THE ROLE OF ATTENTION ON SCENE RECOGNITION

BERHAN ŞENYAZAR

BOĞAZİÇİ UNIVERSITY

2017

THE ROLE OF ATTENTION ON SCENE RECOGNITION

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Cognitive Science

by

Berhan Şenyazar

Boğaziçi University

2017

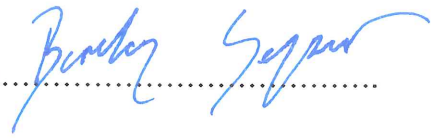
DECLARATION OF ORIGINALITY

I, Berhan Şenyazar, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;

- this thesis contains no material that has been submitted or accepted for a degree or a diploma in any other educational institution;

- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature…  ……

Date 23.08.2017

ABSTRACT

The Role of Attention on Scene Recognition

It has been shown that little if any attention is required for scene recognition (Li, VanRullen, & Koch, 2002). The absence of the role of attention in scene recognition, however, has been challenged by Cohen, Alvarez, and Nakayama (2011) showing that basic-level scene categorization and object identification performance degrade while simultaneously performing an attention-demanding task. Here, we use the same dual-task paradigm but in conjunction with a more reliable psychophysical method (Greene & Oliva, 2009a) to measure and compare scene recognition performance in different conditions of a broad range of scene recognition tasks, including detection, recognition of spatial structure and scene function, superordinate- and basic-level categorizations. Analysis of minimum duration at which the percentage of correct answers reached 75% showed a threshold increase in scene recognition performance from single- to dual-task conditions, suggesting a respective degradation in scene recognition performance. The performance of the secondary multiple-object tracking task also got worse in dual-task condition, implying that scene recognition and multiple-object tracking tasks may share an attentional capacity resource. A computational model was used to test whether a feedforward model lacking attentional modulation can account for our findings and the results showed that human scene recognition performance fits to the predictions of the model only in the dual-task conditions, where the attentional mechanisms are already occupied to facilitate scene recognition. For scene images categorized as "hard to be recognized" by the model in the single task blocks, however, the

behavioral performance did not change, providing evidence for a potential attentional facilitation.

ÖZET

Sahne Tanımada Dikkatin Rolü

Li, VanRullen, ve Koch'un 2002 yılında yaptıkları çalışma sahne tanımada dikkatin rolünün çok az, belki de hiç olmadığını göstermiştir. Sahne tanımada dikkatin rolünün olmadığını ileri süren bu görüşle çelişir şekilde Cohen, Alvarez, ve Nakayama (2011), basit-anlam seviyesinde sahne sınıflandırması ve nesne tanıma performansının aynı anda dikkat gerektiren başka bir görev ile birlikte gerçekleştirildiğinde düştüğünü göstermiştir. Biz bu tezde, Cohen ve grubunun kullanmış olduğu ikili-görev paradigmasını daha güvenilir (Greene & Oliva, 2009a) bir psikofizik yöntemi ile bir arada kullanarak sahne tanıma performansını farklı dikkat yükü durumlarında ve sahne tespit etme, sahnenin uzamsal yapısı ve işlevini tanıma, ve yüksek- ve basit-anlam seviyelerinde sınıflandırma gibi geniş bir aralıktaki sahne tanıma görevleri bağlamında ölçüp karşılaştırdık. Sonuçlar, anlamlı sahne uyaranlarını tanımada doğru yanıtların %75'e ulaştığı en düşük gösterim süresinin tekli-görev durumdan ikili-görev duruma geçişte arttığını ve buna bağlı olarak sahne tanıma performansının düştüğünü gösterdi. Benzer bir düşüşün ikincil görev olarak sunulan çoklu-nesne takibi performansında da gözlemlenmesi, sahne tanıma ve çoklu-nesne takibi görevlerinin aynı dikkat kaynağını paylaşabiliyor olabileceğine işaret etti. Çalışmamızın son aşamasında, literatürde var olan bir bilişimsel model kullanarak, dikkat modülasyonu içermeyen ileri beslemeli bir modelin davranışsal sonuçları ne kadar açıklayabileceğini test ettik. Sonuçlar, davranışsal sahne tanıma performansının ileri beslemeli modelin tahminlerine sadece ikili-görev durumlarında, dikkat mekanizmaları halihazırda meşgulken uyduğunu, tekli-ödev durumlarındaysa modelin “zor” addettiği sahneler için davranışsal

performansta düşüşün olmadığını ve büyük ihtimalle dikkat mekanizmalarının devreye girmiş olabileceğini gösterdi.

ACKNOWLEDGEMENTS

I am extremely grateful to my advisors İnci Ayhan and Albert Ali Salah for the independence they allowed and the guidance they provided. Your confidence in me from the beginning has been invaluable for me and sincerely, shaped my future.

I gratefully acknowledge the support from The Scientific and Technological Research Council of Turkey (TÜBİTAK) as part of the 2210/B National Scholarship Programme for MA Students.

As in every aspect of my life, the support of my mother Nazlı Şenyazar, my father Turan Şenyazar and my sister Gizem Şenyazar Altunordu gave me the confidence and power to keep on my master's studies. I cannot thank you enough for your love, and describe how lucky I feel thanks to you.

For bearing my endless anxiety, firing my imagination, and giving meaning to my life with their friendships, I want to express my deepest gratitude to Özgür Gök, Ecem Baysal, Ozan Barlas, and Gupse Korkmaz. You have made me feel alive.

The precious times I spent together with Doğa Gülhan, Elif Canseza Kaplan, Erdem Ozan Meral, Gizem Ünlü, Melissa Kurtcan, and Umay Şen were, perhaps, the most important thing I gained in this period. It was the best of times, it was the worst of times...

The support of the members of VisionLab and the working environment they created were essential and I feel indebted to them. Especially, I want to thank Doğa Gülhan for making everything easier, nicer and more joyful. I also have to thank Ayşenur Akyüz for her timely help on data collection.

Finally, I owe a huge debt of thanks to my patient and benevolent participants. Your efforts made everything possible.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: EXPERIMENT 1: THE EFFECT OF DUAL-TASK ON SCENE RECOGNITION TASKS 13

2.1. Methods 17

2.2. Results 23

CHAPTER 3: EXPERIMENT 2: COMPARING PERFORMANCE PATTERNS OF A FEEDFORWARD MODEL AND HUMAN PARTICIPANTS 30

3.1. Methods 32

3.2. Results 34

CHAPTER 4: DISCUSSION 39

CHAPTER 5: CONCLUSION 46

APPENDIX A: INDIVIDUAL DATA OF EXPERIMENT 1 47

APPENDIX B: INDIVIDUAL DATA OF EXPERIMENT 2 48

APPENDIX C: TOP-DOWN EXPECTATIONS EXPERIMENT 49

APPENDIX D: TEMPORAL ATTENTION EXPERIMENT 51

REFERENCES 53

LIST OF FIGURES

Fig. 1 Examples of synthesized masking images..... 19

Fig. 2 The general procedure of Experiment 1 21

Fig. 3 Minimum presentation duration thresholds of Experiment 1 24

Fig. 4 d' values of Experiment 1 26

Fig. 5 MOT results of Experiment 1 29

Fig. 6 Distribution of posterior probabilities of beach images 33

Fig. 7 Minimum presentation thresholds of Experiment 2 35

Fig. 8 d' values of Experiment 2 37

Fig. 9 MOT results of Experiment 2 38

LIST OF APPENDIX FIGURES

Fig. A1 Individual psychometric functions of Experiment 1 47

Fig. B1 Individual psychometric functions of Experiment 2..... 48

Fig. C1 The procedure of the top-down expectations experiment..... 49

Fig. C2 Minimum presentation durations of the top-down expectations experiment50

Fig. D1 The procedure of the temporal attention experiment..... 51

Fig. D2 Minimum presentation durations of the temporal attention experiment 52

CHAPTER 1

INTRODUCTION

We have a rich visual experience of the world around us. While the information our visual system extracts from the outer world consists of contours, edges, luminance- and special-contrasts, our phenomenal experience is shaped by the semantic context of the complex combinations of those basic visual features. To describe our surroundings, for example, we might name nearby objects such as a computer, a desk or a coffee cup. Another way of describing such scene, however, could be to name the whole scenery as a "study room", which would implicitly tell us more than having a list of relevant objects there. In the visual recognition literature, the term *scene recognition* corresponds to the latter approach. A scene is defined as a view of some part of the world containing objects, surfaces and background elements in such a meaningful arrangement that it creates together a namable entity (Henderson & Hollingworth, 1999; Oliva, 2013). A proposed simple heuristic to distinguish a scene from an object is that while we take action *on* an object, we act *within* a scene (Epstein, 2005). Scene recognition, in its most general definition, is the identification of a semantic category, also called a gist, of a scene (Malcolm, Groen, & Baker, 2016).

The names we use in daily language to describe scenes have been defined as basic-level categorization (Rosch, 1973; Rosch & Mervis, 1975). At basic-level, there is a high amount of similarity amongst the members of a category. Discriminability from other categories, however, is also at its highest, which together might explain why basic-level categories, such as a forest or a store, are the most commonly used categorical descriptions (Rosch, Mervis, Gray, & Johnson, 1976).

Common attributes of some basic-level categories constitute another higher-level, more abstract categorical description that is called superordinate-level categorization. A forest and a beach, for example, can be categorized together as natural scenes, whereas a store and a playground may be the members of the opposite superordinate-level category, namely as man-made or urban scenes. Among the members of a basic-level category, such as a store, we might also make further specifications like a rainforest or grocery store, with even more similarity in the specified sub-category. Such finer distinctions are called as subordinate-level categorizations. It has been claimed that categorization starts with a distinction at basic-level, an abstraction process over which results in a later superordinate-level categorization (Rosch et al., 1976). Recent studies, however, found that superordinate-level natural/man-made distinction can be made even before basic-level categorization, suggesting that there is a precedence of superordinate- over basic-level categorization (Loschky & Larson, 2010; Kadar & Ben-Shahar, 2012). The formation of categorizations at different levels was also studied by a free-recall study in which participants were presented with scene images for randomly changing brief presentation durations. The results of this study, where participants were asked to describe freely what they had just seen showed that these hierarchical scene categories manifest themselves also in behavioral reports (Fei-Fei, Iyer, Koch, & Perona, 2007). Greene et al. have proposed another level of scenes description in which spatial structures, such as openness, or functions, such as navigability, of scenes might be mid-level global properties mediating scene recognition (Greene & Oliva, 2009b; Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016). Detection of the presence of a scene has also been suggested to be a distinct step in the recognition process (Mack & Palmeri, 2010), although, it has been reported that categorization is accomplished at the same

moment as detection (Grill-Spector & Kanwisher, 2005). These studies together suggest the following five general scene recognition tasks at different levels: basic- and superordinate-level categorization, scene spatial structure and function recognition, and detection of the presence of a scene.

Considering the large number of components in a scene that define its category, such as objects, surfaces, and three-dimensional structures, scene recognition may be seen as a very complex and time-demanding task. The seminal study of Potter and Levy (1969), however, showed that people are able to recognize scenes by viewing them for only 125 ms, suggesting that scene understanding can be completed in a single glance. Thorpe, Fize, and Marlot (1996) have further investigated the processing speed of recognition using a go/no-go task in an ERP paradigm setup. In their study, photographs of natural scenes were serially presented for 20 ms without a following masking and participants were required to give a response when the target category (animal) was present (go response). Remarkably, mean accuracy was 94%, demonstrating that the minimum presentation duration required for recognition was even shorter than 125 ms if not followed by a mask. The analysis of ERP data also showed that the earliest significant difference between the potentials generated on the go and no-go tasks was around 150 ms, which further supports that the categorical discrimination had already been made by that time. In visual recognition studies, one paradigm to prevent the processing of a stimulus further in the system is visual backward masking, where researchers introduce a structurally similar masking image following the target onset. This allows researchers to control for the presentation duration of a stimulus more accurately (Breitmeyer & Ögmen, 2000). Bacon-Macé, Macé, Fabre-Thorpe, and Thorpe (2005), for example, have studied scene recognition performance in a go/no-go task

using a backward masking method, where the presentation duration was fixed at 6 ms and the manipulation was made on the stimulus onset asynchrony (SOA) of masking images. The results showed that participants were able to recognize scenes at around 40 ms SOA with an accuracy over 75%, suggesting that such a brief interval was sufficient for scene recognition. Instead of measuring the scene recognition performance on different scene recognition tasks with a fixed presentation duration, Greene and Oliva (2009a) introduced a different experimental regime where they measured the minimum presentation durations required to reach a 75% accuracy rate using a psychophysical methodology. In their study, participants were presented with a scene image for varying levels of presentation durations and the scene image was immediately followed by four masking images. The percentage of correct answers for each presentation duration were then fit into a function (Weibull function), previously reported to fit well on psychometric data (Klein, 2001), in order to determine the minimum presentation duration thresholds as the point that corresponded to 75% accuracy. Using this paradigm, Greene and Oliva (2009a) found mean threshold value to be 34 ms for global scene properties and 50 ms for basic-level categorization. The results were comparable to the 40 ms SOA condition of Bacon-Macé et al. (2005), not unexpectedly, considering that in both studies, participants were left with approximately the same duration for processing images.

Such a rapid recognition of scenes presents a rather different picture than the theories of visual recognition which suggests that the visual system reaches at a more complex and abstract representation through a hierarchy consisting of following the steps, respectively: extraction of the basic image features, composition of those features into object representations, understanding relations among these object

representations, and a holistic representation of a scene (Biederman, 1972, 1987, 1976). As an alternative to this hierarchical explanation of scene recognition, it has been proposed that it might rather be the global features of images which facilitate scene recognition (Oliva & Schyns, 1997; Oliva & Torralba, 2001; Torralba & Oliva, 2003). Images can be represented, as any other two-dimensional signal, as the sum of sinusoidal signals following Fourier's theorem. The low-spatial frequency signals represent components of an image that change on a larger spatial scale such as large objects or the spatial structure of a scene, whereas high-spatial frequency signals represent finer details in an image. Single-cell recording studies have shown that neurons in the visual system have different spatial frequency preferences (Webster & De Valois, 1985). Following this physiological evidence, Oliva and Schyns (1997) tested at what extent low- and high-spatial frequency information are used during scene recognition. The authors used hybrid stimuli, which consisted of images with different category information or noise at different spatial frequency bands. Participants were able to recognize scenes using the information at all frequency bands, suggesting that the visual system may use information at different spatial frequencies in parallel using specialized spatial channels. Recognition of scene categories without object recognition has also been supported by a computational model that used global spatial features of a scene such as naturalness, openness and depth to understand a scene category (Oliva & Torralba, 2001). Global spatial features were estimated from the dominant orientations and the rate of change at each orientation, the parameters calculated, respectively, from the energy spectra, which is the global distribution of the amplitudes of the sinusoidal signals, and the energy spectrograms, which are the localized energy distributions, of the scene images. It has been shown that such low-dimensional, global image features could accurately

discriminate scene categories, might be computed in a rapid, feedforward manner and may facilitate object recognition by providing context (Torralba & Oliva, 2003).

Supportingly, neuroscientific studies have revealed scene-selective regions in the brain that process scenes from global features rather than following an object-based approach. The first observation was on "parahippocampal place area" (PPA), which had a greater response to scenes than single objects and faces (Epstein & Kanwisher, 1998). Another area with scene-selective response has been demonstrated to be "retrosplenial complex" (RSC), which showed a stronger response to familiar scenes and thus, was considered to play a role in navigation (Epstein, Higgins, Jablonski, & Feiler, 2007). Similarly, impairing activity on "occipital place area" (OPA) using a transcranial magnetic stimulation (TMS) resulted in a decreased performance on recognition of scenes but not of faces or objects, suggesting a critical role of OPA in scene recognition (Dilks, Julian, Paunov, & Kanwisher, 2013). A multivoxel pattern analysis (MVPA) showed that activity in PPA and RSC might in fact predict scene categories and account for human performance at the behavioral level (Walther, Caddigan, Fei-Fei, & Beck, 2009). In the same study, activity in V1 and lateral occipital complex (LOC), a region which was reported to have object selectivity, also showed significant predictive power, suggesting that to some extent, basic image features might be sufficient to discriminate scene categories (Oliva & Torralba, 2001) and that there might also be a complementary role of object information in scene recognition (Quattoni & Torralba, 2009).

Computational studies suggest that such a rapid recognition of scenes without object information could be accomplished by using global image features (Oliva & Torralba, 2001, 2002; Torralba & Oliva, 2003) and texture analysis (Renninger &

Malik, 2004) in a feedforward manner. It has been shown that bio-inspired feedforward models of object recognition could in fact account for human performance in rapid recognition tasks (Serre, Oliva, & Poggio, 2007; Serre, Kreiman, Kouh, & Cadieu, 2007). Recent goal-driven computational models, which have a feedforward architecture similar to that of previous models, have reached near-human performance in object recognition (LeCun, Bengio, & Hinton, 2015; Guo et al., 2015) and have been shown to have similar representations to those at human ventral stream (Guclu & van Gerven, 2015; Yamins, Hong, Cadieu, & DiCarlo, 2013; Cadieu et al., 2014), which together suggest that a feedforward architecture could account for rapid visual recognition. In addition, it has also been shown that human performance patterns can be accounted by scene-based models with similar features and architecture (Xiao, Ehinger, Hays, Torralba, & Oliva, 2014; Xiao et al., 2013; Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014).

The very rapid nature of scene recognition, together with relevant computational studies provide consistent evidence for a parallel, feedforward and low attention demanding account of scene recognition (Fabre-Thorpe, 2011). Lack of attentional effects in behavioral experiments further supports this account of scene recognition (Rousselet, Fabre-Thorpe, & Thorpe, 2002; Li et al., 2002; Greene & Fei-Fei, 2014). Rousselet et al. (2002), for example, tested whether recognition could be performed in parallel in a similar manner to low-level image features. In a go/no-go task, the authors presented participants with either a peripheral single image or two images corresponding to different visual fields and asked them to respond when a target category is present (go task). While there was a performance decrease in two image condition, mean reaction times and d′ scores over reaction times were found to be very similar. Analysis of ERP data did not reveal a difference in the activation

patterns in single and two image conditions, either, implying that a high-level process such as scene recognition could be performed in a parallel, feedforward manner, at least across different brain hemispheres. Presenting two scenes, but on different visual fields, however, may not be the right approach to test the attentional limits on scene recognition as attentional resources might be distributed in a hemisphere-specific manner.

Li et al. (2002) has suggested that scene recognition is robust to dual-task interference. In their paradigm, participants were asked to perform a centrally presented digit discrimination task while trying to recognize scenes presented in the periphery. They found that performances on scene recognition task were similar in both single- and dual-task conditions, suggesting that scene recognition may not require attentional resources. In this study, Li et al. (2002) controlled that the central task demanded enough attention by testing low-level letter and color discriminations in the periphery, the results of which showed a decrease in the peripheral task performances from single- to dual-task conditions. Their design, however, did still not allow to reach a reliable conclusion that scene recognition requires no attention as one might argue that whereas central digit discrimination task allocates a significant amount of attention from the common resources shared with other discrimination tasks, it may not be using similar resources with scene recognition, a task completely different in nature.

Greene and Fei-Fei (2014) provided further evidence that would support a rather low-attention-demanding account of scene recognition in a study where they tested the automaticity of scene recognition using a Stroop-like paradigm. Here, participants were asked to categorize the words presented at the center of the screen as object names or scene names at basic-level categories. Images of isolated objects

and scenes were presented on a background and the participants were instructed to focus their attention only on the word categorization task. Greene and Fei-Fei (2014) found a significant difference between the performances in word categorization task in congruent (scene image and name of the scene) and incongruent trials (scene image and name of another scene), implying that scene recognition was automatically performed at the expense of a performance decrease on the main task. This interference effect, however, was not observed when participants were asked to perform the same task in a trial where central words were rather at superordinate-level categorization (adjective or noun), presented on a background with incongruent scene images. The authors interpreted these results such that scene recognition require minimal effort and is obligatory at the basic- but not at the superordinate-level. The uncontrolled presentation durations of the images in the study, however, makes it hard to reach a certain conclusion. The mean reaction times were around 750 ms and images were shown until a response was obtained from the participants. Compared to the previous studies in which presentation durations were on a range from 6 to 50 ms, long presentation durations around 750 ms might have allowed participants to accomplish the background scene recognition task by dedicating relatively low attentional resources — be consciously or out of awareness.

Despite the aforementioned accumulated evidence that support a feedforward, low-attention demanding account of scene recognition, there are also some studies reporting that in some conditions scene recognition performance could be significantly impaired (Rousselet, Thorpe, & Fabre-Thorpe, 2004b; Evans & Treisman, 2005; Cohen et al., 2011). Using a go/no-go method, Rousselet et al. (2004b), for example, studied the limits of parallel processing in scene recognition by asking participants to report the existence of a target scene (containing an animal)

among an array of 1, 2 and 4 scene images presented at a fixed 26 ms presentation duration. Behavioral data of this study showed a decrease in performance with an increasing array size. The authors, however, have proposed this decrease could still be accounted by a feedforward, parallel model of scene processing. In signal detection theory, responses are classified according to the true value of the stimuli. If a go response (an animal is present) is given when a target stimulus is actually present, it is called a hit. In contrast, false alarms are the go responses when a target image is absent. Here, by taking into account the different response patterns to targets (proportion of hits) and distractors (proportion of false alarms), together with the increased number of distractors in 2 and 4 array size conditions, the authors adjusted the accuracies of the multiple image conditions to that of the single image condition and observed no difference in the adjusted accuracy rates. The ERP analysis on occipital regions showed a reduced differential activity for the condition where the array was composed of 4 images compared to those with only 1 and 2 images conditions. Interestingly, the location of the activation has also followed the positions of the images, indicating a parallel, retinotopic processing of scenes, rather than a limited recognition process. Taken together, the authors concluded that the performance decrease in conditions with multiple images might be the result of a late response selection mechanism rather than an early-stage attentional limit as was suggested by the parallel, feedforward account of scene recognition.

A different account of the possible source of attentional limit has been given by Evans and Treisman (2005). In a series of attentional blink experiments, the authors have shown that the recognition performance of a scene is decreased in conditions where another scene is to be recognized shortly before the target, with only a brief interstimulus interval in between (varying between 220 to 880 ms),

suggesting that the recognition performance of the second scene might have declined because of a temporal recovery limit, where it takes time for the system to direct the attentional resources already allocated to the first image onto the subsequent one. The authors claimed that the source of the attentional limit in this task was at a stage where low-level features extracted in a parallel manner were bound together to build higher-level representations. This interpretation, in contrast to the feedforward model proposed by Rousselet et al. (2004b), assigns a role to the attention in the recognition process.

On the basis of a theoretical assumption that awareness should require attention, Cohen et al. (2011) hypothesized that scene recognition performance in the previous studies not because scene recognition does not require attention but because the secondary tasks used in the literature may not have been attentionally challenging enough to reduce resources allocated to the scene recognition tasks significantly. In inattentional blindness paradigm of Cohen et al. (2011), participants were presented with a background scene image, of which no prior notice has been given. On top of those background images were moving objects which were to be tracked as part of a well-known multiple-object tracking (MOT) task, or alternatively, was a sequence of rapidly flashed numbers and letters as part of a rapid serial visual presentation (RSVP) task, both of which are known to be highly attention demanding tasks. In MOT task, participants were asked to track the position of four target discs among four identical distractor discs, whereas in RSVP task, the participants counted the number of digits in a stream of letters and digits. Results demonstrated that most of the participants did not detect the scene image on the background, a finding controversial to what a low attention demanding model of scene recognition would predict. In another experiment, to test the hypothesis that scene recognition

performance might be degraded with a sufficiently challenging dual-task paradigm, Cohen et al. (2011) used two MOT conditions manipulating disc speeds as slower and faster, a manipulation which has been reported that makes MOT task more difficult, hence require more attentional resources (Alvarez & Franconeri, 2007). As predicted by their hypothesis, Cohen et al. (2011) showed a decrease in the scene recognition performance only in the more attention demanding condition (faster discs) compared to the baseline conditions. In this study, they also showed that the accuracy of MOT task was significantly degraded in the difficult condition but not in the easy condition, providing evidence that scene recognition and MOT tasks might in fact share a common attentional resource. To control that the performance difference was due to a higher attentional resource rather than a lower-level motion component of the MOT task, a similar dual-task experiment was run using an RSVP paradigm, the results of which provided further evidence that both the scene recognition and RSVP performances were degraded in dual- compared to single-task conditions.

In the present study, we use a reliable psychophysical methodology to measure and compare scene recognition performance in different conditions and test the role of attention, for the first time in the literature, on a broad range of scene recognition tasks (basic- and superordinate-level categorization, scene spatial structure and function recognition, detection of the presence of a scene). Using a state-of-the-art computational model, we also test whether feedforward models, that lack any attentional modulation, can account for our behavioral data.

CHAPTER 2

EXPERIMENT 1: THE EFFECT OF DUAL-TASK ON SCENE RECOGNITION TASKS

In Experiment 1, using a large range of scene recognition tasks, including detection, recognition of spatial structure and scene function, superordinate- and basic-level categorizations, we aim to test whether attention affects the minimum duration at which a scene can be recognized. One way of testing the attentional demand of a task is to use a dual-task paradigm (Braun & Julesz, 1998; Pashler, 1994), where attention is modeled as the allocation of a limited resource for a given task. In dual-task paradigms, it is hypothesized that if a target task requires attention, then the resulting performance should get degraded as the capacity of the limited resource is shared with a competing secondary task. Studies using dual-task paradigm reported that while processing of some low-level visual features such as orientation and color is completed in a parallel manner without a significant attentional cost, processing of higher-level, feature-bound combinations that, at the end, produce an object or a scene shows performance degradation in dual-task conditions (McElree & Carrasco, 1999; Treisman, 1998). In the context of these results, one might expect to see an attentional cost of a high-level task, where participants are asked to recognize complex natural scene images. Surprisingly, though, it has been documented that the gist of scenes can be obtained at even brief presentation durations (Thorpe et al., 1996). This led some researchers to conceptualize scene recognition as a rather feedforward and low attention demanding task (Fabre-Thorpe, 2011). This view was also empirically supported by studies showing the robustness of scene recognition performance in parallel processing (Rousselet et al., 2002) and dual-task conditions

(Li et al., 2002). In one of these studies, where participants were asked to do a main task while passively viewing the secondary scene recognition task on the background, Greene and Fei-Fei (2014) have demonstrated a drop in the accuracy with which participants completed the main task, implying that scene recognition not only does not demand an intentional effort, it is in fact an automatic operation processed spontaneously.

Conceptualization of scene recognition as a low attention demanding, automatic task, however was challenged by a number of studies using attention demanding parallel performance conditions (Rousselet, Thorpe, & Fabre-Thorpe, 2004a), attention blink in rapid serial visual presentation (RSVP) (Evans & Treisman, 2005) and high attention demanding dual-task conditions (Cohen et al., 2011). In the study of Rousselet et al. (2004a), for example, increase in the number of concurrently presented distractor scenes led to a decrease in the recognition performance of a target category (i.e. animal) in the array. Seemingly controversial to the automatic processing account of scene recognition, the authors suggested that the decrease in performance could result from a late selection mechanism, where visual features and semantic information might still be automatically processed in parallel at an early categorization level but that the selection of a target in a particular category might be subject to a later-stage, spatially-selective attentional limit. A similar performance degradation, but this time in temporal domain was observed by Evans and Treisman (2005) using an attentional blink paradigm in RSVP. In their interpretation, however, the source of this attentional limitation was at a stage where low-level features are bound together even before a categorization is made, an account which assigns attention a critical role on scene recognition. Such interpretation argues against the low attention-demanding parallel accounts of scene

recognition. Cohen et al. (2011) suggested that the reason scene recognition might appear to be robust to the dual-task interference might be because employed tasks in the literature do not require a sufficient amount of attentional capacity sharing. To test this hypothesis, they used a multiple-object tracking (MOT) paradigm with varying degrees of difficulty by changing the speed at which the objects moved (Alvarez & Franconeri, 2007). While the scene recognition performance did not differ between the slow-motion-speed MOT condition and the single-task baseline, the performances in both scene recognition and MOT tasks had decreased in the fast-motion-speed MOT condition, suggesting that scene recognition and MOT tasks may in fact share a common attentional resource.

There is no comprehensive study in the literature, which looked at the effect of attention on scene recognition using the same paradigm at all conceptual levels including detection, recognition of spatial envelope and scene function, superordinate- and basic-level categorizations. Inconsistent usage and comparisons of these various tasks in scene recognition literature made it harder to reach a reliable conclusion from previous studies. As it has been suggested that there might be a hierarchical processing between different conceptual levels of scene recognition, mainly superordinate- (the most abstract, e.g. natural vs. artificial) and basic-level (the most common usage, e.g. mountain vs. beach) categorizations (see, Fabre-Thorpe, 2011). Thus, here, we tested a range of tasks while studying the effect of attention on scene recognition using a single reliable method. In one variation of the hierarchical processing hypotheses of scene recognition, it is claimed that coarse visual information is processed in a fast, parallel route facilitating superordinate-level categorization, which is then followed by a more resource demanding stage that results in basic-level categorization (Macé, Joubert, Nespoulous, & Fabre-Thorpe,

2009; Poncet & Fabre-Thorpe, 2014). Some empirical evidence, however, was contradictory to the prediction of such account in that it showed an obligatory automaticity for scene recognition at basic- but not at superordinate-level (Greene & Fei-Fei, 2014). As well as these contradictions in the literature, there are also some missing links. Although mid-level features such as spatial properties and scene functions/affordances, for example, were proposed to facilitate scene recognition (Greene & Oliva, 2009b; Greene et al., 2016; Oliva & Torralba, 2001); no study was conducted, though, in order to investigate attentional modulation at those levels. Another debate in scene recognition literature is on whether detection of the presence of a scene is a distinct process (Mack & Palmeri, 2010) or whether it is an operation coupled with categorization (Grill-Spector & Kanwisher, 2005). In Experiment 1, therefore, we tested the role of attention including all of these five scene recognition tasks (basic- and superordinate-level categorization, scene spatial structure and function recognition, detection of the presence of a scene) to make a relevant comparison.

Most of the previous studies in the literature follow a similar methodology to measure the effect of dual-task interference. In this paradigm, single- and dual-task performances are compared in terms of accuracy or reaction time for a fixed-length presentation duration. From a psychophysical point of view, this paradigm might be misleading because it does not take into account the shape of the performance curve as a function of different presentation durations but rather relies on a single sampling point. This approach would produce reliable results only if — by any chance — the selected interval happens to fall in a small range around the 75% accuracy threshold in the relevant task. Following (Greene & Oliva, 2009a), here, we employ a more reliable measure of comparison by determining the minimum duration at which a

scene can be recognized, finding for each experimental condition, the 75% accuracy thresholds separately. Using this method, for the first time in literature, we make a reliable standardized comparison of the involvement of attention in different scene recognition tasks.

2.1. Methods

2.1.1 Participants

Six psychophysically trained Boğaziçi University members participated, four of whom were naive to the purpose of the experiment (two other participants were the author and one of the supervisors). The study was conducted in accordance with the Declaration of Helsinki and approved by the Boğaziçi University Human Research Ethics Committee. All participants had normal or corrected-to-normal vision.

2.1.2 Apparatus

Experiment was conducted in a dark room using a Philips 109B40/20 CRT monitor (1024 × 768 pixels screen resolution at 85 Hz refresh rate). A chin rest was used to keep head steady at a viewing distance of 45 cm. Stimuli were generated in the Matlab environment (The Mathworks, Natick, MA) using Psychtoolbox (Brainard & others, 1997; Pelli, 1997; Kleiner et al., 2007) and responses were collected using a common keyboard.

2.1.3. Stimuli

105 target and 105 distractor scene images were used for each task. Images were selected from SUN Database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010) and

SUN Attribute Database (Patterson & Hays, 2017). Categories and attributes determined as targets were namely: beach category for basic-level categorization, naturalness attribute for superordinate-level categorization, sports attribute for scene function recognition, and openness attribute for scene spatial structure recognition (Table 1). For detection task, random images from SUN Attribute Database that were not selected for other tasks were used as targets and masking images were the distractors. All images were resized to 256 × 256 pixels. At a distance of 45 cm and given spatial resolution, images subtended 10.9 × 10.9 degrees of visual angle.

Table 1. Selected Image Categories for Scene Recognition Tasks

Task	Target	Distractor
Scene Function	Highest Sports	Lowest Sports
Spatial Structure	Highest Openness	Lowest Openness
Superordinate-level Categorization	Highest Natural	Highest Man-made
Basic-level Categorization	Beach	Mountain
Scene Detection	Unused Images	Masking Images

In order to limit sensory processing of the images, a paradigm with both backward (Bacon-Macé et al., 2005) and forward maskings (Cohen et al., 2011) was used. For each scene recognition task, selected images from that particular category or attribute of the database were synthesized to generate masking images using a procedure introduced by Greene and Oliva (2009a) following the parametric texture model of Portilla and Simoncelli (2000). This model aims to capture the transformations applied to a stimulus in the early stages of the visual system by extracting a set of statistics from the image. In the synthesis phase of this model, the

input image is first linearly decomposed using multi-orientation, multi-scale filter outputs and a set of marginal statistics, coefficient correlations, magnitude correlation and cross-scale phase statistics are collected. The collected statistics are then imposed on a noise image iteratively, which finally results in a synthesized image maintaining the statistics of the input image, yet decomposed in terms of its semantic content (Fig. 1).

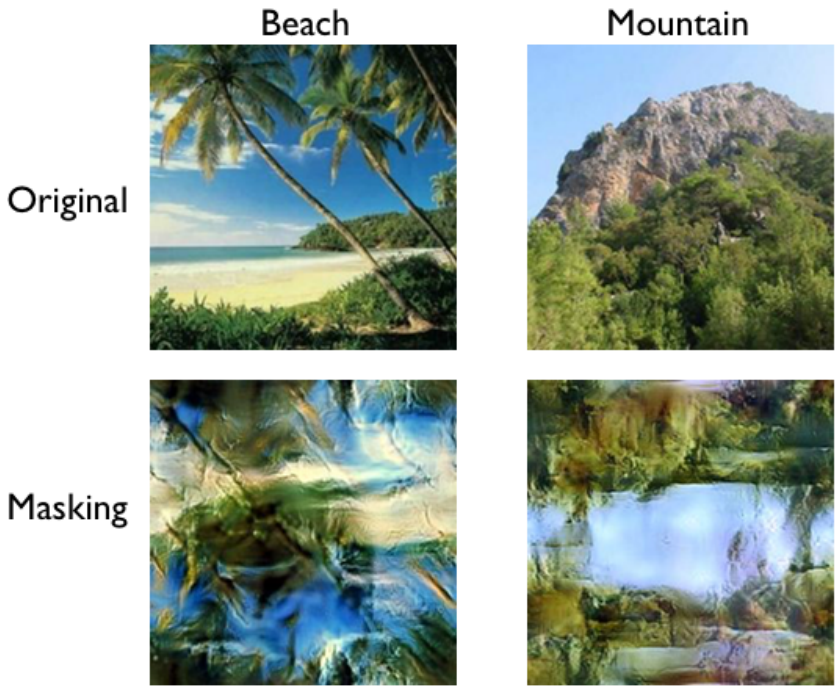


Fig. 1 Examples of synthesized masking images

During MOT task, 6 identical white discs with a radius of 1 deg moved with a constant velocity of 10.5 deg/sec in a 7 deg-radius circular zone at the center of the screen. The initial positions of the discs were assigned in such a way that each would be presented in a random location within the specified circular region, 1 deg away from the center of the screen without any overlap with the rest. First moved in random directions, their direction of motion would change after each collusion, either

with another disc or with the outer boundaries of the circular zone. In addition to the visible six, there were also three invisible discs to make visible discs to change direction suddenly without an apparent collusion.

2.1.4. Procedure

Participants completed five scene recognition tasks. Each scene recognition task was composed of 2 blocked-trial conditions, in which participants focused either on a single scene recognition task with MOT discs totally ignored on the screen or on dual-tasks, making decision on both scene recognition and MOT tasks at the end of each trial. As a control condition, participants were also asked to complete two single MOT tasks, where attention was supposed to be given to the MOT task only. Together these made a total of 12 experimental conditions, the order of which was counterbalanced across participants.

Before each block, participants were first given the procedural instructions and the description of the relevant target category. They were then asked to press a key on the keyboard to proceed and read onscreen instructions. Instructions were followed by the examples of possible target and distractor images. 28 training trials were introduced to get participants used to the procedure. Each trial started with the presentation of a gray screen, on which six identical discs were randomly located. Three of these discs flickered for two seconds to signal them as to-be-tracked discs. As soon as the discs began moving, a train of masking images appeared sequentially on the background, each with a presentation duration of 117.6 ms (10 frames). The durations of the trials were chosen randomly from a range of 3 to 6 seconds and in each trial, a target or distractor scene image was presented before the last two masking images (235.2 ms) (see Fig. 2). The presentation duration of the scene

images varied in seven levels (11.8, 23.5, 35.3, 47, 70.1, 94, 117.6 ms) using the method of constant stimuli to generate a psychometric function indicating the minimum duration at which a scene category or attribute was recognized.

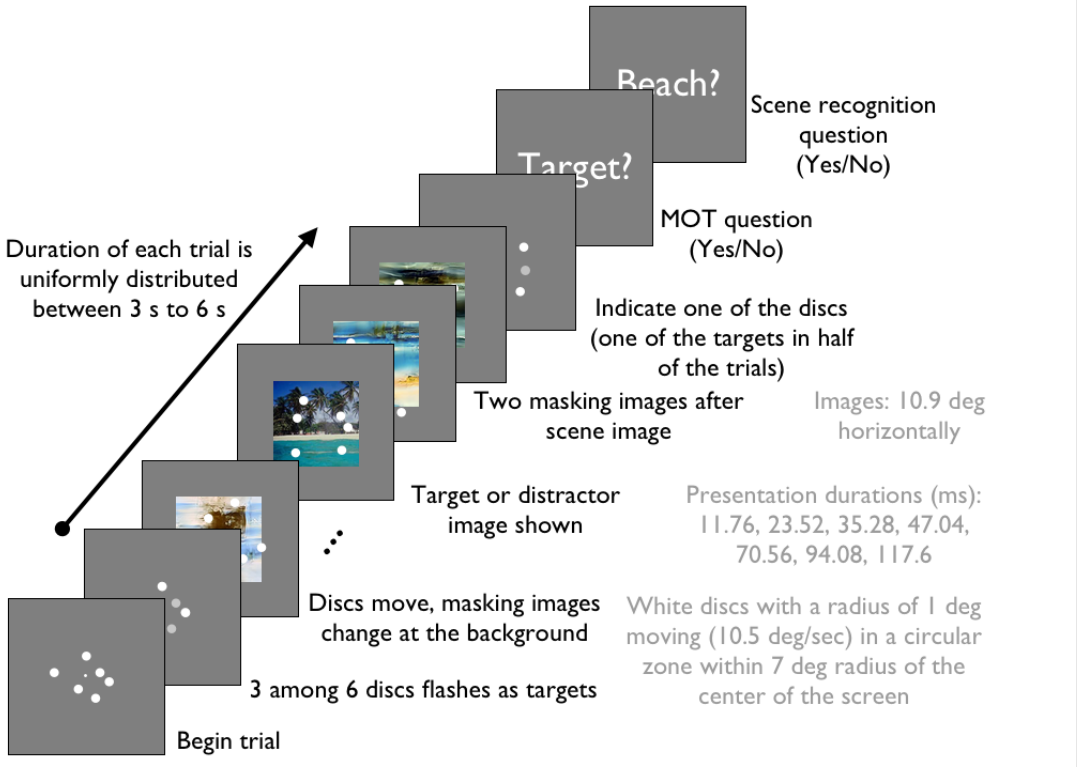


Fig. 2 The general procedure of Experiment 1

At the end of a trial, discs stopped altogether and one of them flickered for two seconds as a target MOT stimulus. In a 2-AFC task, subjects were asked to report whether the flickered disc was one of to-be-tracked discs and to make a scene recognition judgement as to whether scene image was a target or distractor. In half of the trials for each presentation duration, the scene image was from the target category and the target disc was from the group of to-be-tracked discs. Feedback (a "wrong answer" text) was presented at the end of the trials, where participant's decision was not correct. The 75% point on the psychometric function provided an

estimate of the duration thresholds for scene recognition. Each data point, to which a Weibull function was fit, was composed of 30 trials. As a result, each block consisted of 210 trials (7 × 30) in total.

For single-task scene recognition, at the beginning of each block, participants were instructed to ignore the moving discs and focus solely on the background task to recognize the scene image as a target or a distractor. They were informed that the scene image would always going to be presented before the last two masking images of the sequence but that the trial durations and thus, the number of masking images would be random. They were also informed that the presentation duration of the scene images would going change randomly across trials. At the end of each trial, they were expected to report whether the scene image belonged to the target category or not.

For single-task MOT blocks, participants were instructed to ignore the background images and follow to-be-tracked discs with a sustained attention. They were informed that the discs were identical with a constant speed, and that their movement were limited within a 7 deg-radius region at the center of the screen. The task of participants was to report whether the target disc flashed at the end of a trial was amongst the three to-be-tracked discs or not.

For dual-task blocks, participants were instructed to focus on both the MOT and the scene recognition tasks simultaneously. Decisions with regards to scene recognition and MOT tasks were semi-randomly collected to ensure balance across different presentation durations and target/distractor image trials.

Participants were allowed to have breaks during a block to keep themselves alert. They were also required to have at least 15 minutes rest across different blocks.

2.2. Results

If there was a role of attention on scene recognition, one would expect a significant difference between minimum presentation duration thresholds in single- and dual-task conditions of a scene recognition task. Following dual-task paradigm, if this difference is the result of an attentional capacity sharing between the two tasks, single- and dual-task MOT performances would also be expected to differ. Here, the results of Experiment 1 satisfied both of these two criteria, which together suggests that there is a role of attention on scene recognition.

In the data analysis, minimum presentation duration thresholds were determined using the Palamedes toolbox (Prins & Kingdon, 2009). First, we calculated d′ values for each presentation duration. Because d′ values are unbiased sensitivity measures, we could convert them into unbiased percentage of correct responses. Finally, the percentage of correct responses were fit into a Weibull function (see Fig. A1) to determine the minimum duration at which the percentage of correct responses reached 75% point as the minimum presentation duration threshold for each task.

Statistical analyses were conducted in R (R Core Team, 2016) using "ez" package (Lawrence, 2016). We reported effect sizes in both partial eta squared and generalized eta squared (η_G^2; Olejnik & Algina, 2003), the latter of which is a more suitable effect size measure for comparisons between experiments, especially for repeated measures designs (Bakeman, 2005). A 5 × 2 repeated measures ANOVA on minimum presentation duration threshold values showed that to complete scene recognition tasks in dual-task conditions (M = 87.8 ms, SD = 9.0), a longer presentation duration was required than in single-task conditions (M = 72.4 ms, SD = 16.1), $F(1, 5) = 7.45$, $MSe = 475.67$, $p = .041$, $\eta_p^2 = .60$, $\eta_G^2 = .19$ (Fig. 3). There was

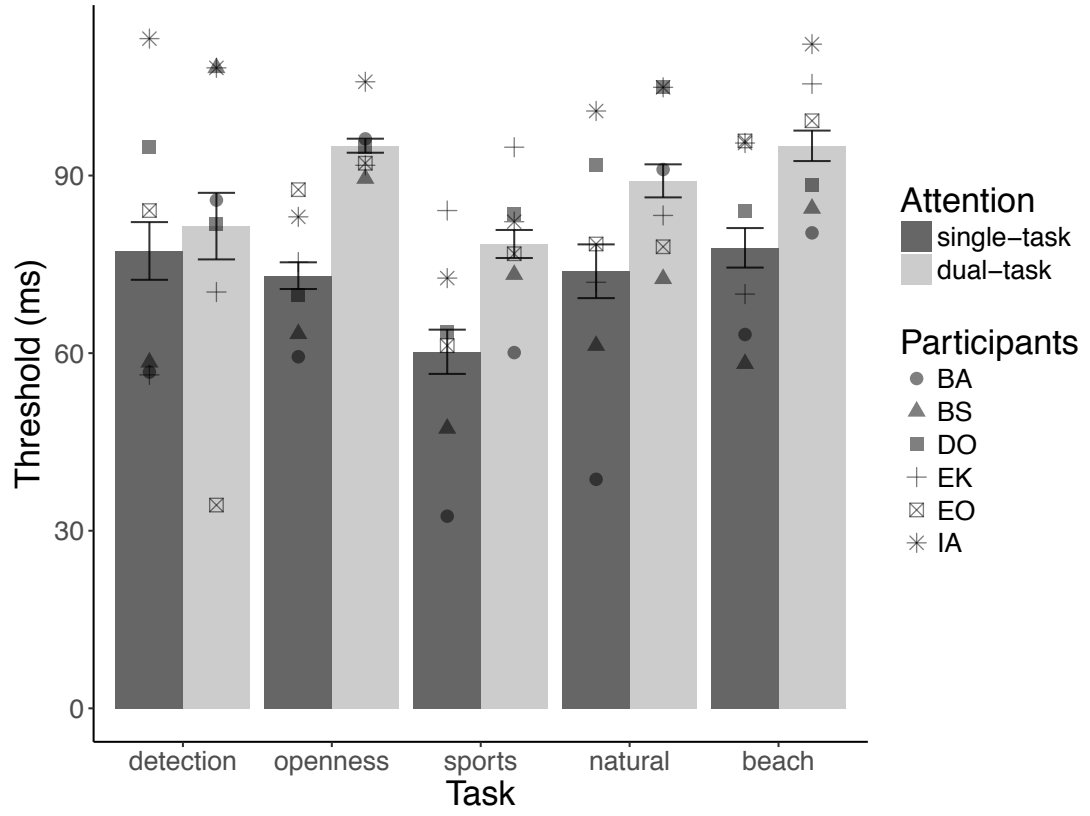


Fig. 3 Minimum presentation duration thresholds of Experiment 1

Note. Bar graph shows means of thresholds and symbols show individual data points of participants. Error bars indicate the standard error of the mean. While a paired t-test showed that the effect of attention was absent in detection task, $p > .05$, a repeated analysis of ANOVA without the inclusion of detection task confirmed the main effect of attention, $p = .005$. The main effect of the type of scene recognition task was also significant in the ANOVA analysis, $p = .011$.

no significant difference, though, among minimum presentation durations required to complete different scene recognition tasks, $F(4, 20) = 2.33$, $MSe = 223.38$, $p > .05$, $\eta_p^2 = .32$, $\eta_G^2 = .12$. The interaction effect between attention and different recognition conditions was not significant, either, $F(4, 20) = .23$, $MSe = 118.81$, $p > .05$, $\eta_p^2 = .18$, $\eta_G^2 = .03$.

As the nature of the detection task was different than that of the other recognition tasks and the error bars of the box plots for the single- and dual-task conditions of the detection task showed a significant overlap (Fig. 3), we ran a separate t-test for that particular condition regardless of a non-significant interaction between the attention (single-task; double-task) and the type of task.

The paired t-test showed that, the difference between single- (M = 77.3 ms, SD = 23.9) and dual-task (M = 81.5 ms, SD = 27.5) conditions were indeed not significant, $t(5) = -.29$, $p > .05$. We repeated our analysis by excluding detection task. A 4 × 2 repeated measures ANOVA showed an increased effect of attention compared to the previous analysis between dual- (M = 89.8 ms, SD = 8.0) and single-task (M = 71.2 ms, SD = 15.1) conditions, $F(1, 5) = 22.59$, $MSe = 175.28$, $p = .005$, $\eta_p^2 = .82$, $\eta_G^2 = .31$. In contrary to the previous analysis, without detection task, there was also a significant difference among presentation durations thresholds of scene recognition tasks, $F(3, 15) = 5.27$, $MSe = 131.29$, $p = .011$, $\eta_p^2 = .51$, $\eta_G^2 = .19$. A post-hoc pairwise comparison analysis using Bonferroni correction on significance level (alpha = .05 / 6 = .008) showed that scene function categorization (sports task) (M = 69.3 ms, SD = 14.8) required significantly less presentation duration than basic-level categorization (beach task) (M = 86.4 ms, SD = 13.2), $p = .001$; and scene spatial structure recognition (openness task) (M = 84.1 ms, SD = 7.0), $p = .002$. The interaction effect between attention and different recognition conditions was again not significant, $F(3, 15) = .42$, $MSe = 56.06$, $p > .05$, $\eta_p^2 = .08$, $\eta_G^2 = .01$.

We also calculated d′ values for each task by pooling all data points of presentation durations to measure difficulty and precision A 5 × 2 repeated-measures ANOVA on d′ values confirmed that the difficulty of dual-task conditions (M = .9, SD = .2) was higher than that of single-task conditions (M = 1.1, SD = .3), $F(1, 5)$ =

18.16, *MSe* = .04, p = .008, η_p^2 = .78, η_G^2 = .14. The main effect on the difficulty of different scene-recognition tasks was also significant, $F(4, 20) = 3.04$, *MSe* = .04, p = .041, η_p^2 = .38, η_G^2 = .11 (Fig. 4). A post-hoc pairwise comparison analysis using Bonferroni correction on significance level (alpha = .05 / 10 = .005) showed that

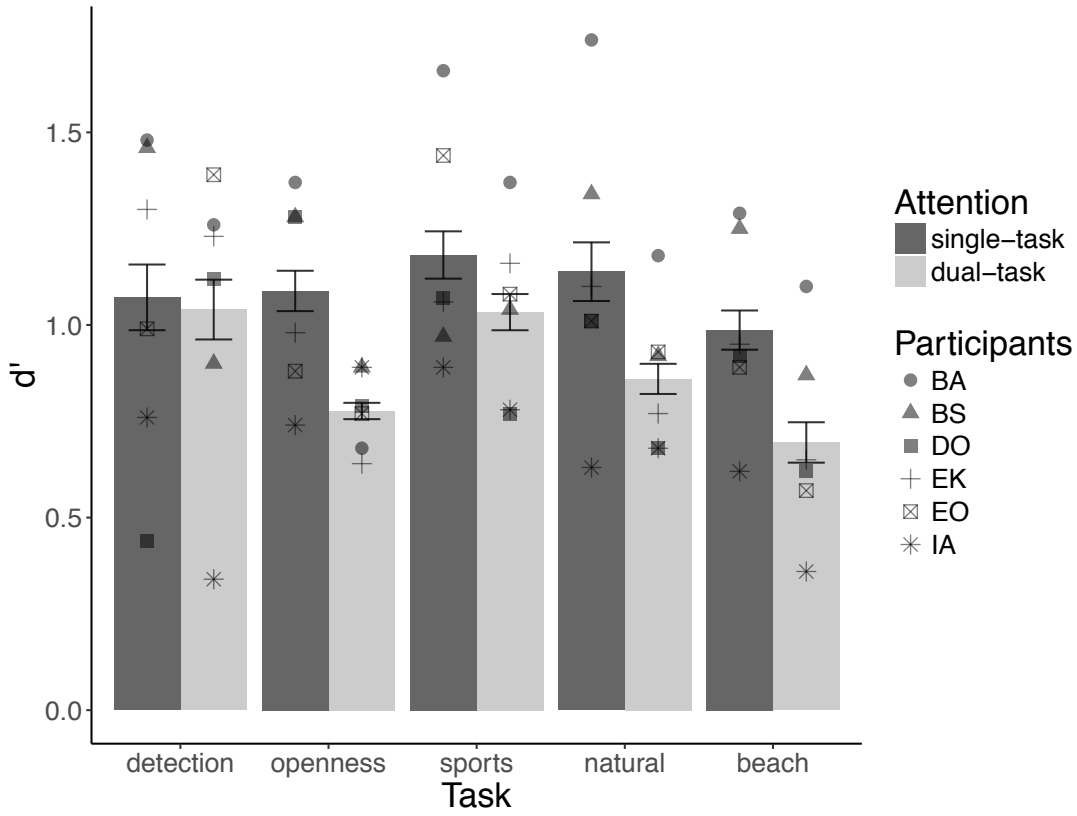


Fig. 4 d' values of Experiment 1

Note. Bar graph shows means of d' values and symbols show individual data points of participants. Error bars indicate the standard error of the mean. There were main effects of attention, p = .008, and task, p = .041. Beach task was significantly more difficult than natural and sports tasks.

basic-level categorization (beach task) (M = .8, SD = .3) was significantly more difficult than superordinate-level categorization (natural task) (M = 1.0, SD = .3), p =

.002; and scene function recognition (sports task) (M = 1.1, SD = .2), p = .002. There was no other significant difficulty difference across tasks. The interaction between the difficulty in different attention and recognition conditions was not statistically significant, either, $F(4, 20) = .17$, $MSe = .04$, $p > .05$, $\eta_p^2 = .17$, $\eta_G^2 = .04$.

Similar to minimum presentation duration analysis, we further investigated detection task because of the overlapping error bars in Fig. 4. A paired t-test showed that difference of d' values between single- and dual-task conditions for detection task was also not significant, $t(5) = 0.16$, $p > .05$. A 4 × 2 repeated measures ANOVA without detection task again showed an increased effect of attention compared to the previous analysis between dual- (M = .8, SD = .1) and single-task (M = 1.1, SD = .3) conditions, $F(1, 5) = 22.09$, $MSe = .04$, $p = .005$, $\eta_p^2 = .82$, $\eta_G^2 = .23$. As in the previous d' analysis, the main effect of task was significant, $F(3, 15) = 4.08$, $MSe = .04$, $p = .027$, $\eta_p^2 = .45$, $\eta_G^2 = .15$. A post-hoc pairwise comparison analysis using Bonferroni correction on significance level (alpha = .05 / 6 = .008) again showed that basic-level categorization (beach task) (M = .8, SD = .3) was significantly more difficult than superordinate-level categorization (natural task) (M = 1.0, SD = .3), p = .002; and scene function recognition (sports task) (M = 1.1, SD = .1), p = .002. The interaction effect was again not significant, $F(3, 15) = .90$, MSe = .02, $p > .05$, $\eta_p^2 = .15$, $\eta_G^2 = .02$.

If the performance difference between single- and dual-task scene recognition conditions were due to the capacity sharing with the MOT task, a similar performance degradation would also be expected to be observed in MOT performance. Thus, we measured dual-task MOT performance together with five scene recognition tasks. A one-way repeated measures ANOVA showed no significant difference in MOT accuracy when it was performed concurrently with

different scene recognition tasks, $F(4, 20) = .42$, $MSe < .01$, $p > .05$, $\eta_p^2 = .08$, $\eta_G^2 = .01$, a result consistent with the threshold analysis which suggested that different scene recognition tasks do not require significantly different amount of attentional resources. We had conducted two single-task MOT conditions, one with images from a detection and the other with images from a basic-level categorization (beach) task. Supporting previous analyses, a paired t-test showed that the MOT performance in these two conditions did not significantly differ, $t(5) = 1.96$, $p > .05$. Since the difference between the MOT performance across conditions was not significant, we binned their data to compare single- and dual-task MOT performance. A paired t-test showed that the percentage of correct answers in a single-task MOT condition ($M = .87$, $SD = .05$) was significantly higher than that in a dual-task condition ($M = .84$, $SD = .05$), $t(5) = 3.89$, $p = .01$ (Fig. 5). That MOT performance decreased together with scene recognition performance in dual-task condition confirmed that both tasks share a common attentional capacity.

Previously, we ran the same experiment without randomizing the order in which the scene recognition and the MOT questions were asked, and without control blocks for the MOT task. Since such a design could produce memory effects and not allow the comparisons of the MOT performances, we modified our procedure and presented its results above, which we consider more reliable. The results of the older experiment were in fact quite similar to the modified version. A 5 × 2 repeated measures ANOVA on minimum presentation duration threshold values showed that dual-task conditions required significantly more presentation duration, $F(1, 5) = 13.46$, $MSe = 1.74$, $p = .014$, $\eta_p^2 = .73$. No significant difference was observed among minimum presentation durations required to complete different scene recognition

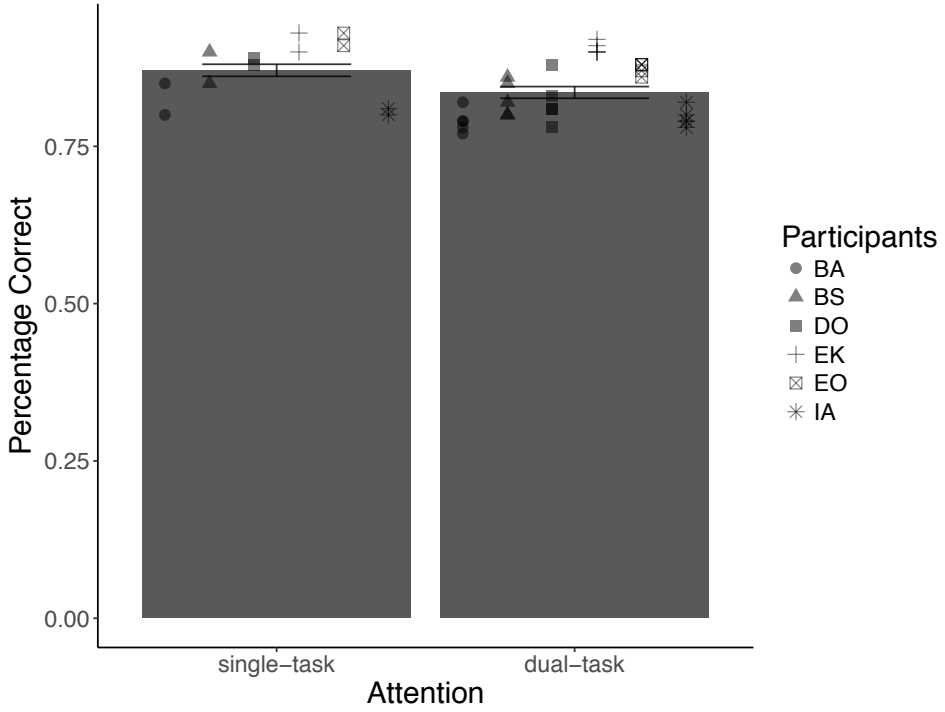


Fig. 5 MOT results of Experiment 1

Note. Bar graph shows means of the percentage of correct answers for MOT task and symbols show individual data points of participants. Error bars indicate the standard error of the mean. There was a main effect of attention, $p = .01$.

tasks, $F(4, 20) = 2.14$, $MSe = 1.50$, $p > .05$, $\eta_p^2 = .30$. The interaction effect between attention and different recognition conditions was not significant, either, $F(4, 20) = .68$, $MSe = .91$, $p > .05$, $\eta_p^2 = .12$.

The analysis of d′ values with a 5 × 2 repeated-measures ANOVA showed that the difficulty of dual-task conditions was higher than that of single-task conditions in the previous experiment, too, $F(1, 5) = 21.19$, $MSe = .04$, $p = .006$, $\eta_p^2 = .81$. The main effect on the difficulty of different scene-recognition tasks, however, was not significant, $F(4, 20) = 2.38$, $MSe = .04$, $p > .05$, $\eta_p^2 = .32$. There was also no interaction effect, $F(4, 20) = .65$, $MSe = .03$, $p > .05$, $\eta_p^2 = .11$.

CHAPTER 3

EXPERIMENT 2: COMPARING PERFORMANCE PATTERNS OF A FEEDFORWARD MODEL AND HUMAN PARTICIPANTS

The results of Experiment 1 showed that there is a capacity sharing between the scene recognition and multiple-object tracking (MOT) tasks. This suggests that scene recognition is an operation requiring flexible visual attention resources similar to the MOT task (Alvarez & Franconeri, 2007; Cohen et al., 2011). Previous visual recognition studies, however, have provided dual-task (Li et al., 2002), reaction time (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Rousselet et al., 2002; Thorpe et al., 1996; VanRullen & Thorpe, 2016), presentation duration (Keysers, Xiao, Földiák, & Perrett, 2001), event-related potential (ERP) (Rousselet et al., 2002; Thorpe et al., 1996), spatial frequency analysis (Oliva & Schyns, 1997; Oliva & Torralba, 2001; Torralba & Oliva, 2003) evidence that support a massively parallel and feedforward account of scene recognition (Fabre-Thorpe, 2011).

Following the seminal findings of Hubel and Wiesel (1965), ventral stream of the visual system has been modeled as a hierarchical system with increasing complexity at each further stage to reach a final view-invariant categorical representation. Progress of a visual stimulus into these stages is called a feedforward pass. Computational models have been developed to simulate computations of neurons in the ventral stream with feedforward connections (Riesenhuber & Poggio, 1999; Fukushima, 1980). These feedforward models were also shown to account for human visual recognition performance (Serre, Oliva, & Poggio, 2007).

Recently, however, hierarchical convolutional neural networks (LeCun & Bengio, 1995) have become more popular as they could successfully model sensory

systems with a near-human performance (LeCun et al., 2015; Guo et al., 2015). These networks are generalized architectures of previous hierarchical feedforward models with more stages (*layers* in deep learning literature) than their bio-inspired counterpart models. In this new paradigm, rather than building the model by implementing neural computations in a bottom-up manner as had been in the traditional approach, researchers rather put constraints on the goal and the architecture of the model and use huge sets of training data to get models develop their own solutions to the computations of the network (Yamins & DiCarlo, 2016). At first, it may seem that such an approach would yield divergent representations than those of biological systems. However, studies showed that they reach similar representations (Yamins et al., 2013; Cadieu et al., 2014), and that they can make successful predictions of neural responses across ventral stream (Guclu & van Gerven, 2015).

If there is a capacity-limited part of scene recognition as suggested by our first experiment, we thought there might also be a systematic shortcoming of feedforward models to explain human behavior. Attention demanding processes, by definition, requires more attentional capacity in challenging situations. In the MOT task, for example, it has been shown that increasing the number of targets or the speed of objects makes the task more attention demanding (Alvarez & Franconeri, 2007). Similarly, one might expect images that are hard to recognize to require more attention.

A machine learning algorithm has been shown that it may be used to assess discriminability or difficulty to categorize a stimulus (Sofer, Crouzet, & Serre, 2015). Sofer et al. has demonstrated that such an approach could also be used to analyze human performance pattern. In convolutional neural networks (CNN), the last layer

provides a posterior probability distribution of an image for a set of categories. CNN performances, however, are not measured using this number directly but rather by analyzing whether the posterior probability of the target category is among the highest of all categories. The end result provides a measure of how close an image to the learned representation of the category. In Experiment 2, we tested whether a set of challenging images for a feedforward model are also challenging to recognize for human participants. We hypothesized that as feedforward models can account for human scene recognition, participants would have a lower performance when the targets are among the set of hard images (low posterior probability) not only in dual-task but also in single-task conditions. However, with the role of attention in scene recognition, we would see a larger decrease in performance from easier to harder target trials in dual-task conditions than in baseline conditions, because participants would be left with less attentional capacity to overcome the challenge introduced by hard images. In other words, their performance in the reduced-attention condition would have a similar pattern to a feedforward model, which lacks an attention component.

3.1. Methods

The experimental setup was identical to the one described above except of the specified differences.

3.1.1. Participants

Five psychophysically trained Boğaziçi University students and the author participated to experiment voluntarily. All participants had normal or corrected-to-normal vision.

3.1.2. Stimuli

Posterior probability of Places-CNN model for a set of beach images were taken from SUN Database as our difficulty benchmark. Places-CNN is a scene recognition convolutional neural network (CNN) that uses ImageNet-CNN architecture (Krizhevsky, Sutskever, & Hinton, 2012) and is trained on Places database with a state-of-the-art performance on scene recognition benchmarks (Zhou et al., 2014). Since there is no "beach" label in Places-CNN, we rather used "coast" and "ocean" as target categories for the model and chose the higher probability among them as the posterior probability for our "beach" label. Resulting distribution of posterior probabilities of beach images had a normal-like shape (M = .28, SD = .14) (Fig. 6).

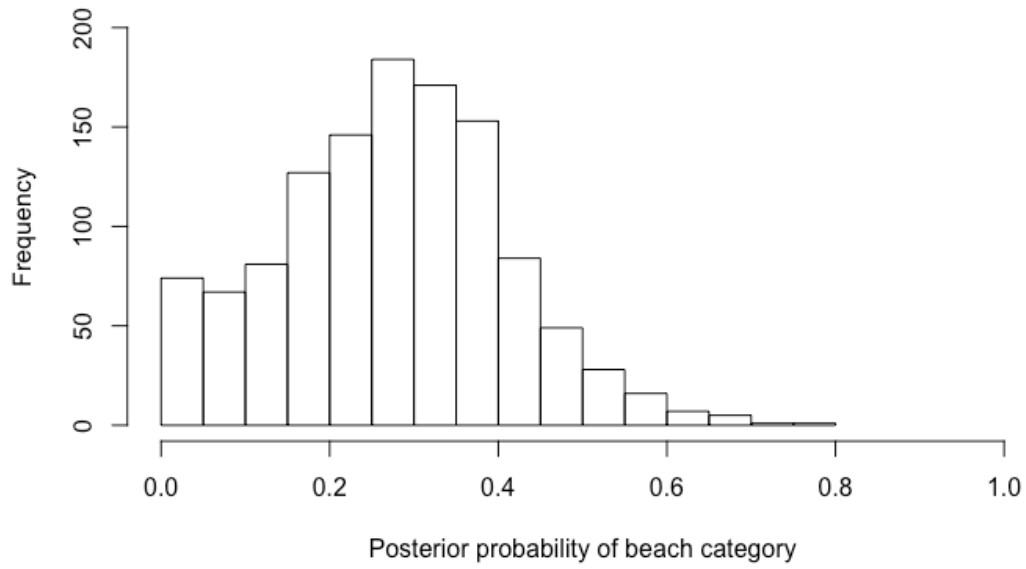


Fig. 6 Distribution of posterior probabilities of beach images

To have enough images for hard and easy conditions, whereas the images with less than .15 posterior probability were categorized as hard images, those with more than

.4 posterior probability were categorized as easy ones. 98 images from those categories were then selected as target images for each condition. Masking images were synthesized using the same method in Experiment 1.

3.1.3. Procedure

Participants completed basic-level categorization (beach) tasks. Easy- and hard-category images were used in blocked trials in both single-scene recognition and dual-task conditions, making a total of 4 experimental runs. A control single-MOT block was also run for each observer. Order of the blocks were counterbalanced across different participants.

Presentation durations of the scene images were as same as those used in Experiment 1. Each data point was composed of 28 trials, which made a total of 7 × 28 = 196 trials in each run. Experimental blocks started with 28 training trials to prepare participants to the procedure.

3.2. Results

Experiment 1 showed that performing scene recognition concurrently with a MOT task makes scene recognition tasks more difficult, indicated by longer presentation durations. If the observed effect is the result of a higher-level attentional resource sharing, then, one might expect performance of a feedforward computational model, which lacks such an attentional component, to be similar to the performance of the participants in the dual-task condition, where their attentional capacity is also reduced by the concurrent MOT task.

A 2 × 2 repeated-measures ANOVA on minimum presentation duration threshold values showed a main effect of attention that supported the result of

Experiment 1, such that to complete scene recognition tasks in dual-task condition (M = 85.6 ms, SD = 12.2), a longer presentation duration was required than that in single-task condition (M = 69.2 ms, SD = 12.4), $F(1, 5) = 11.75$, $MSe = 138.71$, p = .019, η_p^2 = .70, η_G^2 = .29 (Fig. 7). Unlike the computational model, though, there was no significant difference between the minimum presentation durations required for

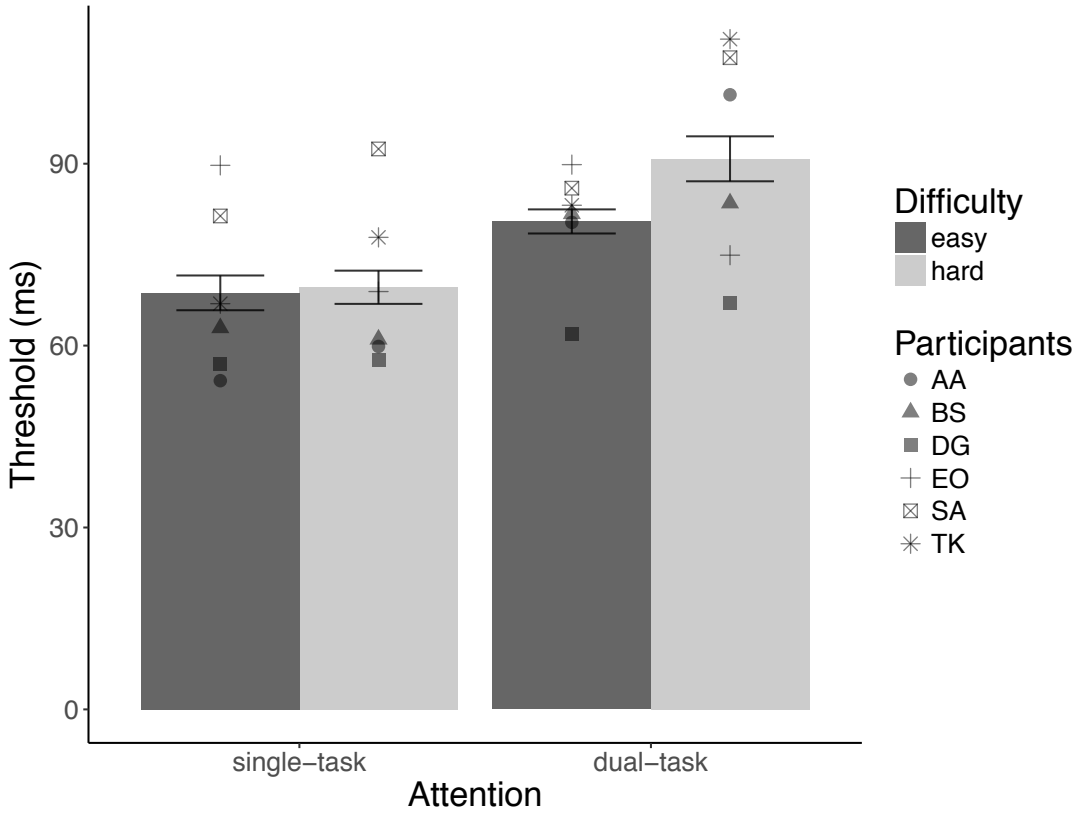


Fig. 7 Minimum presentation thresholds of Experiment 2

Note. Bar graph shows means of thresholds and symbols show individual data points of participants. Error bars indicate the standard error of the mean. There was a main effect of attention, p = .019. The interaction between attention and difficulty was also significant, such that hard images required significantly more presentation duration for participants only in dual-task condition, p = .009.

participants to categorize hard and easy images, $F(1, 5) = 1.00$, $MSe = 189.77$, $p > .05$, $\eta_p^2 = .16$, $\eta_G^2 = .05$. There was, however, an effect of interaction showing that for participants, the recognition duration thresholds of hard images ($M = 90.1$ ms, $SD = 18.2$) were significantly higher than those of easy images ($M = 80.5$ ms, $SD = 9.7$) only in dual-task but not in single-task conditions, suggesting that the locus of the effect observed in Experiment 1 might be higher in the visual hierarchy than the layers modeled by feedforward models, $F(1, 5) = 17.05$, $MSe = 7.79$, $p = .009$, $\eta_p^2 = .77$, $\eta_G^2 = .03$ (see Fig. B1 for individual data).

We also analyzed d′ values by pooling all data points of presentation durations to measure the perceived task difficulty in different conditions. A 2 × 2 repeated-measures ANOVA on d′ values showed that the difficulty of dual-task conditions ($M = .9$, $SD = .2$) was significantly higher than single-task conditions ($M = 1.2$, $SD = .3$), $F(1, 5) = 9.67$, $MSe = .08$, $p = .027$, $\eta_p^2 = .66$, $\eta_G^2 = .38$ (Fig. 8). Participants, however, did not have significantly different difficulty in completing blocks of trials with either hard or easy images, $F(1, 5) = .03$, $MSe = .05$, $p > .05$, $\eta_p^2 = .05$, $\eta_G^2 = .01$. The interaction effect was not significant, either, for d′ values, $F(1, 5) = 1.73$, $MSe = .01$, $p > .05$, $\eta_p^2 = .26$, $\eta_G^2 < .01$.

While the percentage of correct answers in a single-task MOT condition ($M = .87$, $SD = .05$) was higher than that in a dual-task condition ($M = .83$, $SD = .07$), the difference was not significant, $t(5) = 1.53$, $p > .05$ (Fig. 9).

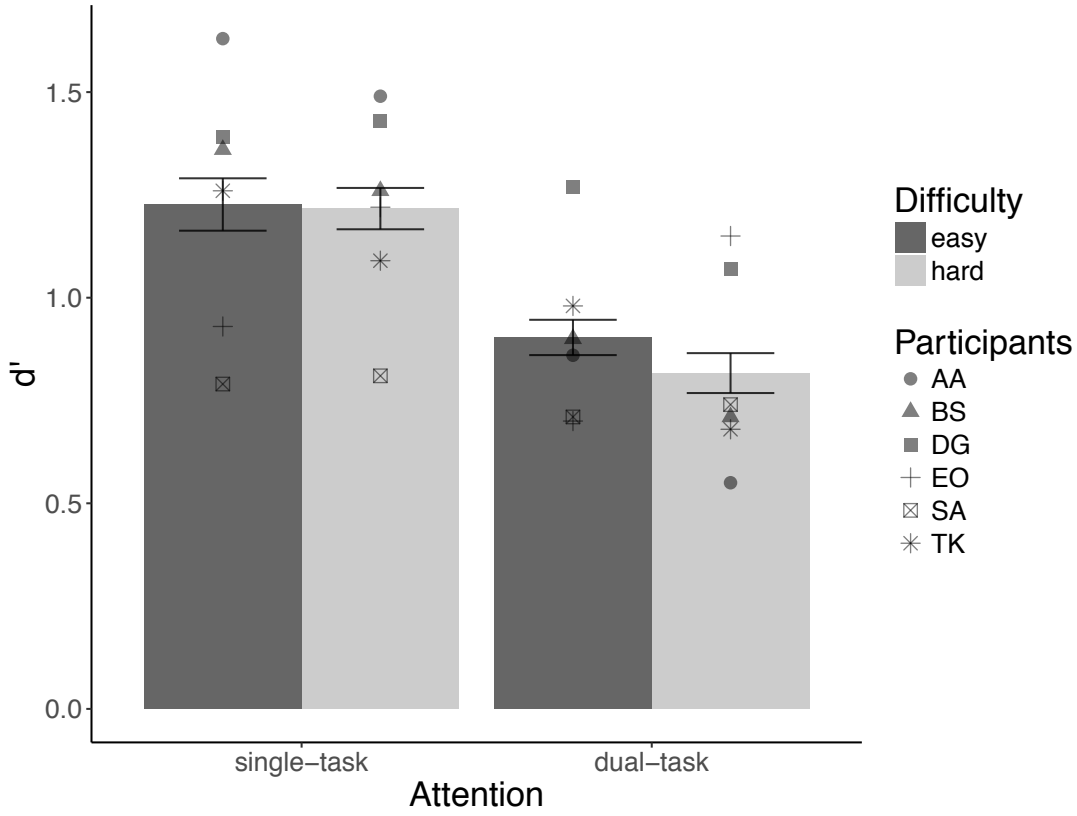


Fig. 8 d' values of Experiment 2

Note. Bar graph shows means of d' values and symbols show individual data points of participants. Error bars indicate the standard error of the mean. There was a main effect of attention, $p = .027$.

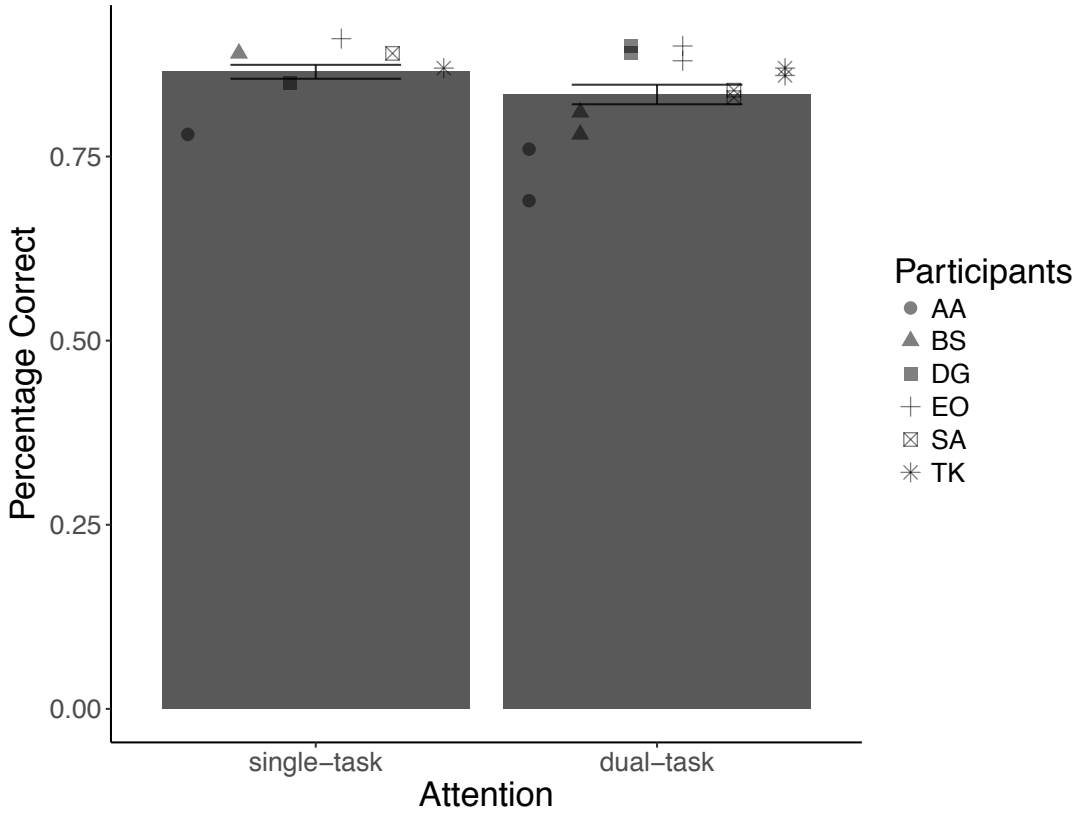


Fig. 9 MOT results of Experiment 2

Note. Bar graph shows means of the percentage of correct answers for MOT task and symbols show individual data points of participants. Error bars indicate the standard error of the mean.

.

CHAPTER 4

DISCUSSION

In Experiment 1, we tested the effect of attention on a broad range of scene recognition tasks using a dual-task paradigm and a reliable psychophysical methodology. Our results showed a performance decrease in both the scene recognition tasks and the MOT task from single- to dual-task conditions. Despite the evidence in literature that support a parallel, pre-attentive, feedforward account of scene recognition (Rousselet et al., 2002; Li et al., 2002; Oliva & Torralba, 2001; Torralba & Oliva, 2003; Fabre-Thorpe, 2011), our results supported the findings of Cohen et al. (2011) and suggested that there might be an attentional capacity sharing between the scene recognition and the MOT tasks.

The computational analysis of global visual features that can discriminate scene categories (Oliva & Torralba, 2001; Torralba & Oliva, 2003) and the hierarchical models of visual recognition (Serre, Oliva, & Poggio, 2007; Riesenhuber & Poggio, 1999) have shown the feasibility of feedforward accounts in scene recognition (Fabre-Thorpe, 2011). One of the main problems of visual recognition is to reach invariant representations of target categories such that objects and scenes could be recognized under various angles of views and lighting conditions (Riesenhuber & Poggio, 2000). Computational studies showed this problem can be handled by hierarchical feedforward models, in which neural receptive fields get increased up through the hierarchical system, where neurons start to be selective to more complex features (Riesenhuber & Poggio, 2000; Serre, Kreiman, et al., 2007). If global spatial features proposed by Oliva and Torralba (2001) such as naturalness, openness and depth were sufficient for scene recognition, then a group of neurons

that are specialized on those features could facilitate scene recognition in a purely feedforward manner. Such model might in fact explain the results of behavioral studies that failed to show attentional effects on scene recognition (Rousselet et al., 2002; Li et al., 2002; Greene & Fei-Fei, 2014) and the brief processing times reported in the ERP studies (Thorpe et al., 1996; Rousselet, Joubert, & Fabre-Thorpe, 2005). Another implication of a global feature-based account of scene recognition would be a system with a fixed performance characteristic that could not be improved by familiarity or experience, a prediction also supported by behavioral human data (Fabre-Thorpe et al., 2001). When the information in relation to a feature is coded in the pattern of firing across cells, one would observe aftereffects as a result of diminished activity of neurons following adaptation. In this context, Greene and Oliva (2010) used an RSVP paradigm to test whether global scene properties are susceptible adaptation effects and found that performances of observers' basic-level scene categorization are modulated after adapting to a global property, indicating the role of global properties in rapid scene categorization. It is important to note, however, that although these models seem to provide a relatively low-level account of scene recognition, the role of attention is still not completely eliminated. Some researchers have argued that attention is involved in a rather later stage, when recognition process is already completed and the system is occupied making target selection from competing stimuli, or while in the process of response preparation (Rousselet et al., 2004b; Reynolds, Chelazzi, & Desimone, 1999).

It is claimed that top-down modulation might facilitate object recognition via feedback connections by providing a context, such as a scene category, that may limit the number of possible object categories (Bar & Aminoff, 2003; Torralba, Oliva, Castelhano, & Henderson, 2006; Bar & Ullman, 1996). Such models depend

on a concept of a fast-track stream, processing scene contexts pre-attentively (Kveraga, Ghuman, & Bar, 2007). Our results demonstrating a clear role of attention on scene recognition, however, is rather compatible with an alternative account, where scene recognition is supposed to be top-down modulated in a manner similar to object recognition, where outputs of the initial fast-processing stream is later compared to the learnt representations held in the memory to minimize mismatches between high-level predictions and low-level activations (Ullman, 1995; Friston, 2005; Hinton, Dayan, Frey, & Neal, 1995).

The onset of the feedback activity from frontal and parietal areas which are associated with attentional modulation (Bressler, Tang, Sylvester, Shulman, & Corbetta, 2008) was reported to be within the 100-300 ms period following the stimulus onset (Mehta, Ulbert, & Schroeder, 2000). This range of interval is larger than most of the presentation durations we used in this study; thus, it is unlikely that it is such a top-down modulation which underlines our observed attentional effects. Because of our blocked design in which participants had already known what the target category was before they started doing the experiments, however, one might argue that in our paradigm top-down modulation might have occurred even before the stimulus presentation, making recognition faster during the actual trial. Using a pre-cue/post-cue paradigm, Evans, Horowitz, and Wolfe (2011) have shown that the number of to-be-recognized target categories and the level of the similarity across those categories have an effect on recognition performance, indicating that the visual system might in fact be prepared to recognize target categories before the actual stimulus presentation. In a pilot experiment, using a similar pre-cue/post-cue design, we tested whether the existence of attentional effect in our data might also be explained by such a predictory top-down modulation. For this experiment, we chose

three target categories (playground, highway, and castle) and asked participants to perform a basic-level scene categorization task. The design of the pre-cue condition was similar to that of Experiment 1, except that target categories were not presented in a blocked order but rather at the beginning of each trial (Fig. C1). The critical condition was the post-cue condition, though, where we did not pre-inform our participants on the target category and they made judgements as to whether the target was a member of a particular category with a possibility that it might have come from either of the three categories: playground, highway, or castle. For both pre-cue and post-cue conditions, we also separately ran single- and dual-task blocks using the same MOT task as was used in Experiment 1. Our hypothesis was that if there were a preparatory top-down modulation in the visual system before the actual stimulus onset and the additional MOT task in Experiment 1 introduced noise in the preparatory signal, then, we would replicate the results of Experiment 1 only in the pre-cue condition and would not see a degradation in performance form single- to dual-task conditions in the post-cue condition. We observed no significant performance difference, though, between pre- and post-cue conditions contrary to the results reported by Evans et al. (2011), suggesting that our manipulation had either no effect as an independent variable, or more possibly, using a set of three categories was simply not challenging enough to reduce preparatory signal in relation to each category (Fig. C2). We were not able to increase the number of categories, however, as it would get the duration of a block longer than what participants could reliably perform. Such a possible source of attention, therefore, remains as an open question for future studies.

It has been shown that attending a temporally preceding stimulus might decrease recognition performance for succeeding stimuli (Evans & Treisman, 2005).

In a second pilot experiment, we tested whether it was the sustained-attention allocated to the MOT task, disrupting the temporal attention given to the target scene was the cause of the performance drop observed in the dual-task condition of Experiment 1. While there was only a single scene image in our procedure, the MOT task might have interfered with a suppression mechanism which reduced the effect of masking images or might have alternatively caused participants to miss the short temporal window when the target was presented, a particularly possible account as the number of masking images and thus the trial length was randomized across trials; thus, participants could not reliably predict when exactly the scene image was going to appear. In this experiment, we modified our procedure in Experiment 1 by adding a fixed trial duration condition and asked participants to perform a basic-level categorization using playground images as targets (Fig. D1). We hypothesized that if the performance decrease in dual-task condition was due to the MOT task allocating from the common temporal attention resources, then, we would see an enhancement in the performance degradation from single- to dual-task conditions when the trial duration was kept at a fixed rather than a randomized value. The results, however, have controversially shown a trend of higher thresholds in the fixed trial duration condition (Fig. D2). Predicting that it would be fruitless, we left this data as it is and did not pursue this issue in further experiments.

Following the inconclusive aforementioned experiments, while the type of attentional effect on scene recognition observed in Experiment 1 had remained at a speculative level, the source was most likely to be in the feedback connections in the visual system (Kveraga et al., 2007; Kastner & Ungerleider, 2000). Thus, we thought feedforward models might have systematic shortcomings while trying to explain human performance. Serre, Oliva, and Poggio (2007) have proposed a feedforward

model for visual recognition and shown that the types of images the model had a difficulty to recognize were also challenging for their human participants. In Experiment 2, by choosing a set of images that was challenging enough to a state-of-the-art feedforward model (Zhou et al., 2014), we tested to what extent the model would be able to account for human performance. Our results showed that the hard images classified by the model were significantly harder to recognize for participants only in the dual-task, but not in the single-task condition, suggesting that participants could fail to compensate for the challenging nature of the images only when some of their attention were already allocated to the MOT task.

One of the recent developments in the computational object recognition models is the inclusion of attentional mechanisms that serially select some portions of the images to allocate more processing resources with an increased performance (Cao, Liu, Yang, Yu, & Wang, 2015; Eslami, Heess, Weber, & Tassa, 2016; Mnih, Heess, Graves, & Kavukcuoglu, 2014). While these models show how attentional mechanisms could improve feedforward models, such a serial attention modulation akin to the spatial attention in humans seems unlikely to explain the observed effect in Experiment 1, as we only used brief presentation durations (around 120 ms at maximum) that would not allow more than a glance of the stimuli. The recurrent neural network (RNN) model proposed by Liao and Poggio (2016) has shown another approach to include feedback connections to the computational models. In their model, a recurrent network builds more complex representations by simulating deeper layers of a feedforward model with a shallower network after an initial feedforward feature extraction. While such an approach could explain the performance increase in human participants with longer presentation durations, it is unclear how a modulatory processing could be integrated into this model to

implement capacity sharing between the scene recognition and the MOT tasks. Such connections might have similar functions to the local feedback connections in the visual system that improve object representations and could be dissociated from attentional top-down modulations on the basis of their early onset characteristic (Wyatte, Jilk, & O'Reilly, 2014; Bachmann, 2014).

Next step to understand the role of attention on scene recognition might be running computational analyses on the features of the images that are challenging for the current feedforward models as done by Oliva and Torralba (2001) for the feedforward part. Determination of computations that can facilitate the discrimination of such stimuli may allow us to model the attentional resources that can be shared across different tasks and modulated in a top-down manner.

CHAPTER 5

CONCLUSION

Despite the accumulated evidence in the literature for a feedforward, low attention demanding account of scene recognition, here, using a dual-task paradigm and a reliable psychophysical method, we showed that there might be a role of attention in scene recognition. While the possible source and the function of the attentional modulation, such as top-down preparation and temporal attention, is still open to debate, we compared the performance pattern of a feedforward model to human participants, and showed that current feedforward models cannot account for human performance in scene recognition. These results suggest that the source of the observed effect might be a high-level attentional process that can modulate and be shared across different tasks.

APPENDIX A

INDIVIDUAL DATA OF EXPERIMENT 1

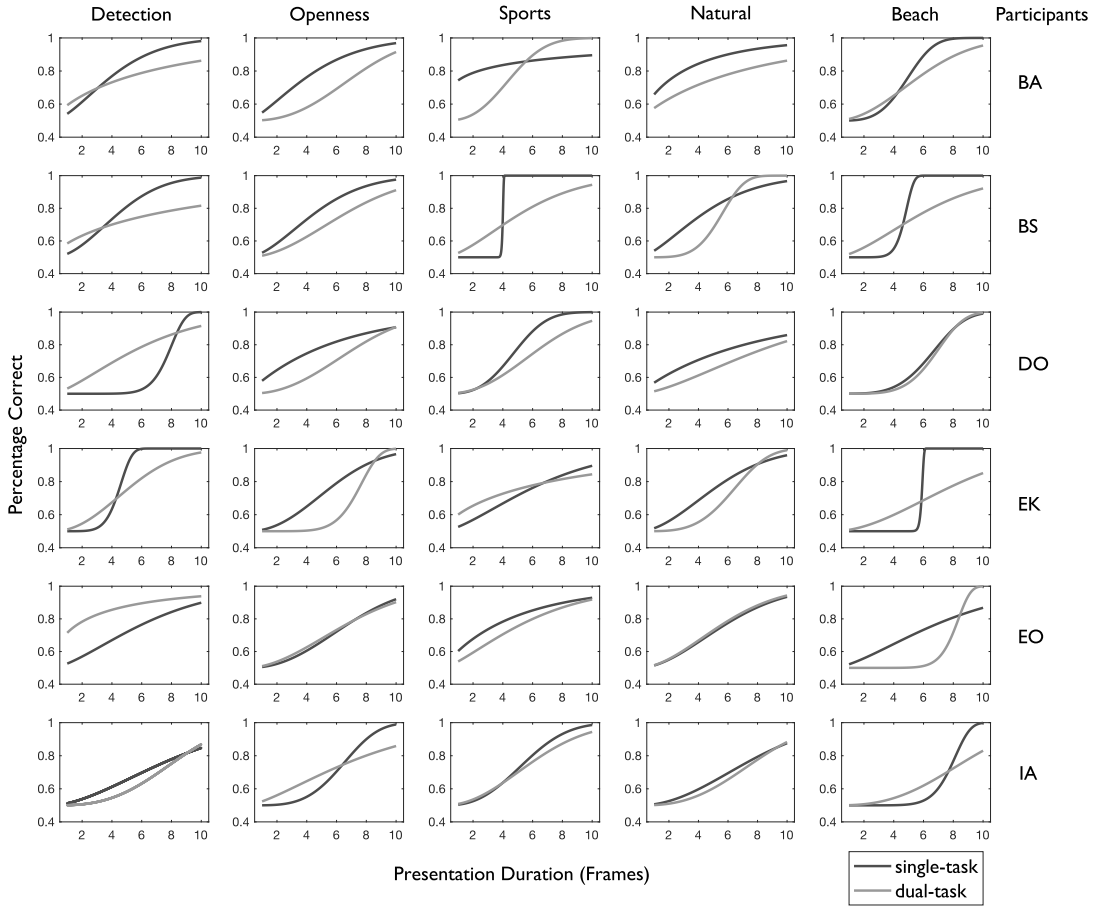


Fig. A1 Individual psychometric functions of Experiment 1

Notes. Individual data points were fit into Weibull function. Darker lines indicate single-task, whereas lighter lines indicate dual-task conditions.

APPENDIX B

INDIVIDUAL DATA OF EXPERIMENT 2

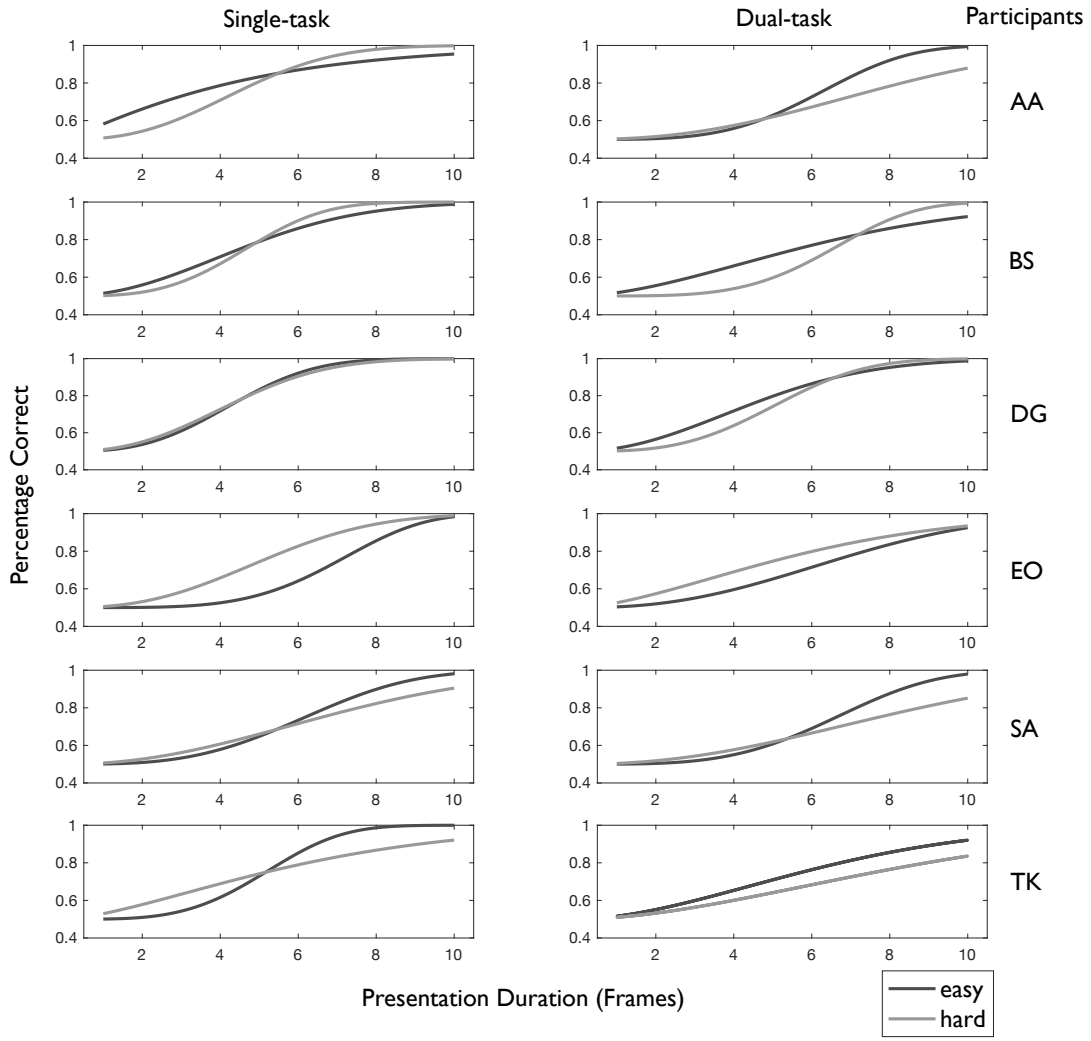


Fig. B1 Individual psychometric functions of Experiment 2

Note. Individual data points were fit into Weibull function. Darker lines indicate easy images, whereas lighter lines indicate hard images conditions.

APPENDIX C

TOP-DOWN EXPECTATIONS EXPERIMENT

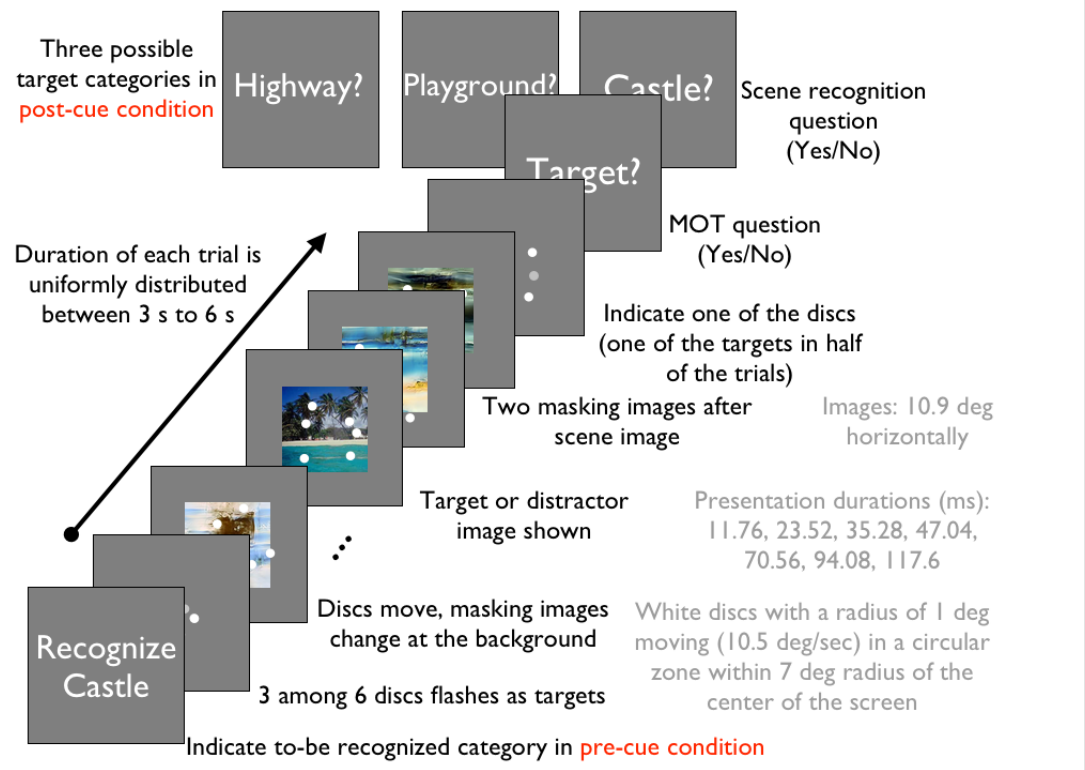


Fig. C1 The procedure of the top-down expectations experiment

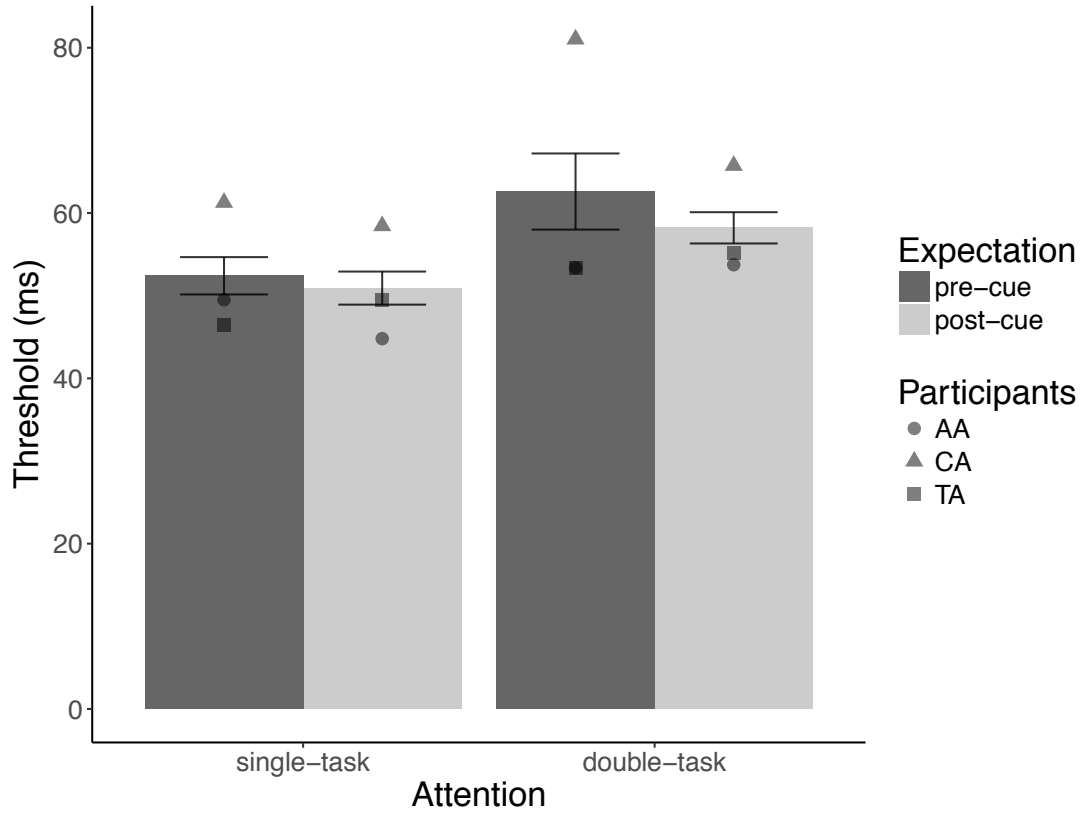


Fig. C2 Minimum presentation durations of the top-down expectations experiment

Note. Bar graph shows means of thresholds and symbols show individual data points of participants. Error bars indicate the standard error of the mean. There was a trend of higher presentation durations from single- to dual-task conditions $F(1, 2) = 13.36$, $MSe = 17.16$, $p = .067$, $\eta_p^2 = .87$, $\eta_G^2 = .22$. We did not observe, however, a significant effect of pre-/post-cue manipulation, which was critical to test our hypothesis, $F(1, 2) = .78$, $MSe = 33.31$, $p > .05$, $\eta_p^2 = .28$, $\eta_G^2 = .03$. There was also no significant interaction effect, $F(1, 2) = .32$, $MSe = 19.67$, $p > .05$, $\eta_p^2 = .14$, $\eta_G^2 = .01$.

APPENDIX D

TEMPORAL ATTENTION EXPERIMENT

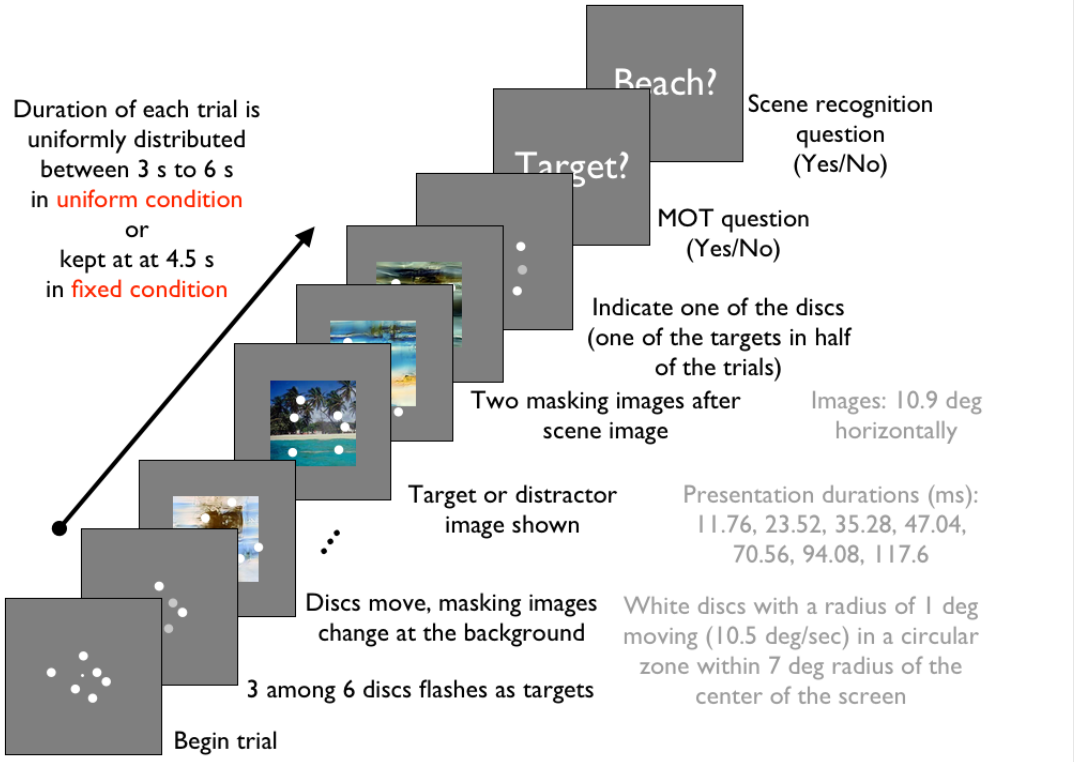


Fig. D1 The procedure of the temporal attention experiment

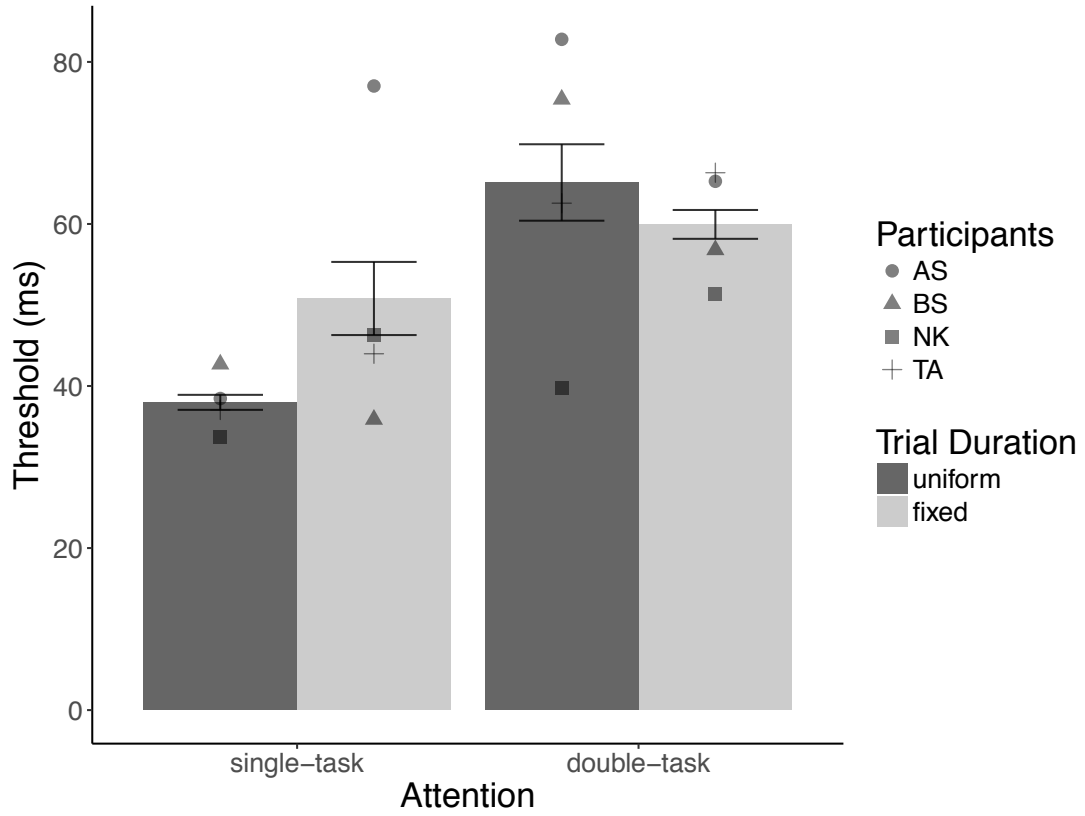


Fig. D2 Minimum presentation durations of the temporal attention experiment

Note. Bar graph shows means of thresholds and symbols show individual data points of participants. Error bars indicate the standard error of the mean. There was a significant main effect of attention, $F(1, 3) = 14.56$, $MSe = 90.42$, $p = .032$, $\eta_p^2 = .83$, $\eta_G^2 = .37$. We did not observe, however, a significant effect of uniform/fixed trial duration manipulation, which was critical to test our hypothesis, $F(1, 3) = .55$, $MSe = 129.67$, $p > .05$, $\eta_p^2 = .13$, $\eta_G^2 = .03$. There was also no significant interaction effect, $F(1, 3) = 1.94$, $MSe = 166.75$, $p > .05$, $\eta_p^2 = .39$, $\eta_G^2 = .13$.

REFERENCES

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, *7*(13), 14–14.

Bachmann, T. (2014). A hidden ambiguity of the term "feedback" in its use as an explanatory mechanism for psychophysical visual phenomena. *Frontiers in Psychology*, *5*, 780.

Bacon-Macé, N., Macé, M. J. M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, *45*(11), 1459–1469.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, *38*, 347–358. Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception*, *25*(3), 343–352.

Biederman, I. (1972). Perceiving Real-World Scenes. *Science*, *177*(4043), 77–80.

Biederman, I. (1976). On processing information from a glance at a scene: Some implications for a syntax and semantics of visual processing. *Proceedings of the ACM/SIGGRAPH Workshop on User-Oriented Design of Interactive Graphics Systems*, 75–88.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Brainard, D. H., & others. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Braun, J., & Julesz, B. (1998). Withdrawing attention at little or no cost: detection and discrimination tasks. *Perception & Psychophysics*, *60*(1), 1–23.

Breitmeyer, B. G., & Ögmen, H. (2000, December). Recent models and findings in visual backward masking: A comparison, review, and update. *Perception & Psychophysics*, *62*(8), 1572–1595.

Bressler, S. L., Tang, W., Sylvester, C. M., Shulman, G. L., & Corbetta, M. (2008, October). Top-Down Control of Human Visual Cortex by Frontal and Parietal Cortex in Anticipatory Visual Spatial Attention. *The Journal of Neuroscience*, *28*(40), 10056–10061.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., . . . DiCarlo, J. J. (2014, December). Deep Neural Networks Rival the

Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Cao, C., Liu, X., Yang, Y., Yu, Y., & Wang, J. (2015). 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).

Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011, August). Natural-Scene Perception Requires Attention. *Psychological Science*, *22*(9), 1165–1172.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013, January). The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. *The Journal of Neuroscience*, *33*(4), 1331–1336.

Epstein, R. A. (2005, August). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954–978.

Epstein, R. A., Higgins, J. S., Jablonski, K., & Feiler, A. M. (2007, March). Visual Scene Processing in Familiar and Unfamiliar Environments. *Journal of Neurophysiology*, *97*(5), 3670–3683.

Epstein, R. A., & Kanwisher, N. (1998, April). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Eslami, S., Heess, N., Weber, T., & Tassa, Y. (2016). Attend, infer, repeat: Fast scene understanding with generative models. *Advances in Neural Information Processing Systems*, 3225–3233.

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011, June). When Categories Collide Accumulation of Information About Multiple Categories in Rapid Scene Perception. *Psychological Science*, *22*(6), 739–746.

Evans, K. K., & Treisman, A. (2005, December). Perception of Objects in Natural Scenes: Is It Really Attention Free? *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1476–1492.

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology*, *2*, 243.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. J. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience*, *13*(2), 1–10.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007, January). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 10–10.

Friston, K. (2005, April). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815-836.

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82–94.

Greene, M. R., & Fei-Fei, L. (2014, January). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 14–14.

Greene, M. R., & Oliva, A. (2009a, April). The Briefest of Glances The Time Course of Natural Scene Understanding. *Psychological Science*, *20*(4), 464–472.

Greene, M. R., & Oliva, A. (2009b, March). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176.

Greene, M. R., & Oliva, A. (2010, December). High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(6), 1430–1442.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152–160.

Guclu, U., & van Gerven, M. A. J. (2015, July). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of Neuroscience*, *35*(27), 10005–10014.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2015, November). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*(1), 243–271.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *195*, 215–243.

Kadar, I., & Ben-Shahar, O. (2012, December). A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *Journal of Vision*, *12*(13), 16–16.

Kastner, S., & Ungerleider, L. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*(1), 315–341.

Keysers, C., Xiao, D. K., Földiák, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience*, *13*(1), 90–101.

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, *63*(8), 1421–1455.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., & others. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007, November). Top-down predictions in the cognitive brain. *Brain and Cognition*, *65*(2), 145–168.

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10).

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444.

Li, F. F., VanRullen, R., & Koch, C. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596–9601.

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv.org*.

Loschky, L. C., & Larson, A. M. (2010, April). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*(4), 513–536.

Macé, M. J. M., Joubert, O. R., Nespoulous, J.-L., & Fabre-Thorpe, M. (2009, June). The Time-Course of Visual Categorizations: You Spot the Animal Faster than the Bird. *PLoS ONE*, *4*(6), e5927–12.

Mack, M. L., & Palmeri, T. J. (2010, October). Decoupling object detection and categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1067–1079.

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016, November). Making Sense of Real-World Scenes. *Trends in Cognitive Sciences*, *20*(11), 843–856.

McElree, B., & Carrasco, M. (1999). The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. *Journal of Experimental Psychology*, *25*(6), 1517–1539.

Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2000). Intermodal selective attention in monkeys. I: distribution and timing of effects across visual areas. *Cerebral Cortex*, *10*, 343–358.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems*, 2204–2212.

Olejnik, S., & Algina, J. (2003). Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, *8*(4), 434–447.

Oliva, A. (2013). Scene perception. In J. S. Werner & L. M. Chalupa (Eds.), *The new visual neurosciences.*

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*, 72–107.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2002). Scene-centered description from spatial envelope properties. *Biologically motivated computer vision*.

Pashler, H. (1994, September). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244.

Patterson, G., & Hays, J. (2017, March). The SUN Attribute Database: Organizing Scenes by Affordances, Materials, and Layout. In *Visual attributes* (pp. 269–297). Cham: Springer International Publishing.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Poncet, M., & Fabre-Thorpe, M. (2014, May). Stimulus duration and diversity do not reverse the advantage for superordinate-level representations: the animal is seen before the bird. *European Journal of Neuroscience*, *39*(9), 1508–1516.

Portilla, J., & Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15.

Prins, N., & Kingdon, F. (2009). Palamedes: Matlab routines for analyzing psychophysical data.

Quattoni, A., & Torralba, A. (2009). Recognizing indoor scenes. *Computer Vision and Pattern Recognition*. R Core Team. (2016). R: A Language and Environment for Statistical Computing.

Renninger, L. W., & Malik, J. (2004, September). When is scene identification just texture recognition? *Vision Research*, *44*(19), 2301–2311.

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, *19*(5), 1736–1753.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*, 1199–1204.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., & Johnson, D. M. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rousselet, G., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, *5*(7), 629–630.

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005, August). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852–877.

Rousselet, G., Thorpe, S. J., & Fabre-Thorpe, M. (2004a, August). How parallel is visual processing in the ventral pathway? *Trends in Cognitive Sciences*, *8*(8), 363–370.

Rousselet, G., Thorpe, S. J., & Fabre-Thorpe, M. (2004b, April). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision research*, *44*(9), 877–894.

Serre, T., Kreiman, G., Kouh, M., & Cadieu, C. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, *165*, 33–56.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. In *Proceedings of the national academy of sciences* (pp. 6424–6429).

Sofer, I., Crouzet, S. M., & Serre, T. (2015, September). Explaining the Timing of Natural Scene Understanding with a Computational Model of Perceptual Categorization. *PLoS Computational Biology*, *11*(9), e1004456–20.

Thorpe, S. J., Fize, D., & Marlot, C. (1996, June). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Torralba, A., & Oliva, A. (2003, January). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *353*(1373), 1295–1306.

Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, *5*(1), 1–11.

VanRullen, R., & Thorpe, S. J. (2016, June). Is it a Bird? Is it a Plane? Ultra-Rapid Visual Categorisation of Natural and Artifactual Objects. *Perception*, *30*(6), 655–668.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009, August). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *The Journal of Neuroscience*, *29*(34), 10573–10581.

Webster, M. A., & De Valois, R. L. (1985). Relationship between spatial-frequency and orientation tuning of striate-cortex cells. *JOSA A*, *2* (7), 1124–1132.

Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014, July). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, *5*(4), 760.

Xiao, J., Ehinger, K. A., Hays, J., Torralba, A., & Oliva, A. (2014, August). SUN Database: Exploring a Large Collection of Scene Categories. *International Journal of Computer Vision*, 1–20.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3485–3492.

Xiao, J., Hays, J., Russell, B. C., Patterson, G., Ehinger, K. A., Torralba, A., & Oliva, A. (2013). Basic level scene understanding: categories, attributes and structures. *Frontiers in Psychology*, *4*.

Yamins, D. L. K., & DiCarlo, J. J. (2016, February). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. *Advances in Neural Information Processing Systems*, 3093–3101.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems*, 487–495.