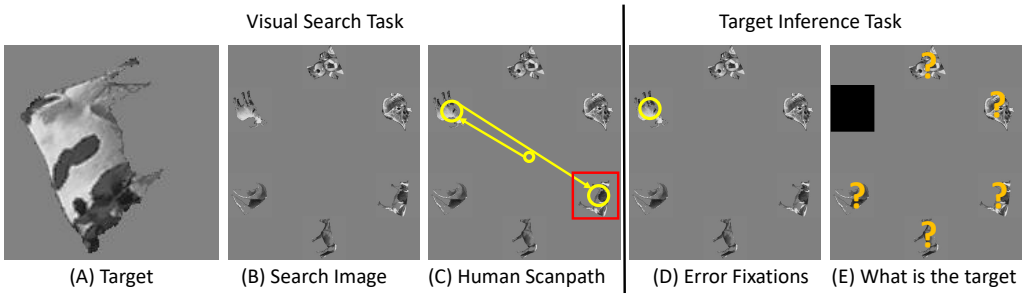
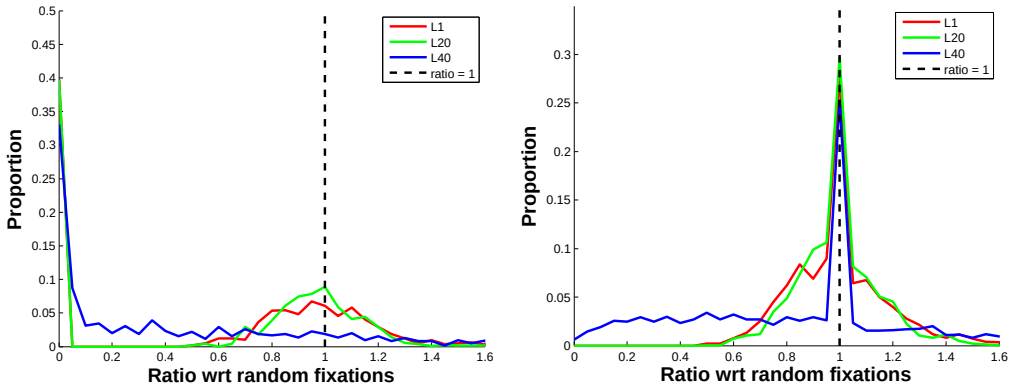


Figure 7.1: Illustration of the target inference problem.



(a) Last fixations on target objects (b) Error fixations on non-target objects

Figure 7.2: Feature similarity analysis between error fixations or last on-target fixations and the given target exemplars across layers in object arrays in human visual search tasks.

According to [52, 53, 54], the error fixations share more target-similar features than distractors. Thanks to the rich deep features from VGG16 pre-trained on object classification task, we provided a detailed analysis of the feature similarity between pairs of error fixations-target and pairs of random fixations-target in object arrays within layers and across layers in VGG16.

Figure 7.2(b) showed the distribution of the ratio between L2 distance of last on-target fixation-target pair versus random fixation-target pairs in subplot (a) and error fixation-target pair versus random fixation-target pair in subplot (b) using features extracted from different layers in pre-trained VGG16. Similar trends are observed across all layers. For simplicity, we only presented results on Layer 1 (red), Layer 20 (green) and Layer 40 (blue). The dash line denotes the ratio equal to 1. If there are more error fixations-target pairs which are more similar than random fixation-target pairs, the distribution of the ratios is skewed to the left of dash line (ratio $= 1$) and vice versa.

If error fixations share more target-similar features, the distance between error fixations and the targets is expected to be smaller than random fixation-target pairs which makes the ratio less than 1. We observed that the ratio distribution is skewed to the left of dash line (ratio equals 1) within layers which validates the point that there are more error fixations which share more similarities with the targets than random fixations versus the targets. The ratio distribution in higher layers (from 1 to 40) becomes more dispersed with larger variance which suggests features in higher layers have more discriminative power to distinguish target-similar objects among other distractors. The area under the curve to the left of the dash line becomes more dominant than the other half as the layer number increases.

For comparison purposes, we also provided the ratio distribution between last on-target fixation-target pairs versus random fixation-target pairs (See Figure 7.2(a)). As expected, since the last fixations are on the target, they should share more feature similarities than random fixation-target pairs. Thus, most of the pairs should have ratio less than 1. Across all layers, we observed that the area under the curve for ratio < 1 is far greater than the other half. As the layer number increases, the area under the curve on the left of the dash line (ratio = 1) increases.

Our model is based on the idea that the location with more feature similarities for all error fixations is more likely to be the search target location. We approximate the target inference problem in feature similarity space among targets and distractors: given T error fixations with coordinates (x_i, y_i) where $1 \leq i \leq T$, the task is to predict a 2D probabilistic map M_f of where the search target is most likely to be (Figure 7.3). We take the maximum on M_f as the current guess location. If the cropped area centered at the current guess location overlaps with the ground truth bounding box encompassing the whole target object, the inference is deemed successful; otherwise, after each incorrect guess, the map is updated by removing the erroneous inference location on M_f.

7.1 Zero-shot Target Inference Model

We provide an overview of the model, followed by a more detailed description of our proposed zero-shot deep network (InferNet, Figure 7.3).

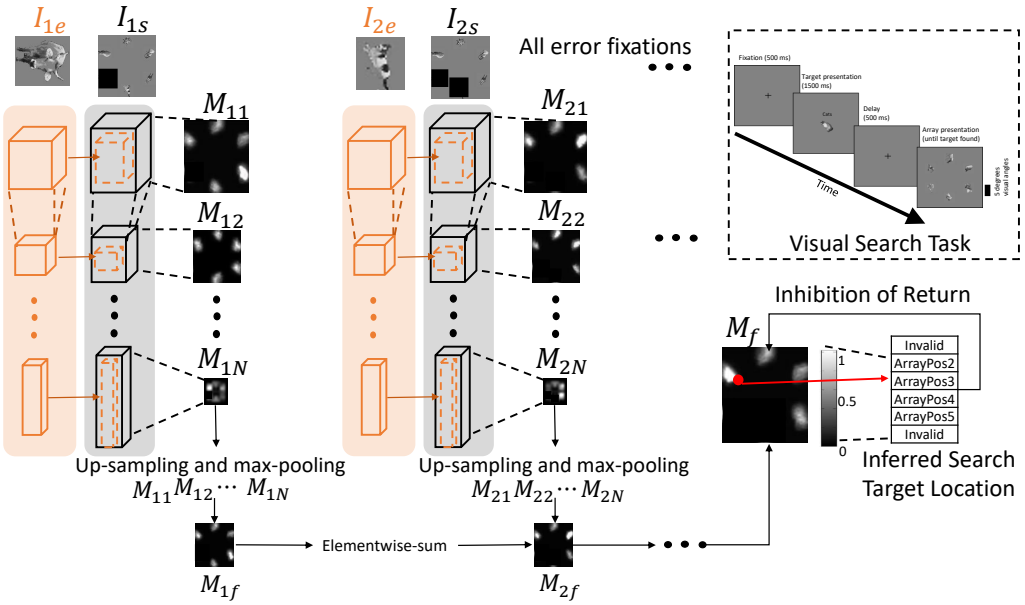


Figure 7.3: Architecture of InferNet.

The model is based on a pre-trained deep convolutional network that is applied to the error fixations (**Prior Nework (PN)**) and to the search image (**Likelihood Netwrok (LN)**). **PN** takes the cropped area I_{ie} of size 28×28 pixels centered at error fixation i as input and outputs feature maps across layers. We define I_{is} as the search image which has the objects at all past error fixations $1, ..., i$ inhibited with a black mask. **LN** modulates the feature maps from I_{is}, generating a series of likelihood maps (M_{i1}, M_{i2}, ..., M_{ij}, ...,M_{iN}) across different layers where j denotes the index of the jth likelihood map M_{ij} for error fixation i. These maps are concatenated and max-pooled to produce the final likelihood map M_{if} for error fixation i which tracks the parts of the image that are most similar between I_{ie} and I_{is}. InferNet integrates these likelihood maps M_{if} across all T error fixations via elementwise-sum by assuming all the error fixations play equally important roles in contributing to the final inference map M_f.

7.1.1 Prior Network

We used a deep feed-forward network, implemented in VGG16 [185], and pre-trained for image classification on the ImageNet dataset [224]. We show that the invariant features from VGG16 can be directly used for target inference task without any additional training. Given I_{ie} at error fixation i, the network weights W learnt from image

classification extract feature maps $\varphi_j^{PN}(I_{ie}, W)$ at layer j (orange boxes in Figure 7.3).

7.1.2 Likelihood Network

Given I_{is}, **LN** has the same network parameters W as **PN** and extracts the feature representation of I_{is} at layer j, $\varphi_j^{LN}(I_{is}, W)$ (gray boxes in Figure 7.3). The weights are shared between **PN** and **LN**, and both are pre-trained for image classification, not for target inference. The weights W do *not* depend on I_{is} or I_{ie}. The InferNet network has no prior training with the objects or images in this study. The locations of the error fixations in I_{is} are blacked out (so that the model does not indicate that the most similar location to an error fixation is the error fixation itself). The input to **PN** is smaller than the input to **LN**, hence the output $\varphi_j^{PN}(I_{ie}, W)$ is smaller than $\varphi_j^{LN}(I_{is}, W)$. The activity of the units in **LN** in response to the search image is modulated by those in **PN**, which contain features more similar to the visual search target than distractors.

The modulation in the activation map is achieved by convolving the representation of the error fixation with the representation of the search image at multiple scales:

$$M_{ij} = m(\varphi_j^{SN}(I_{is}, W), \varphi_j^{PN}(I_{ie}, W)) \tag{7.1}$$

where $m(\cdot)$ is the error fixation modulation function defined as 2D convolution with kernel $\varphi_j^{PN}(I_{ie}, W)$ on the search feature map $\varphi_j^{LN}(I_{is}, W)$ where j denotes the index of the jth feature similarity map M_{ij} for error fixation i.

Inspired by neurophysiological recordings during visual search and attentional modulation in visual cortex [22, 59, 19], and with the goal of capturing target properties at multiple scales and with different features, modulation is applied across multiple layers. Intuitively, if the target object shares more similarities with the error fixations in low-level features, such as similar orientations, error fixation modulation on M_{ij} may be sufficient; however, if high-level features are shared between the target and the error fixations, such as surface texture, feature similarity maps at higher levels may be required. We empirically selected $N = 7$ feature similarity maps. In InferNet, the following specific layers were selected: M_{i1} (layer $j = 5$ of VGG16, size 101×101), M_{i2} (layer $j = 10$, 52×52), M_{i3} (layer $j = 17$, 27×27), M_{i4} (layer $j = 23$, 27×27), M_{i5} (layer $j = 24$, 15×15), M_{i6} (layer $j = 30$, 15×15), M_{i7} (layer $j = 31$, 9×9). The

layer number refers to the layer index in VGG16 [185].

Each of these feature similarity maps is up-sampled to 224×224 pixels and the final feature similarity map is max pooled at each location (x, y) on M_{ij} over all the N intermediate maps (Table 7.2 reports performance separately for each feature similarity map). The model thus keeps track of all the locations which share similar sub-patterns including both low-level and high-level feature descriptors:

$$M_{if}(x, y) = \max_{j=1}^{N} M_{ij}(x, y) \tag{7.2}$$

7.1.3 Combination of Maps and Target Inference

The feature similarity maps M_{if} are summed over all T error fixations:

$$M_f(x, y) = \sum_{i=1}^{T} M_{if}(x, y) \tag{7.3}$$

We assume all error fixations play equally important roles in inferring the search target. In general, it is possible to use a weighted summation where some error fixations are more important than the rest depending on the applications. InferNet selects the maximum of the M_f map. If the cropped area centered at the current guess location overlaps with the ground truth bounding box encompassing the whole target object, the inference is deemed successful and the inference stops. Otherwise, that location is inhibited and the next maximum is selected.

7.1.4 Evaluation

To evaluate performance of InferNet, we computed the average number of guesses required over all the trials with different images as a function of the number T of error fixations. The less number of guesses required, the more effective the inference process is. However, since the target inference difficulty varies, we report the relative performance P_r defined as the average number of guesses required by the computational model $A_m(T)$ relative to the average number of guesses required by a chance model $A_c(T)$ on the same image and task (the chance model is defined below):

$$P_r(T) = \frac{A_c(T) - A_m(T)}{A_c(T)} \times 100\% \tag{7.4}$$

If the computational model requires less number of guesses on average, $P_r(T)$ is greater than zero. The larger $P_r(T)$, the more efficient the inference process is.

7.2 Experiments on Target Inference

We tested InferNet on images containing object arrays and in natural images by evaluating the number of guesses required to correctly infer the sought target, $P_r(T)$. As benchmarks, we compared our model with other alternative null models.

7.2.1 Datasets

We used the dataset introduced in the previous Chapter 6 on object arrays and natural images. We excluded the Waldo dataset because the target to search for remains the same throughout the experiment.

7.2.2 Comparative Null Models

We compared our model with several alternative null models. In all cases, the alternative models proposed an inference map and the procedure to select a target was the same as with InferNet, including infinite inhibition-of-return (*i.e.* never selecting the same location twice).

Chance. We considered a model where the target location was chosen at random. For object arrays, we randomly chose one out of the remaining possible locations. For the natural images dataset, a random location was selected for each guess. This random process was repeated 20 times.

Template Matching. To evaluate whether pixel-level features of the error fixations were sufficient for guiding inference, we introduced a pixel-level template matching model where the inference map was generated by sliding the canonical target of size 28×28 pixels over the whole search image of size 224×224 pixels. Compared to

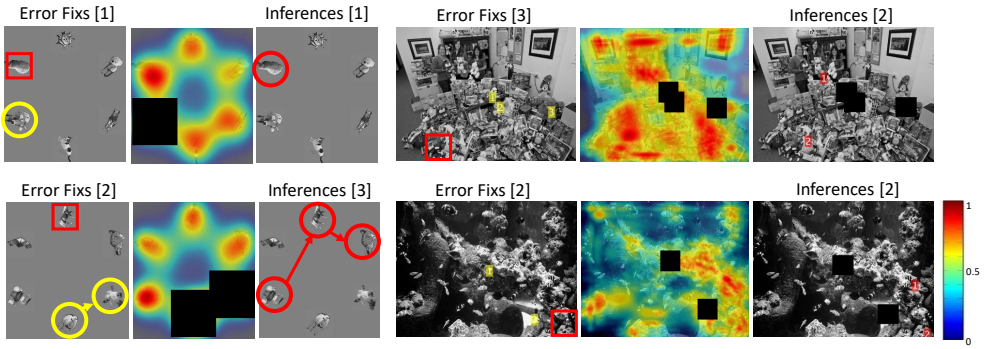
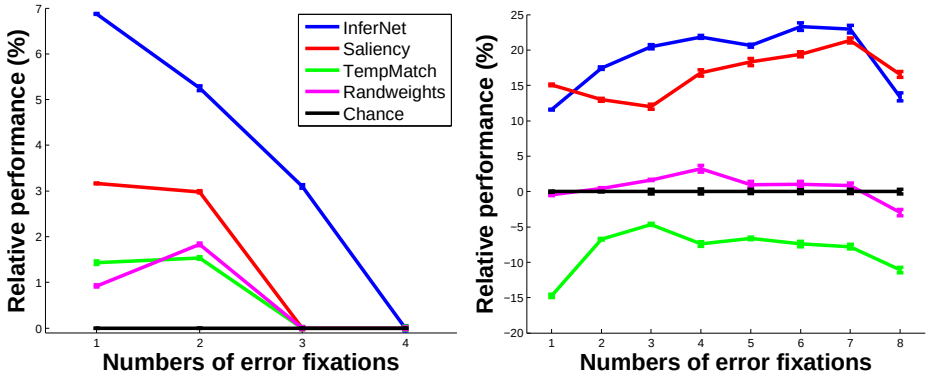


Figure 7.4: Two example results of target inference in object arrays (first 3 columns) and two examples in natural images (last 3 columns).



(a) Object arrays (b) Natural images

Figure 7.5: Evaluation of model inference performance for object arrays (a) and natural images (b).

the classical sliding window models in computer vision, this can be interpreted as an "attentional" sliding window.

IttiKoch. We considered a pure bottom-up saliency model that has no information about the error fixations [3].

RanWeight. Instead of using VGG16 [185] pre-trained for image classification, we randomly picked weights W from a gaussian distribution with mean 0 and standard deviation 1000. The network was otherwise identical to InferNet. The random selection of weights was repeated 100 times.

7.3 Model Inference Performance

In this section, we provided the evaluation results of our model on object arrays and natural images in terms of the location and category information of the targets.

Figure 7.4 shows example results of target inference by our InferNet model on object arrays and natural images. Given the "error fixations" (yellow circles, column 1 and 4), the InferNet model predicts the 2D probabilistic map M_f overlaid on the stimuli (Columns 2 and 5, scale on the right). The red bounding box (Column 1, 4) denotes the ground truth area encompassing the search target. The red circles (Column 3) and black boxes (Column 6) show the successive maxima of the final inference map. InferNet correctly determined the target at the 1st and 3rd guess (Column 3) and in the 2nd guess (Column 6).

Figure 7.5 shows quantitative evaluation of model inference performance. Relative performance improvement for the computational model relative to the chance model is as a function of the number of error fixations. The smaller the number of guesses, the better the inference algorithm is and the higher the relative performance improvement. The different colors denote different models: InferNet model (blue), bottom-up IttiKoch saliency (red), template matching (green), RanWeight (magenta), Chance (black). See Section Comparative Null Models for descriptions. Error bars are standard error of the mean for all trials.

7.3.1 Object Arrays

Figure 7.4 shows examples illustrating how the model efficiently inferred the target location given only one or two fixations on object arrays. In the first example (Column 1-3, Row 1), a subject made one error fixation on the cow which looks visually similar to the sheep before finding the sheep. Given this single error fixation, InferNet determined that the subject was probably looking for a sheep among all the five remaining distractors (red circle, Column 3, Row 1). In the second example (Column 1-3, Row 2), a subject made 2 error fixations before finding the target (horse). In this case, InferNet correctly determined the target at the 3rd guess (Column 3, Row 2).

InferNet showed an overall improvement of $3.8 \pm 3\%$ with respect to the chance model over all error fixations (Figure 7.5(a), blue). Even with a single error fixation as

input data, InferNet could infer the target 6.87% faster than the chance model. That is, while random guessing would correctly land on the target within 3 guesses, InferNet only required 2.80 ± 0.01 guesses on object arrays.

In Figure 7.5(a), none of the null models reached the level of relative performance improvement shown by InferNet ($P < 4.6 \times 10^{-20}$, two-tailed t-test, $t = -9.2$, $df = 12128$) for all the numbers of error fixations except for the case of 4 error fixations where none of the models were above chance. Though we took precautions to normalize *average* low-level features on arrays, for goal inference, on any trial, InferNet can capitalize on shared IttiKoch features between error fixations and the target. Performance for the bottom-up saliency model (IttiKoch) is better than the chance model but still below InferNet which suggests that the target information embedded in error fixations is useful for target inference. The model with random weights (RanWeight) and the model with template matching (TempMatch) on pixel levels show minimal improvements from selecting random locations (Figure 7.5(a)), suggesting the discriminative features learnt from a hierarchical network for image classification are important for target inference.

7.3.2 Natural Scenes

The experiment reported so far focused on images consisting of segmented objects at discrete locations, presented on a uniform background, at fixed positions equidistant from the center of the image. In the real world, visual search happens most of the time in cluttered environments involving non-segmented objects amidst a complex background. As the inference space becomes continuous (the target object could be anywhere on the search image), the inference problem becomes more challenging and hence, there is higher demand for computational models to assist in target inference in these scenarios. To evaluate whether our model could generalize to complex natural scenes, we extended the previous results by evaluating the relative performance of InferNet in the natural images (Figure 7.4 and Figure 7.5(b)).

Figure 7.4 shows two examples where InferNet successfully determined the target in natural images. The appearance of the target in the search image is notably different from that in the target image due to changes in size and 3D rotation. Yet, the examples

in Figure 7.4 show that InferNet can still effectively use features from error fixations to infer what the target is. For example, in Row 1, column 4, the error fixations fall on plush toys, such as teddy bears. Based on the characteristics of all plush toys, InferNet outputs an inference map with high activations around all the plush toys regions. In this example, InterNet correctly inferred the target within 2 guesses. In another example (Row 2, column 4), all the high activations on M_f focused on ground regions, such as the surface of coral reefs. InferNet can extract the essential texture information of ground surface under the sea and consider the features shared across all error fixations.

Figure 7.5(b) shows that InferNet was successful at inferring the target in natural images with significant improvements of $19 \pm 4\%$ compared with the chance model. In general, InferNet required an average of 16.2 ± 0.07 guesses given only one error fixation and 15 ± 0.6 guesses given 8 error fixations (blue) while the chance model required 18.2 guesses given only one error fixation and 17.3 guesses given 8 error fixations. As we observed in Figure 7.5(b), InferNet outperformed all the alternative null models ($P < 4 \times 10^{-27}$, two-tailed t-test, $t = -10.8$, $df = 140422$). Performances for the bottom-up saliency model (IttiKoch) was relatively high among all the null models because target objects were typically salient and they occupied a large percentage of the image. We also noted that template matching under-performs the chance model. It is possible that pixel-level matching misleads the model towards wrong locations and wrong cues are worse than random cues.

We also observed that given more error fixations, the average number of guesses required to infer the target of interest was reduced. This effect can be ascribed to two factors: (i) the hypothesis space, *i.e.* number of location choices on the search image, is reduced with more error fixations, and (ii) more error fixations provide richer information that is useful for target inference.

7.3.3 Target Category Inference

In addition to inferring the target location, we tested InferNet on sought object category inference task. Out of 240 natural images, we selected 100 images where the target categories belong to ImageNet. For each subject, InferNet predicts the belief of possible target categories out of total 1000 categories by leveraging on the weights pre-trained

Error Fixations	Top N category inference accuracy $\%$							
	1	2	4	8	16	32	64	128
1	6	8	11	13	17	29	38	55
2	4	9	14	20	23	33	46	65
3	3	10	16	25	28	38	51	72
4	0	13	20	20	30	39	54	74
5	3	11	23	28	34	42	56	74
6	0	14	23	31	37	45	54	77
7	1	15	25	31	37	47	53	78
8	4	17	28	37	40	48	57	80

Table 7.1: Our model performance of top N inferred target category accuracy across error fixations (rows) where $N = 1, 2, ..., 128$ (columns) is shown.

on ImageNet and accumulates these belief across error fixations. Table 7.1 reports the accuracy of top N most probable target categories inferred by InferNet based on the accumulated belief across error fixations. We have two observations. First, given even only one error fixation, the inference accuracy of InferNet surpasses the chance model ($1/1000$). As N increases, the target category inference accuracy increases. Ideally, the accuracy of inferred top 1000 probable target categories should be 1 as the target always belong to at least one of the 1000 categories from ImageNet. Given 8 error fixations, InferNet is capable of inferring the target category correctly with accuracy of around 50% for top 32 most probable categories out of 1000 categories. Second, as InferNet takes more number of error fixations as inputs, the belief gets constantly updated and the inference becomes more accurate. This validates the error fixations carry important information revealing the target identity during visual search.

7.4 Ablation Study

In this section, we provided detailed model analysis via a series of ablation studies and compared the model performance with humans.

7.4.1 Effect of Low and High-level Features from Error Fixations

To evaluate the contribution of different layers of InferNet, we tested each individual feature similarity map M_j and their different combinations in object arrays and natural images. Table 7.2 shows our ablated models' relative performance (%) compared with

Table 7.2: Target inference relative performance (%) of ablated models compared with the chance model in object arrays and natural images given T error fixations.

	Object Arrays				Natural Images							
#Error Fixations	1	2	3	4	1	2	3	4	5	6	7	8
InferNet (our model)	6.87	5.25	3.10	-	12.83	19.67	22.48	24.20	24.35	25.59	28.28	18.14
Layer 5	1.88	3.51	1.58	-	9.35	15.91	17.11	14.70	17.24	13.18	20.91	9.56
Layer 10	3.98	4.07	0.67	-	14.69	21.26	24.82	23.18	25.16	23.82	26.97	15.98
Layer 17	5.96	5.64	1.99	-	16.50	22.51	19.28	23.50	22.42	19.17	26.43	14.38
Layer 23	7.46	6.13	0.01	-	13.32	22.44	24.72	22.33	28.07	25.00	23.56	16.93
Layer 24	6.60	6.74	3.28	-	18.53	25.73	28.04	28.10	30.59	28.37	30.42	27.61
Layer 30	8.21	5.77	3.08	-	-	7.04	4.45	0.51	6.03	0.02	3.36	-
Layer 31	7.56	3.78	2.34	-	-	6.15	4.60	-	5.00	2.26	3.93	-
Max + Max	6.87	3.99	1.13	-	12.84	19.40	21.11	22.13	22.96	21.75	24.49	20.01
Mean + Max	7.01	4.48	2.63	-	8.67	11.60	11.97	12.66	14.22	11.87	16.05	7.92
Mean + Mean	7.01	6.24	3.68	-	8.67	10.60	9.68	9.78	10.61	8.71	13.31	6.30
Human performance	-	-	-	-	60.87	74.27	67.33	66.20	38.18	43.29	35.65	44.47
Model using common fixations	10.96	1.29	1.52	2.22	20.09	-	30.35	26.30	24.98	36.34	25.77	3.90

the chance model using feature similarity maps (M_j) at different layers j for T error fixations. The larger, the better. (-) denotes performance not significantly better than chance. The layer number refers to the index in the VGG16 network [185]. The first row M_f corresponds to our full model considering all feature similarity maps across layers whereas the other rows show the predictions using either only one feature similarity map from M_{i1} to M_{i7} in Figure 7.3 or their combinations.

From Table 7.2, we have several observations: (1) Compared to the individual maps, target inference performance was generally more effective using the feature similarity maps M_j in higher layers which implies that high-level features extracted at error fixations are more reliable for target inference. (2) We are also interested in exploring how the compositionality of feature similarity maps across layers reveals the identity of the target. InferNet takes max-pooling of M_{ij} for error fixation i and averages M_{if} for all T error fixations. Instead of max-pooling across layers, we also evaluated ablated models where the max-pooling across N layers is replaced by averaging and vice versa. We did not observe any significant improvements in object arrays but different combination methods of feature similarity maps contribute dramatically differently in natural images. Our InferNet model outperforms the rest which suggests error fixations seem not to be guided by the overall target features as a whole (taking average across N layers) but by sub-patterns of the search target (max-pooling across N layers) which aligns with [180].

7.4.2 Effect of Locations and Sequence Orders of Error Fixations

Our InferNet model treats all error fixations equally and only utilizes the visual feature information at the error fixations. In the last ablated model, we study the role of the

Table 7.3: Target inference performance of ablated model after taking into account of error fixation locations and fixation order information in two datasets: object arrays and natural images.

	Object Arrays				Natural Images							
# Error Fixation	1	2	3	4	1	2	3	4	5	6	7	8
InferNet (our model)	6.87	5.25	3.10	-	12.83	19.67	22.48	24.20	24.35	25.59	28.28	18.14
Location + Sequence	7.14	10.61	7.87	0.78	-37.78	-26.14	-30.16	-31.91	-28.65	-34.25	-30.92	-34.81

locations and the sequence order of error fixations in target inference. We created the ablated model and trained it using supervised learning to predict the final inference map directly: (1) generate a binary error fixation map masked with gaussian kernels to denote the locations of error fixations and the magnitude of gaussian kernels vary depending on the fixation order. The higher intensity of the gaussian mask is applied at the error fixation, the more recent the error fixation is. (2) concatenate this fixation map with the search image as inputs to a feed-forward 2D convolution neural network. (3) KullbackLeibler divergence loss is computed between the predicted inference map and the ground truth map where 1 denotes the target location and 0 otherwise.

In Table 7.3, we reported the result (Location+Sequence) in both object arrays and natural images. This ablated model which takes location and order information into account performs equally well as InferNet in object arrays but much worse than InferNet in natural images. It is surprising that the experimental result seems to suggest the location and order information of error fixations do not matter much in target inference task.

7.4.3 Comparison of Human and Model Performance

Human visual search is variable both within-subject and between-subjects [19]. We conducted additional psychophysics experiments to explore the question whether human subjects could correctly infer what the target is on the search image. We reported the results in Table 7.2 (last two rows). Humans were not able to solve the inference problem in object arrays but were better than InferNet in natural images, perhaps by using contextual cues (second last row). To investigate the between-subject variability, we created a new model using only error fixations that are common across subjects. The result (last row) shows that in some (but not all) cases, InferNet can overcome the consequences of variability in human scanpath patterns. However, in general, we need algorithms that can predict *individual intentions in single trials*, which is the goal for

InferNet.

Chapter 8

Conclusions and Perspectives

Within the context of visual attention, we are interested in developing a computational visual attention model integrated with memory, which can adaptively apply both bottom-up and top-down attention modulations in various tasks.

8.1 Summary of Bottom-up Visual Attention Models

In the first half of the thesis, we design several attention subsystems to explore the role of the fovea, the bottom-up pathway in the ventral stream of the visual cortex as well as their interaction with the working memory. Our experimental results have shown that our current models have outperformed state-of-the-art methods in scanpath prediction on static images, gaze prediction and anticipation in egocentric videos. The contribution in each chapter is summarized below:

In Chapter 3, we go beyond the current deep neural network-based saliency map prediction and extend it to visual scanpath prediction. We introduced DSNN, the first RNN on scanpath prediction. It integrates the sequence of fixations to estimate the temporal saliency maps, and it makes decision on where the human subjects may look next. In addition to substantial improvements on scanpath prediction compared with the state-of-the-arts, DSNN also obtains a competitive predictive accuracy of the saliency map with state-of-the-art models. Our analysis on the learnt model demonstrates the utility of recurrent connections in the predictive scanpath accuracy and the emergence of a temporally changing spatial bias during the scanpath prediction.

In Chapter 4, we present a novel foveated neural network for gaze prediction on

egocentric videos. Evaluation results on the publicly available dataset demonstrate that FNN outperforms the state-of-the-art methods. The integration process of proposing, attending and analysing ROI on the previous frame as well as the feature extraction from the current frame helps gaze prediction performance. We also incorporate head movement to FNN by introducing the dense optical flow as the additional feature inputs. We will extend FNN to more than two adjacent frames by introducing a memory module in the near future.

In Chapter 5, we present a new challenging gaze anticipation problem on future frames as an extension of the gaze prediction problem on current frames on both egocentric and third person videos. We develop an integrated framework, named as Deep Future Gaze (DFG), consisting of two pathways: bottom-up pathway **DFG-G** built upon Generative Adversarial Network (GAN) and task-specific pathway **DFG-P** generating gaze spatial prior maps which modulate the bottom-up saliency prediction. We evaluate our integrated model using standard metrics and our performance surpasses all the competitive baselines significantly in both egocentric and third-person videos covering various activities, such as cooking and object search tasks. Moreover, we investigate the potential factors contributing to better gaze anticipation performance and justify the importance of the individual component in our proposed architecture. Though our model is not specifically trained for gaze prediction problem on current frames, DFG performs better compared with the state-of-the-art. Different from all the existing methods, DFG does not require explicit egocentric cues or any past information.

8.2 Summary of Top-down Visual Attention Models

In the second half of my thesis, we develop our attention model and integrate it with memory functions as well as top-down modulation mechanism. At last, we introduce how we apply the top-down attention modulation in the target inference task.

In Chapter 6, we show for the first time that humans can efficiently and invariantly search for natural objects in complex scenes. To gain insight into the mechanisms that guide visual search, we propose a biologically inspired computational model, Invariant Visual Search Network (IVSN) model. The current model provides a reasonable initial sketch that captures how humans can selectively localize a target object amongst

distractors, the efficiency of visual search behavior, the critical ability to search for an object in an invariant manner, and zero-shot generalization to novel objects including the famous Waldo. Waldo cannot hide any more.

Even when our Invariant Visual Search Network (IVSN) model may approximate human search behavior, the model may not be searching in the same way that humans do. First, IVSN shows constant acuity over the entire visual field, which is clearly not the case for human vision where acuity drops rapidly from the fovea to the periphery. Second, humans must decide after each saccade whether the target is present or not. The default IVSN model executed this decision through an "oracle" (the same oracle was used for the human data for fair comparison). As a proof-of-principle, we implemented a recognition step for each fixation, a step that can be improved through the extensive work on invariant visual recognition systems [15, 185, 227, 226]. Humans also make recognition mistakes (*e.g.*, visual search experiments in natural images and Waldo images where subjects fixated on the target yet did not click the mouse). Third, humans also revisit the same location even if the target is not there. Yet, the default IVSN model implements infinite inhibition of return as a simplifying assumption that could also be improved upon by including a memory decay function, as shown in $IVSN_{fIOR}$. Fourth, there is no learning in the current model. The visual system could learn the interaction of the different bottom-up, top-down, memory and recognition components. An elegant idea on how learning could be implemented was presented in **??** where the authors proposed an architecture that can learn to generate eye movements via reinforcement learning with a system that is rewarded when the target is found. IVSN can be improved by training or fine-tuning for various search tasks. Fifth, the model assumes that each saccade is independent of the previous one except for the inhibition-of-return mechanism and the saccade distance constraints. A complete model should incorporate inter-dependences across saccades such that visual information obtained during previous fixations can be used to guide the next saccade. Finally, subjects may capitalize on high-level knowledge about scenes [18, 234] including statistical correlations in object positions (e.g., car keys are usually not glued to the ceiling), physical properties (keys are more likely on top a desk rather than floating in the air), correlations in object sizes (the size of a phone may set an expectation for the size of the keys), etc.

In Chapter 7, we proposed a computational model to infer intentions from behaviors in the context of a visual search task. InferNet can determine what the sought target is, in object array images as well as in natural images, by using the prior set of non-target fixations. InferNet is based on transfer-learning in that it uses weights learnt for a different task. InferNet is a "zero-shot" architecture: there is no training with the specific objects or images that the model analyzes during the inference process. Leveraging on the idea that error fixations share feature similarities with the targets, InferNet builds an implicit relationship between the inference problem and the feature similarity problem. The experimental results show that InferNet significantly outperforms the comparative null models.

There are many areas where the model could be improved. Most notably, inference could be enhanced by incorporating intuitive semantics in the real world (*e.g.* if the error fixations are mostly distributed on the ground, one could deduce that the target of interest would most likely not be the airplanes in the sky). Problem-specific training (e.g. weights for each layer, or weights for each error fixation) could also improve performance. The proof-of-principle demonstration in this study provides a possible inference solution to effectively guess what the subject is searching for in complex images and suggests that computational models can make reasonable conjectures to read the subject's mind purely based on behavioral data.

Bibliography

[1] J. R. Anderson, *Cognitive psychology and its implications* . WH Freeman/Times Books/Henry Holt & Co, 1990.

[2] R. A. Rensink, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *IEEE ICCV*, pp. 262–270, 2015.

[5] Y. Lin, S. Kong, D. Wang, and Y. Zhuang, "Saliency detection within a deep convolutional architecture,"

[6] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[7] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*, pp. 115–141, Springer, 1987.

[8] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention prediction in egocentric video using motion and visual saliency," in *Pacific-Rim Symposium on Image and Video Technology*, pp. 277–288, 2011.

[9] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *ICCV*, pp. 3216–3223, 2013.

[10] R. Ohme, M. Matukin, and B. Pacula-Lesniak, "Biometric measures for interactive advertising research," *Journal of Interactive Advertising*, vol. 11, no. 2, pp. 60–72, 2011.

[11] W. Ding, P. Chen, H. Al-Mubaid, and M. Pomplun, "A gaze-controlled interface to virtual reality applications for motor-and speech-impaired users," *HCI International*, 2009.

[12] M. Kumar, *Gaze-enhanced user interface design*. PhD thesis, Citeseer, 2007.

[13] F. Multon, L. France, M.-P. Cani-Gascuel, and G. Debunne, "Computer animation of human walking: a survey," *The journal of visualization and computer animation*, vol. 10, no. 1, pp. 39–54, 1999.

[14] R. C. Zeleznik, A. S. Forsberg, and J. P. Schulze, "Look-that-there: Exploiting gaze in virtual reality interactions," tech. rep., Technical Report CS-05, 2005.

[15] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[16] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, "A quantitative theory of immediate visual recognition," *Progress in brain research*, vol. 165, pp. 33–56, 2007.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] J. M. Wolfe and T. S. Horowitz, "Five factors that guide attention in visual search," *Nature Human Behaviour*, vol. 1, no. 3, p. 0058, 2017.

[19] T. Miconi, L. Groomes, and G. Kreiman, "There's waldo! a normalization model of visual search predicts single-trial human fixations in an object search task," *Cerebral Cortex*, vol. 26, no. 7, pp. 3064–3082, 2015.

[20] R. P. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard, "Eye movements in iconic visual search," *Vision research*, vol. 42, no. 11, pp. 1447–1463, 2002.

[21] T. J. Buschman and E. K. Miller, "Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations," *Neuron*, vol. 63, no. 3, pp. 386–396, 2009.

[22] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[23] N. P. Bichot, M. T. Heard, E. M. DeGennaro, and R. Desimone, "A source for feature-based attention in the prefrontal cortex," *Neuron*, vol. 88, no. 4, pp. 832–844, 2015.

[24] F. Tong and K. Nakayama, "Robust representations for faces: evidence from visual search.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, no. 4, p. 1016, 1999.

[25] J. M. Wolfe, "Saved by a log: How do humans perform hybrid visual and memory search?," *Psychological Science*, vol. 23, no. 7, pp. 698–703, 2012.

[26] O. Hershler and S. Hochstein, "The importance of being expert: Top-down attentional control in visual search with photographs," *Attention, Perception, & Psychophysics*, vol. 71, no. 7, pp. 1478–1486, 2009.

[27] J. M. Wolfe, T. S. Horowitz, N. Kenner, M. Hyle, and N. Vasan, "How fast can you change your mind? the speed of top-down guidance in visual search," *Vision research*, vol. 44, no. 12, pp. 1411–1426, 2004.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[29] J. Yang and M.-H. Yang, "Top-down visual saliency via joint crf and dictionary learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 3, pp. 576–588, 2017.

[30] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3743–3752, 2015.

[31] T. Gevers and A. W. Smeulders, "Pictoseek: Combining color and shape invariant features for image retrieval," *IEEE transactions on Image Processing*, vol. 9, no. 1, pp. 102–119, 2000.

[32] A. J. Rodriguez-Sanchez, E. Simine, and J. K. Tsotsos, "Attention and visual search," *International Journal of Neural Systems*, vol. 17, no. 04, pp. 275–288, 2007.

[33] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[34] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of vision*, vol. 9, no. 3, pp. 5–5, 2009.

[35] J. W. Bisley, "The neural basis of visual attention," *The Journal of physiology*, vol. 589, no. 1, pp. 49–57, 2011.

[36] T. Yao, S. Treue, and B. S. Krishna, "Saccade-synchronized rapid attention shifts in macaque visual cortical area mt," *Nature communications*, vol. 9, no. 1, p. 958, 2018.

[37] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annual review of neuroscience*, vol. 24, no. 1, pp. 167–202, 2001.

[38] J. Martinez-Trujillo, "Searching for the neural mechanisms of feature-based attention in the primate brain," *Neuron*, vol. 70, no. 6, pp. 1025–1028, 2011.

[39] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision research*, vol. 45, no. 2, pp. 205–231, 2005.

[40] B. T. Vincent, R. J. Baddeley, T. Troscianko, and I. D. Gilchrist, "Optimal feature integration in visual search," *Journal of Vision*, vol. 9, no. 5, pp. 15–15, 2009.

[41] L. J. Lanyon and S. L. Denham, "A model of active visual search with object-based attention guiding scan paths," *Neural Networks*, vol. 17, no. 5, pp. 873–897, 2004.

[42] F. H. Hamker, "The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas v4, it for attention and eye movement," *Cerebral cortex*, vol. 15, no. 4, pp. 431–447, 2004.

[43] B. Chen and P. Perona, "Speed versus accuracy in visual search: Optimal performance and neural architecture," *Journal of vision*, vol. 15, no. 16, pp. 9–9, 2015.

[44] M. S. Castelhano, M. L. Mack, and J. M. Henderson, "Viewing task influences eye movement control during active scene perception," *Journal of vision*, vol. 9, no. 3, pp. 6–6, 2009.

[45] G. Busswell, "How people look at pictures: A study of the psychology of perception in art," 1935.

[46] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Image processing, 2003. icip 2003. proceedings. 2003 international conference on*, vol. 1, pp. I–253, IEEE, 2003.

[47] J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk, "Predicting cognitive state from eye movements," *PloS one*, vol. 8, no. 5, p. e64937, 2013.

[48] T. Betz, T. C. Kietzmann, N. Wilming, and P. König, "Investigating task-dependent top-down effects on overt visual attention," *Journal of vision*, vol. 10, no. 3, pp. 15–15, 2010.

[49] A. Borji and L. Itti, "Defending yarbus: Eye movements reveal observers' task," *Journal of vision*, vol. 14, no. 3, pp. 29–29, 2014.

[50] S. T. Iqbal and B. P. Bailey, "Using eye gaze patterns to identify user tasks," in *The Grace Hopper Celebration of Women in Computing*, pp. 5–10, 2004.

[51] A. Yarbus, "Eye movements and vision. 1967," *New York*, 1967.

[52] M. P. Eckstein, B. R. Beutter, B. T. Pham, S. S. Shimozaki, and L. S. Stone, "Similar neural representations of the target for saccades and perception during search," *Journal of Neuroscience*, vol. 27, no. 6, pp. 1266–1270, 2007.

[53] R. G. Alexander and G. J. Zelinsky, "Visual similarity effects in categorical search," *Journal of Vision*, vol. 11, no. 8, pp. 9–9, 2011.

[54] J. M. Wolfe, "Guided search 4.0," *Integrated models of cognitive systems*, pp. 99–119, 2007.

[55] J. Braun, C. Koch, and J. L. Davis, *Visual attention and cortical circuits*. MIT Press, 2001.

[56] W. X. Schneider, "Vam: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action," *Visual Cognition*, vol. 2, no. 2-3, pp. 331–376, 1995.

[57] E. K. Miller and T. J. Buschman, "Cortical circuits for the control of attention," *Current opinion in neurobiology*, vol. 23, no. 2, pp. 216–222, 2013.

[58] D. Amso and G. Scerif, "The attentive brain: insights from developmental cognitive neuroscience," *Nature Reviews Neuroscience*, vol. 16, no. 10, pp. 606–619, 2015.

[59] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 350–363, 2013.

[60] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex.," *Cerebral cortex (New York, NY: 1991)*, vol. 1, no. 1, pp. 1–47, 1991.

[61] F. Crick, "Function of the thalamic reticular complex: the searchlight hypothesis," *Proceedings of the National Academy of Sciences*, vol. 81, no. 14, pp. 4586–4590, 1984.

[62] C. D. Gilbert and T. N. Wiesel, "Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex," *Nature*, vol. 280, no. 5718, pp. 120–125, 1979.

[63] C. D. Gilbert and T. N. Wiesel, "Clustered intrinsic connections in cat visual cortex," *Journal of Neuroscience*, vol. 3, no. 5, pp. 1116–1133, 1983.

[64] C. GILBERT and T. N. Wiesel, *Columnar specificity of intrinsic horizontal and corticocortical connection in cat visual cortex*. Society for Neuroscience, 1989.

[65] C. D. Gilbert and T. N. Wiesel, "The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat," *Vision research*, vol. 30, no. 11, pp. 1689–1701, 1990.

[66] W. S. Geisler and D. G. Albrecht, "Visual cortex neurons in monkeys and cats: detection, discrimination, and identification," *Visual neuroscience*, vol. 14, no. 05, pp. 897–919, 1997.

[67] M. Carandini, D. J. Heeger, and J. A. Movshon, "Linearity and normalization in simple cells of the macaque primary visual cortex," *Journal of Neuroscience*, vol. 17, no. 21, pp. 8621–8644, 1997.

[68] S. F. Bowne, "Contrast discrimination cannot explain spatial frequency, orientation or temporal frequency discrimination," *Vision Research*, vol. 30, no. 3, pp. 449–461, 1990.

[69] D. C. Somers, S. B. Nelson, and M. Sur, "An emergent model of orientation selectivity in cat visual cortical simple cells," *Journal of Neuroscience*, vol. 15, no. 8, pp. 5448–5465, 1995.

[70] E. A. DeYoe and D. C. Van Essen, "Concurrent processing streams in monkey visual cortex," *Trends in neurosciences*, vol. 11, no. 5, pp. 219–226, 1988.

[71] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth-anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, 1988.

[72] J. H. MAUNSELL and C. J. MCADAMS, "22 effects of attention on neuronal response properties in visual cerebral cortex," *The new cognitive neurosciences*, p. 315, 2000.

[73] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[74] R. B. Tootell, A. M. Dale, M. I. Sereno, and R. Malach, "New images from human visual cortex," *Trends in neurosciences*, vol. 19, no. 11, pp. 481–489, 1996.

[75] J. Zihl, D. Von Cramon, and N. Mai, "Selective disturbance of movement vision after bilateral brain damage," *Brain*, vol. 106, no. 2, pp. 313–340, 1983.

[76] J. Zihl, D. Von Cramon, N. Mai, and C. Schmid, "Disturbance of movement vision after bilateral posterior brain damage," *Brain*, vol. 114, no. 5, pp. 2235–2252, 1991.

[77] M. Corbetta, F. M. Miezin, S. Dobmeyer, G. L. Shulman, and S. E. Petersen, "Selective and divided attention during visual discriminations of shape, color, and speed: functional anatomy by positron emission tomography," *Journal of neuroscience*, vol. 11, no. 8, pp. 2383–2402, 1991.

[78] M. S. Beauchamp, R. W. Cox, and E. A. Deyoe, "Graded effects of spatial and featural attention on human area mt and associated motion processing areas," *Journal of neurophysiology*, vol. 78, no. 1, pp. 516–520, 1997.

[79] K. M. O'Craven, B. R. Rosen, K. K. Kwong, A. Treisman, and R. L. Savoy, "Voluntary attention modulates fmri activity in human mt–mst," *Neuron*, vol. 18, no. 4, pp. 591–598, 1997.

[80] S. P. Gandhi, D. J. Heeger, and G. M. Boynton, "Spatial attention affects brain activity in human primary visual cortex," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 3314–3319, 1999.

[81] Y. B. Saalmann, I. N. Pigarev, and T. R. Vidyasagar, "Neural mechanisms of visual attention: how top-down feedback highlights relevant locations," *Science*, vol. 316, no. 5831, pp. 1612–1615, 2007.

[82] S. Zeki, J. Watson, C. Lueck, K. J. Friston, C. Kennard, and R. Frackowiak, "A direct demonstration of functional specialization in human visual cortex," *Journal of neuroscience*, vol. 11, no. 3, pp. 641–649, 1991.

[83] R. B. Tootell and J. B. Taylor, "Anatomical evidence for mt and additional cortical visual areas in humans," *Cerebral Cortex*, vol. 5, no. 1, pp. 39–55, 1995.

[84] D. E. Rumelhart, "A multicomponent theory of the perception of briefly exposed visual displays," *Journal of Mathematical Psychology*, vol. 7, no. 2, pp. 191–218, 1970.

[85] G. Rainer, W. F. Asaad, and E. K. Miller, "Selective representation of relevant information by neurons in the primate prefrontal cortex," *Nature*, vol. 393, no. 6685, pp. 577–579, 1998.

[86] T. Shallice and P. W. Burgess, "Deficits in strategy application following frontal lobe damage in man," *Brain*, vol. 114, no. 2, pp. 727–741, 1991.

[87] P. S. Goldman-Rakic, "Topography of cognition: parallel distributed networks in primate association cortex," *Annual review of neuroscience*, vol. 11, no. 1, pp. 137–156, 1988.

[88] J. Duncan, "Cooperating brain systems in selective perception and action.," 1996.

[89] J. D. Schall, "Visuomotor areas of the frontal lobe," in *Extrastriate cortex in primates*, pp. 527–638, Springer, 1997.

[90] J. S. Baizer, L. G. Ungerleider, and R. Desimone, "Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques," *Journal of Neuroscience*, vol. 11, no. 1, pp. 168–190, 1991.

[91] C. Colby, R. Gattass, C. Olson, and C. Gross, "Topographical organization of cortical afferents to extrastriate visual area po in the macaque: a dual tracer study," *Journal of Comparative Neurology*, vol. 269, no. 3, pp. 392–413, 1988.

[92] G. Stanton, C. Bruce, and M. Goldberg, "Topography of projections to posterior cortical areas from the macaque frontal eye fields," *Journal of Comparative Neurology*, vol. 353, no. 2, pp. 291–305, 1995.

[93] K. G. Thompson, N. P. Bichot, and J. D. Schall, "Dissociation of visual discrimination from saccade programming in macaque frontal eye field," *Journal of neurophysiology*, vol. 77, no. 2, pp. 1046–1050, 1997.

[94] T. Moore and K. M. Armstrong, "Selective gating of visual signals by microstimulation of frontal cortex," *Nature*, vol. 421, no. 6921, pp. 370–373, 2003.

[95] C. J. Bruce and M. E. Goldberg, "Primate frontal eye fields. i. single neurons discharging before saccades," *Journal of neurophysiology*, vol. 53, no. 3, pp. 603–635, 1985.

[96] C. E. Connor, J. L. Gallant, D. C. Preddie, and D. C. Van Essen, "Responses in area v4 depend on the spatial relationship between stimulus and attention," *Journal of Neurophysiology*, vol. 75, no. 3, pp. 1306–1308, 1996.

[97] R. Desimone and L. G. Ungerleider, "Neural mechanisms of visual processing in monkeys," *Handbook of neuropsychology*, vol. 2, pp. 267–299, 1989.

[98] M. Harries and D. Perrett, "Visual processing of faces in temporal cortex: Physiological evidence for a modular organization and possible anatomical correlates," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 9–24, 1991.

[99] E. T. Rolls, A. Cowey, and V. Bruce, "Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas [and discussion]," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 335, no. 1273, pp. 11–21, 1992.

[100] E. A. Buffalo, P. Fries, R. Landman, H. Liang, and R. Desimone, "A backward progression of attentional effects in the ventral stream," *Proceedings of the National Academy of Sciences*, vol. 107, no. 1, pp. 361–365, 2010.

[101] K. Tanaka, "Neuronal mechanisms of object recognition," *SCIENCE-NEW YORK THEN WASHINGTON-*, vol. 262, pp. 685–685, 1993.

[102] M. A. Basso and R. H. Wurtz, "Modulation of neuronal activity in superior colliculus by changes in target probability," *Journal of Neuroscience*, vol. 18, no. 18, pp. 7519–7534, 1998.

[103] J. M. Findlay and R. Walker, "A model of saccade generation based on parallel processing and competitive inhibition," *Behavioral and Brain Sciences*, vol. 22, no. 04, pp. 661–674, 1999.

[104] R. M. McPeek and E. L. Keller, "Deficits in saccade target selection after inactivation of superior colliculus," *Nature neuroscience*, vol. 7, no. 7, pp. 757–763, 2004.

[105] R. Desimone, M. Wessinger, L. Thomas, and W. Schneider, "Attentional control of visual perception: cortical and subcortical mechanisms," in *Cold Spring Harbor symposia on quantitative biology*, vol. 55, pp. 963–971, Cold Spring Harbor Laboratory Press, 1990.

[106] M. E. Goldberg, H. Eggers, and P. Gouras, "The oculomotor system," *Principles of neural science. 3ª ed. New York: Appleton & Lange*, pp. 660–676, 1991.

[107] F. Arcizet, K. Mirpour, and J. W. Bisley, "A pure salience response in posterior parietal cortex," *Cerebral Cortex*, vol. 21, no. 11, pp. 2498–2506, 2011.

[108] J. H. Fecteau and D. P. Munoz, "Salience, relevance, and firing: a priority map for target selection," *Trends in cognitive sciences*, vol. 10, no. 8, pp. 382–390, 2006.

[109] A. E. Ipata, A. L. Gee, M. E. Goldberg, and J. W. Bisley, "Activity in the lateral intraparietal area predicts the goal and latency of saccades in a free-viewing visual search task," *Journal of Neuroscience*, vol. 26, no. 14, pp. 3656–3661, 2006.

[110] E. Premereur, W. Vanduffel, and P. Janssen, "Functional heterogeneity of macaque lateral intraparietal neurons," *Journal of Neuroscience*, vol. 31, no. 34, pp. 12307–12317, 2011.

[111] E. Premereur, W. Vanduffel, P. R. Roelfsema, and P. Janssen, "Frontal eye field microstimulation induces task-dependent gamma oscillations in the lateral intraparietal area," *Journal of neurophysiology*, vol. 108, no. 5, pp. 1392–1402, 2012.

[112] R. M. SIEGEL, "Encoding of spatial location by posterior parietal neuro," 1985.

[113] F. A. Wilson, S. P. Scalaidhe, and P. S. Goldman-Rakic, "Dissociation of object and spatial processing domains in primate prefrontal cortex," *SCIENCE-NEW YORK THEN WASHINGTON-*, vol. 260, pp. 1955–1955, 1993.

[114] M. I. Posner, J. A. Walker, F. J. Friedrich, and R. D. Rafal, "Effects of parietal injury on covert orienting of attention," *Journal of neuroscience*, vol. 4, no. 7, pp. 1863–1874, 1984.

[115] M. J. Riddoch and G. W. Humphreys, "Perceptual and action systems in unilateral visual neglect," *Advances in psychology*, vol. 45, pp. 151–181, 1987.

[116] S. M. Szczepanski and Y. B. Saalmann, "Human fronto-parietal and parieto-hippocampal pathways represent behavioral priorities in multiple spatial reference frames," *Bioarchitecture*, vol. 3, no. 5, pp. 147–152, 2013.

[117] S. E. Petersen, D. L. Robinson, and J. D. Morris, "Contributions of the pulvinar to visual spatial attention," *Neuropsychologia*, vol. 25, no. 1, pp. 97–105, 1987.

[118] D. LaBerge and M. S. Buchsbaum, "Positron emission tomographic measurements of pulvinar activity during an attention task," *Journal of neuroscience*, vol. 10, no. 2, pp. 613–619, 1990.

[119] D. L. Robinson and S. E. Petersen, "The pulvinar and visual salience," *Trends in neurosciences*, vol. 15, no. 4, pp. 127–132, 1992.

[120] C. Baleydier and A. Morel, "Segregated thalamocortical pathways to inferior parietal and inferotemporal cortex in macaque monkey," *Visual neuroscience*, vol. 8, no. 05, pp. 391–405, 1992.

[121] R. A. Berman and R. H. Wurtz, "Signals conveyed in the pulvinar pathway from superior colliculus to cortical area mt," *Journal of Neuroscience*, vol. 31, no. 2, pp. 373–384, 2011.

[122] D. L. Robinson and S. E. Petersen, "Responses of pulvinar neurons to real and self-induced stimulus movement," *Brain research*, vol. 338, no. 2, pp. 392–394, 1985.

[123] L. M. Chalupa, "Visual function of the pulvinar," in *The neural basis of visual function*, vol. 4, pp. 141–159, CRC Press Boca Raton, 1991.

[124] P. S. Churchland, V. Ramachandran, and T. J. Sejnowski, "A critique of pure vision1," *OF THE JBRAIN*, p. 23, 1994.

[125] A. Kafkas and D. Montaldi, "Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity," *The Quarterly Journal of Experimental Psychology*, vol. 64, no. 10, pp. 1971–1989, 2011.

[126] M. I. Posner and Y. Cohen, "Components of visual orienting," *Attention and performance X: Control of language processes*, vol. 32, pp. 531–556, 1984.

[127] D. E. Irwin and G. J. Zelinsky, "Eye movements and scene perception: Memory for things observed," *Perception & Psychophysics*, vol. 64, no. 6, pp. 882–895, 2002.

[128] I. Van Der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Visual memory for fixated regions of natural images dissociates attraction and recognition," *Perception*, vol. 38, no. 8, pp. 1152–1171, 2009.

[129] Y. Pertzov, G. Avidan, and E. Zohary, "Accumulation of visual information across multiple fixations," *Journal of Vision*, vol. 9, no. 10, pp. 2–2, 2009.

[130] E. A. Buffalo, S. J. Ramus, R. E. Clark, E. Teng, L. R. Squire, and S. M. Zola, "Dissociation between the effects of damage to perirhinal cortex and area te," *Learning & Memory*, vol. 6, no. 6, pp. 572–599, 1999.

[131] J. R. Manns, C. E. Stark, and L. R. Squire, "The visual paired-comparison task as a measure of declarative memory," *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 12375–12379, 2000.

[132] R. E. Clark, S. M. Zola, and L. R. Squire, "Impaired recognition memory in rats after damage to the hippocampus," *The Journal of Neuroscience*, vol. 20, no. 23, pp. 8853–8860, 2000.

[133] R. R. Althoff and N. J. Cohen, "Eye-movement-based memory effect: a reprocessing effect in face perception.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 25, no. 4, p. 997, 1999.

[134] D. E. Hannula, R. R. Althoff, D. E. Warren, L. Riggs, N. J. Cohen, and J. D. Ryan, "Worth a glance: using eye movements to investigate the cognitive neuroscience of memory," *Frontiers in human neuroscience*, vol. 4, no. 166, 2010.

[135] J. D. Ryan, R. R. Althoff, S. Whitlow, and N. J. Cohen, "Amnesia is a deficit in relational memory," *Psychological Science*, vol. 11, no. 6, pp. 454–461, 2000.

[136] C. N. Smith, R. O. Hopkins, and L. R. Squire, "Experience-dependent eye movements, awareness, and hippocampus-dependent memory," *The Journal of neuroscience*, vol. 26, no. 44, pp. 11304–11312, 2006.

[137] C. N. Smith and L. R. Squire, "Experience-dependent eye movements reflect hippocampus-dependent (aware) memory," *The Journal of Neuroscience*, vol. 28, no. 48, pp. 12825–12833, 2008.

[138] M. J. Jutras, P. Fries, and E. A. Buffalo, "Oscillatory activity in the monkey hippocampus during visual exploration and memory formation," *Proceedings of the National Academy of Sciences*, vol. 110, no. 32, pp. 13144–13149, 2013.

[139] K. L. Hoffman, M. C. Dragan, T. K. Leonard, C. Micheli, R. Montefusco-Siegmund, and T. A. Valiante, "Saccades during visual exploration align hippocampal 3–8 hz rhythms in human and non-human primates," *Eye movement-related brain activity during perceptual and cognitive processing*, p. 80, 2014.

[140] M. J. Jutras and E. A. Buffalo, "Recognition memory signals in the macaque hippocampus," *Proceedings of the National Academy of Sciences*, vol. 107, no. 1, pp. 401–406, 2010.

[141] G. Kreiman, C. Koch, and I. Fried, "Category-specific visual responses of single neurons in the human medial temporal lobe," *Nature neuroscience*, vol. 3, no. 9, pp. 946–953, 2000.

[142] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, no. 11, pp. 1254–1259, 1998.

[143] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[144] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, pp. 545–552, 2006.

[145] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 153–160, 2013.

[146] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of vision*, vol. 14, no. 1, pp. 28–28, 2014.

[147] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in neural information processing systems*, pp. 241–248, 2008.

[148] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using adaboost," *Journal of Vision*, vol. 12, no. 6, pp. 22–22, 2012.

[149] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, pp. 155–162, 2005.

[150] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *NIPS*, pp. 2204–2212, 2014.

[151] T. S. Lee and X. Y. Stella, "An information-theoretic framework for understanding saccadic eye movements.," in *NIPS*, pp. 834–840, 1999.

[152] L. W. Renninger, J. M. Coughlan, P. Verghese, and J. Malik, "An information maximization model of eye movements," in *NIPS*, pp. 1121–1128, 2004.

[153] X. Sun, H. Yao, and R. Ji, "What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency," in *IEEE CVPR*, pp. 1552–1559, IEEE, 2012.

[154] H. Liu, D. Xu, Q. Huang, W. Li, M. Xu, and S. Lin, "Semantically-based human scanpath estimation with hmms," in *ICCV*, pp. 3232–3239, 2013.

[155] S. Mathe and C. Sminchisescu, "Action from still image dataset and inverse optimal control to learn task specific visual scanpaths," in *NIPS*, pp. 1923–1931, 2013.

[156] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[157] M. Jiang, X. Boix, G. Roig, J. Xu, L. Van Gool, and Q. Zhao, "Learning to predict sequences of human visual fixations," *IEEE Transaction On Neural Networks And Learning Systems*, 2016.

[158] S. O. Ba and J.-M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *TPAMI*, vol. 33, no. 1, pp. 101–116, 2011.

[159] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *CVPR, 2012*, pp. 470–477, 2012.

[160] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?," in *CVPR*, pp. 5501–5510, 2015.

[161] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *TPAMI*, vol. 39, no. 5, pp. 893–907, 2017.

[162] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *CVPR, 2013*, pp. 1147–1154, IEEE, 2013.

[163] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," *arXiv preprint arXiv:1603.08199*, 2016.

[164] D. M. Beck and S. Kastner, "Top-down and bottom-up mechanisms in biasing competition in the human brain," *Vision research*, vol. 49, no. 10, pp. 1154–1165, 2009.

[165] L. G. Williams, "The effects of target specification on objects fixated during visual search," *Acta psychologica*, vol. 27, pp. 355–360, 1967.

[166] J. M. Findlay, "Saccade target selection during visual search," *Vision research*, vol. 37, no. 5, pp. 617–631, 1997.

[167] M. V. Peelen and S. Kastner, "A neural basis for real-world visual search in human occipitotemporal cortex," *Proceedings of the National Academy of Sciences*, vol. 108, no. 29, pp. 12125–12130, 2011.

[168] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A bayesian inference theory of attention," *Vision research*, vol. 50, no. 22, pp. 2233–2247, 2010.

[169] A. J Yu and P. Dayan, "Uncertainty, neuromodulation, and attention," *Neuron*, vol. 46, no. 4, pp. 681–692, 2005.

[170] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2956–2964, 2015.

[171] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2147–2154, 2014.

[172] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[173] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[174] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European conference on computer vision*, pp. 584–599, Springer, 2014.

[175] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

[176] A. Haji-Abolhassani and J. J. Clark, "A computational model for task inference in visual search," *Journal of vision*, vol. 13, no. 3, pp. 29–29, 2013.

[177] M. Cerf, J. Harel, A. Huth, W. Einhäuser, and C. Koch, "Decoding what people see from where they look: Predicting visual stimuli from scanpaths," in *International Workshop on Attention in Cognitive Systems*, pp. 15–26, Springer, 2008.

[178] T. O'Connell and D. Walther, "Fixation patterns predict scene category," *Journal of Vision*, vol. 12, no. 9, pp. 801–801, 2012.

[179] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 921–928, IEEE, 2013.

[180] U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis," *Journal of Vision*, vol. 6, no. 4, pp. 7–7, 2006.

[181] M. R. Greene, T. Liu, and J. M. Wolfe, "Reconsidering yarbus: A failure to predict observers task from eye movement patterns," *Vision research*, vol. 62, pp. 1–8, 2012.

[182] A. Borji, A. Lennartz, and M. Pomplun, "What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations," *Neurocomputing*, vol. 149, pp. 788–799, 2015.

[183] G. J. Zelinsky, Y. Peng, and D. Samaras, "Eye can read your mind: Decoding gaze fixations to reveal categorical search targets," *Journal of vision*, vol. 13, no. 14, pp. 10–10, 2013.

[184] Y. LeCun and M. Ranzato, "Deep learning tutorial," in *Tutorials in International Conference on Machine Learning (ICML13)*, Citeseer, 2013.

[185] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[186] R. M. Klein, "Inhibition of return," *Trends in cognitive sciences*, vol. 4, no. 4, pp. 138–147, 2000.

[187] Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," 2014.

[188] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *European Conference on Computer Vision*, pp. 17–32, Springer, 2014.

[189] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision research*, vol. 116, pp. 165–178, 2015.

[190] G. Kootstra, B. de Boer, and L. R. Schomaker, "Predicting eye fixations on complex visual stimuli using local symmetry," *Cognitive computation*, vol. 3, no. 1, pp. 223–240, 2011.

[191] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 5, pp. 802–817, 2006.

[192] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, "What do saliency models predict?," *Journal of vision*, vol. 14, no. 3, pp. 14–14, 2014.

[193] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *ECCV 2010*, (Crete, Greece).

[194] N. Bruce and J. Tsotsos, "Attention based on information maximization," *Journal of Vision*, vol. 7, no. 9, pp. 950–950, 2007.

[195] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *IEEE ICCV*, pp. 921–928, IEEE, 2013.

[196] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[197] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, pp. 32–32, 2008.

[198] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[199] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[200] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[201] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[202] T. Jost, N. Ouerhani, R. Von Wartburg, R. Müri, and H. Hügli, "Assessing the contribution of color in visual attention," *Computer Vision and Image Understanding*, vol. 100, no. 1, pp. 107–123, 2005.

[203] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv preprint arXiv:1604.03605*, 2016.

[204] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[205] T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *CVPR*, pp. 41–48, 2009.

[206] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *ECCV*, pp. 314–327, 2012.

[207] Z. Mengmi, M. Keng Teck, L. Joo Hwee, Z. Qi, and F. Jiashi, "Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks," in *CVPR, 2017*, IEEE, 2017.

[208] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *arXiv preprint arXiv:1609.02612*, 2016.

[209] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[210] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[211] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR, 2009*, pp. 2929–2936, IEEE, 2009.

[212] S. Mathe and C. Sminchisescu, "Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition," *TPAMI*, vol. 37, no. 7, pp. 1408–1424, 2015.

[213] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: state-of-the-art and study of comparison metrics," in *ICCV*, pp. 1153–1160, 2013.

[214] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *ICML*, pp. 233–240, 2006.

[215] P.-H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *JoV*, vol. 9, no. 7, pp. 4–4, 2009.

[216] V. Buso, I. González-Díaz, and J. Benois-Pineau, "Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos," *Signal Processing: Image Communication*, vol. 39, pp. 418–431, 2015.

[217] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *CVPR*, pp. 2714–2721, 2013.

[218] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, pp. 818–833, 2014.

[219] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *CVPR*, pp. 287–295, 2015.

[220] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *arXiv preprint arXiv:1412.0767*, 2014.

[221] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[222] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2005.

[223] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, pp. 267–285, Springer, 1982.

[224] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[225] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[226] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[227] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[228] M. Handford, *Where's Waldo?* Little, Brown Boston, 1987.

[229] J. S. Horst and M. C. Hout, "The novel object and unusual name (noun) database: A collection of novel images for use in experimental research," *Behavior research methods*, vol. 48, no. 4, pp. 1393–1409, 2016.

[230] I. Gauthier and M. J. Tarr, "Becoming a greeble expert: Exploring mechanisms for face recognition," *Vision research*, vol. 37, no. 12, pp. 1673–1682, 1997.

[231] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist, "Scanmatch: A novel method for comparing fixation sequences," *Behavior research methods*, vol. 42, no. 3, pp. 692–700, 2010.

[232] T. S. Horowitz, "Revisiting the variable memory model of visual search," *Visual Cognition*, vol. 14, no. 4-8, pp. 668–684, 2006.

[233] G. J. Zelinsky, "A theory of eye movements during target acquisition.," *Psychological review*, vol. 115, no. 4, p. 787, 2008.

[234] C.-C. Wu, H.-C. Wang, and M. Pomplun, "The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes," *Vision research*, vol. 105, pp. 10–20, 2014.

List of Author's Publications

1. **Mengmi Zhang**, Jiashi Feng, Joo Hwee Lim, Qi Zhao, Gabriel Kreiman, "What am I searching for", International Conference on Inteilligent Robots and Systems **(IROS)**, 2019 **(to submit)**.

2. **Mengmi Zhang**, Jiashi Feng, Karla Montejo, Joseph Kwon, Joo Hwee Lim, Gabriel Kreiman, "Lift-the-Flap: Context Reasoning Using Object-Centered Graphs", IEEE Conference on Computer Vision and Pattern Recognition **(CVPR)**, 2019 **(under review)**.

3. **Mengmi Zhang**, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, Jiashi Feng, "Anticipating Where People Will Look Using Adversarial Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence **(TPAMI)**, 2018 **(impact factor: 9.455)**.

4. **Mengmi Zhang**, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman, "Finding any Waldo with zero-shot invariant and efficient visual search", **Nature Communications**, 2018 **(impact factor: 12.353)**.

5. **Mengmi Zhang**, Keng Teck Ma, Shih-Cheng Yen, Joo Hwee Lim, Qi Zhao, and Jiashi Feng, "Egocentric Spatial Memory", International Conference on Intelligent Robots and Systems **(IROS)**, 2018.

6. **Mengmi Zhang**, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, "Foveated Neural Network: Gaze Prediction on Egocentric Videos", IEEE International Conference on Image Processing **(ICIP)**, 2017.

7. **Mengmi Zhang**, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, Jiashi Feng, "Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Net-

works", IEEE Conference on Computer Vision and Pattern Recognition **(CVPR)**, 2017 **(oral presentation)**.