

Categorization Goals Modulate the Use of Natural Scene Statistics

Andrea De Cesarei, Shari Cavicchi, Antonia Micucci,
and Maurizio Codispoti

Abstract

■ Understanding natural scenes involves the contribution of bottom-up analysis and top-down modulatory processes. However, the interaction of these processes during the categorization of natural scenes is not well understood. In the current study, we approached this issue using ERPs and behavioral and computational data. We presented pictures of natural scenes and asked participants to categorize them in response to different questions (Is it an animal/vehicle? Is it indoors/outdoors? Are there one/two foreground elements?). ERPs for target scenes requiring a “yes” response began to differ from those of nontarget

scenes, beginning at 250 msec from picture onset, and this ERP difference was unmodulated by the categorization questions. Earlier ERPs showed category-specific differences (e.g., between animals and vehicles), which were associated with the processing of scene statistics. From 180 msec after scene onset, these category-specific ERP differences were modulated by the categorization question that was asked. Categorization goals do not modulate only later stages associated with target/nontarget decision but also earlier perceptual stages, which are involved in the processing of scene statistics. ■

INTRODUCTION

Adaptive interaction with the environment requires organisms to perceive external reality through sensory systems and execute meaningful responses. However, the complexity of the environment, as well as the inherent uncertainty in the visual input, prevent the system from analyzing all possible information contained within a view with the same fine-grained level of detail. Critical analysis steps include filtering out irrelevant information and noise and filling in missing information, with the ultimate aim of making sense of a view. Moreover, visual analysis relies on processes that prioritize relevant or novel information (Lang, Bradley, & Cuthbert, 1997; Wolfe, Cave, & Franzel, 1989; Sokolov, 1963) and ultimately achieve a gist and a semantic interpretation of a view (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

The visual input involves the contribution of several structures and processes, which analyze increasingly more complex visual properties. Initial stages of processing analyze visual information, which comprises several spatial scales. It is well known that activity in early visual areas is sensitive to changes in brightness, contrast, orientation, and spatial frequency (Pratt, 2011). Moreover, several studies from the mid 1980s (Field, 1987) emphasized the role of scene statistics in the processing of natural scenes (De Cesarei, Loftus, Mastria, & Codispoti, 2017; Simoncelli & Olshausen, 2001). Scene statistics

can be defined as quantifiable properties of the perceptual appearance of a scene, such as visual clutter, spatial arrangement, naturalness, and others. In the context of scene understanding, scene statistics may aid categorization when they describe regularities, which are common to a class of stimuli. For instance, Torralba and colleagues demonstrated that artificial scenes, compared with naturalistic scenes, contain more edges and sharp contours (Torralba & Oliva, 2003). Scene statistics are particularly important when the visual system deals with natural world views such as natural scenes, rather than with well-controlled but simplified laboratory stimuli (Groen, Silson, & Baker, 2017; Hasson, Malach, & Heeger, 2010). It has recently been suggested that the visual system can make efficient use of scene statistics. For instance, scene statistics of energy and clutter have been shown to modulate the overall activity over occipital scalp electrodes approximately 100 msec after the onset of a visual stimulus (Groen, Ghebreab, Lamme, & Scholte, 2012).

Based on the available information, categorization of the visual input can take place. View-invariant visual representations represent a cortical correlate of abstraction and are first observed at the level of the extrastriate visual areas (Cauchoix & Crouzet, 2013; Miyashita, 1993; Nishijo, Ono, Tamura, & Nakamura, 1993). Focusing on the categorization of objects in scenes, a number of studies have used ERPs to investigate the electrocortical correlates of categorization and revealed differences between a target category (animals) and distractors beginning at 150 msec (Thorpe, Fize, & Marlot, 1996); it is

likely that these differences originated from neural generators in the inferior temporal cortex (Codispoti, Ferrari, Junghöfer, & Schupp, 2006; Fize et al., 2000). Moreover, several studies have further differentiated between an initial modulation, which is tied to sensory differences between stimuli, and a later modulation interval, associated with the target relevance of some stimuli, which is defined by the task that participants are carrying out (Johnson & Olshausen, 2003; VanRullen & Thorpe, 2001; see also Rossion & Caharel, 2011, for similar results in the processing of faces). Importantly, categorization in these studies is observed when participants are told beforehand what should be considered as a target, creating a top-down presetting, which may bias the rapid categorization of natural scenes (De Cesarei, Peverato, Mastria, & Codispoti, 2015; Schendan & Ganis, 2015; Fabre-Thorpe, 2011; Treisman, 2006; Enns, 2004).

Little is currently known about whether rapid categorization is restricted to some categorization tasks, such as naturalistic-artifactual distinction, or whether it is linked to more domain-general processes. A previous study addressed this issue by comparing the effects of different task instructions (“categorize animals” or “categorize vehicles”) on the processing of the same images (animals, vehicles, or various distractors; VanRullen & Thorpe, 2001). When ERPs in response to the same images were compared, depending on whether they were targets or nontargets, the ERP began to differ between 156 and 212 msec, with differences in topography and latency depending on the categorization goal. Furthermore, another study used the same approach to examine the categorization of animals and humans compared with distractors and observed later effects for task-driven effects compared with picture-related effects, for both target categories (Rousselet, Macé, Thorpe, & Fabre-Thorpe, 2007).

Top-down modulations combine with bottom-up processes in the analysis and interpretation of the visual input. In terms of top-down processes, it has often been noted that the brain analyzes incoming input actively and through inferential processes. This concept is not new and dates back to Von Helmholtz’s (1867) concept of unconscious inferences, which was later developed by Gregory (1970) and, more recently, by several models in cognitive neuroscience (Schendan & Ganis, 2015; Friston & Kiebel, 2009; Philiastides & Sajda, 2006; Bar, 2003). Within these views, the system is thought to analyze the incoming input not through a comprehensive and exhaustive analysis but using an inferential processing, which involves the generation and testing of perceptual hypotheses. For this reason, visual perception represents a form of perceptual decision-making (Seger & Peterson, 2013; Philiastides & Sajda, 2006). Several studies have investigated perceptual decision-making in conditions in which the visual input is perceptually degraded, such as in peripheral vision. In all these cases, the brain is called upon for an inferential choice to

compensate for the loss of information, which is due to perceptual degradation. Clear effects of degradation and object category are usually observed at the level of object-specific areas, such as the lateral occipital complex (De Cesarei, Mastria, & Codispoti, 2013; Grill-Spector, Kourtzi, & Kanwisher, 2001).

Recent studies have emphasized the interaction of bottom-up and top-down factors in the interpretation of natural scenes (Groen et al., 2017; Harel, Kravitz, & Baker, 2014; VanRullen & Thorpe, 2001). Scenes are rich in visual information, concerning, for instance, the spatial position of objects, the paths for navigation, the geographic context, the biological relevance, novelty, and so on (Malcolm, Groen, & Baker, 2016). Based on task demands, each of these sources of information may require focused attention, resulting in a variety of perceptual tasks within the same scene. How is the animal in a scene categorized when people are focusing on the environment in which the animal is located? A recent study presented pictures of objects and asked participants to perform in one of six different tasks. It was observed that object information could be decoded from fMRI activity over posterior fusiform areas, but that object information decoding varied between different tasks (Harel et al., 2014). However, a recent study comparing the effects of task goals on the processing of visual scenes failed to observe goal-related modulations in the processing of naturalness and spatial expanse up to the P2 ERP component (Hansen, Noesen, Nador, & Harel, 2018). Taken together, these results suggest that categorization goals may modulate the accumulation of information from the same visual input, at the level of extrastriate structures. However, the identification of the specific scene or object properties for which processing can be modulated remains a matter of debate, as does the timing at which this modulation occurs.

The Research Problem

This study focused on the neural processes that are implied in visual categorization and examined the extent to which they are general or to which they depend on the observer’s goal. To this end, pictures were presented, and participants categorized the pictures according to different goals (Figure 1). This study examined four research questions (summarized in Table 1):

Research Question 1 (RQ1). What is the role of top-down attention, independent of bottom-up features?

Expanding on previous studies (Rousselet et al., 2007; VanRullen & Thorpe, 2001), this study examined the role of task-related relevance on the processing and response to a natural scene. Here, we used a design in which several categorization questions could be asked (Figure 1), and the same picture could serve both as target and as nontarget for different questions. In this way, the overall difference between target and





TRIAL	1	2	3	4
PICTURE				
QUESTION	ANIMAL?	OUTDOOR?	VEHICLE?	ONE?
QUESTION DOMAIN	Content	Scenario	Content	Number
CATEGORY (* = probed by question)	Animal * Outdoor One	Vehicle Indoor * One	Vehicle * Outdoor Two	Animal Outdoor Two *
DECISION	Target	Nontarget	Target	Nontarget

Figure 1. Example of four trials, with coding of the experimental factors.

nontarget scenes is not confounded by any physical difference between scenes, as pictures contributing to the “target” and “nontarget” conditions are the same.

Research Question 2 (RQ2). *What is the role of bottom-up features, independent of attention?* Several previous studies indicated that bottom-up features, including sensory and compositional properties of a scene, are associated with modulation of ERPs (e.g., spectral power, De Cesarei et al., 2013; figure completion, Hazenberg & Van Lier, 2015; symmetry, Bertamini & Makin, 2014). In the current study, we examined ERP differences between physically different scenes independent of categorization goal and target status. The functional meaning of these ERP modulations and the extent to which they vary depending on one’s goals were examined in RQ2.1 and RQ3.

Research Question 2.1 (RQ2.1). *Are bottom-up driven ERP category differences related to a categorization based on scene statistics?* The observation of bottom-up related differences between the processing of two different scenes bears little information concerning which scene features are being analyzed, how the processing of these features contributes to the ERP modulation, and how ERP differences reflect processes that are related to the categorization task. Here, we approached this issue from the point of view of a system engaged in a binary classification task. Specifically, we compared bottom-up driven ERP modulations (RQ2) with the performance of an artificial system, which had to categorize scenes in the same categories as those given to human participants, based solely on scene statistics of energy and clutter (Groen et al., 2012; Scholte, Ghebreab, Waldorp, Smeulders, & Lamme, 2009). A

similarity in the performance of the two systems (ERP differences and artificial classification performance) would support the idea that scene statistics are involved in classification and that ERP modulations reflect categorization-oriented processing of scene statistics.

Research Question 3 (RQ3). *Do categorization goals modulate bottom-up analysis?* The moment in which the observer’s goals and attention begin to modulate visual processing is a matter of debate. Previous studies compared active to passive tasks, in which scenes were either relevant for the task that participants were carrying out or irrelevant and, in some cases, distractors (e.g., Groen, Ghebreab, Lamme, & Scholte, 2016). However, very few studies have, to date, compared scene processing during active categorization tasks while varying the categorization question (Hansen et al., 2018; Harel et al., 2014). In this study, we compared scene processing after three different categorization questions (Is it an animal/vehicle? Is it indoors/outdoors? Are there one/two foreground elements?) and examined the extent to which bottom-up differences (observed in RQ2) were similar across tasks or varied depending on the categorization question.

Research Question 4 (RQ4). *What is the role of memory systems in maintaining the categorization goal?* Finally, we investigated to what extent a sustained categorization goal (i.e., having the same categorization question for a series of trials) modulates the processing of natural scenes, compared with a situation in which the categorization goal varies on a trial by trial basis. If the goal of the observer remains the same throughout a long period, the system can be expected to form a better template of the categorization target,

which may result in earlier or stronger target-related effects, compared with a condition in which the categorization question is constantly changing.

METHODS

Participants

A total of 36 participants (24 women) took part in the study. Ages ranged from 19 to 37 years ($M = 22.19$ years, $SD = 3.21$ years). All participants had normal or corrected-to-normal vision, and none of them reported current or past neurological or psychopathological problems. Participants had no previous experience with the materials used in this experiment. The experimental protocol conforms to the declaration of Helsinki and was approved by the Bioethics Committee of the University of Bologna.

Stimuli

A total of 576 pictures were selected from the Internet for this study. The picture set was created along three orthogonal dimensions: content (animal or vehicle), number of foreground elements (one or two), and scenario (indoors or outdoors). Thus, each picture portrayed one or two foreground elements (animal or vehicle) in an indoor or outdoor scenario, for a total of eight possible combinations, examples of which are reported in Figure 2. For each combination of the three orthogonal factors,

72 scenes were collected. Pictures were in color and were balanced for brightness and contrast (0.5 and 0.8, respectively, on a 0–1 linear scale). Each picture was resized to 383×287 pixels and projected on a 21° monitor placed 60 cm from the participant, yielding a visual angle of 19.8° (horizontal) by 14.7° (vertical). We also used 128 additional pictures portraying pieces of furniture, houses, flowers, and musical instruments for practice phases.

Scene Statistics Calculation and Machine Learning Categorization

For each picture, spatial coherence and contrast energy were calculated using the algorithm developed by Scholte et al. (2009). This procedure estimates the beta (contrast energy) and gamma (spatial coherence) parameters of the Weibull fit to the distribution of contrast.

Scene statistics were used to investigate the extent to which an artificial classifier (support vector machine [SVM]) is able to categorize scenes using the same tasks that human participants were required to carry out: animals versus vehicles, indoors versus outdoors, and one versus two foreground elements. An SVM is a classification algorithm dedicated to finding the best separation (hyperplane) between two classes of data based on one or more continuous predictor variables and has been used in previous research on visual categorization (De Cesare & Codispoti, 2015; Crouzet & Serre, 2011; Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). The picture set was split

Table 1. Main Research Questions

	<i>Question</i>	<i>Section</i>	<i>Critical Comparison</i>	<i>Notes</i>
RQ1	What is the role of top–down attention, independent of bottom–up features?	Research Question 1; Figure 3; Figure 4	Factor decision: target vs. nontarget	
RQ2	What is the role of bottom–up features, independent of attention?	Research Question 2; Figure 5A	Factor category: animal vs. vehicle; indoors vs. outdoors; one vs. two	
RQ2.1	Are bottom–up driven ERP category differences related to a categorization based on scene statistics?	Research Question 2.1; Figure 5B; Figure 5C	Correlation ERP category differences (RQ2) vs. machine learning performance based on scene statistics	
RQ3	Do categorization goals modulate bottom–up analysis?	Research Question 3; Figure 6	Interaction involving category and question: (animal* vs. vehicle*) vs. (animal ^o vs. vehicle ^o)	* = probed by question (e.g., “Is it an animal/vehicle?”) o = not probed by question (e.g., “Is it indoors/outdoors/one/two?”)
RQ4	What is the role of memory systems in maintaining the categorization goal?	Research Questions 1, 3	Interaction with factor block (sustained; randomized; uncued)	

Each research question is listed, along with relevant sections in the paper and figures, and critical comparisons.

Figure 2. Picture examples. One representative example for each combination of the content, scenario, and number domain is displayed.



into a training subset and a test subset, keeping an equal number of pictures from each category in each subset. Contrast energy and spatial coherence for pictures in the training set, along with corresponding category labels, were given as input to an SVM. A total of 288 scenes were used for training, and an equal number of pictures were used for testing. The training and tests were repeated 100 times, and a different sample of pictures was selected each time for training and testing. The results of these runs were collapsed to obtain averages and standard errors of the mean for SVM categorization accuracy.

Procedure

The experiment required participants to match pictures and categorization questions. In each trial, a picture and a question were presented, and participants had to respond as to whether the picture and the question matched each other (Figure 1). For instance, when seeing a picture of two cats in the kitchen, participants could be asked whether the scene was an outdoor scene and, thus, were expected to respond “no.” Alternatively, observers might see the same picture and be asked whether it contained an animal, and therefore, the expected answer was “yes.” Each participant saw each picture once. Across participants, each picture was associated with all questions (animal, vehicle, indoors, outdoors, one, two). Throughout the experiment, each question was associated an equal number of times with each combination

of the three orthogonal dimensions. This design allowed each picture to be “target” and “nontarget” an equal number of times, therefore ruling out any physical difference between targets and nontargets. Moreover, each picture could be a target or nontarget for three different reasons—because of its content, the number of elements, or the scenario—therefore allowing us to investigate the effects of the question domain (scenario, content, or number) on the categorization of natural scenes. The whole experiment lasted about 1 hr, including breaks between blocks.

Block Procedure

The experiment was divided into three blocks (randomized, sustained, and uncued), and the order of blocks was counterbalanced across participants (Table 2). Each block consisted of 192 trials and was preceded by a practice phase. In the randomized and sustained blocks, one of six different questions (animal, vehicle, indoors, outdoors, one, two) was presented before each picture, whereas in the uncued block, the question was presented only after the picture. In all blocks, a yes/no response was required.

In the randomized block, the order of questions was pseudorandomized. In neighboring trials of the randomized block, a different question was asked. More specifically, as questions could pertain to the three orthogonal domains of content, scenario, or number (Figure 1), it was

Table 2. Design of the Experimental Blocks

<i>Block</i>	<i>Description</i>	<i>Categorization Question</i>	<i>Picture</i>	<i>Decision</i>
Sustained	Questions follow each other predictably (same question 32 times in a row)	Presented before the picture	One from the orthogonal combination of content (animal/vehicle), scenario (indoors/outdoors), and number (one/two)	Defined based on the match between question and picture; 50% target, 50% nontarget
Randomized	Questions follow each other unpredictably	Presented before the picture	Same as above	Same as above
Uncued	Questions follow each other unpredictably	Presented after the picture	Same as above	Same as above

decided that questions from the same domain would never be asked in two successive trials. Thus, for instance, if in one trial the question “Is it an animal?” was asked, the next trial could ask whether the picture was indoors, outdoors, with one or with two foreground elements, but not whether the picture portrayed an animal or a vehicle. This was done to avoid repetition and response switch effects between two successive trials. The randomized block contained 192 trials and was preceded by 32 practice trials.

In the sustained block, the same question was asked 32 times in a row. Thus, an “animal” question was always followed by an “animal” question. After all 32 trials, a short break was given to provide instructions concerning the following question, and after this break, new questions that were identical to each other were presented. Similar to the randomized block, it was decided that questions from the same domain would never be asked in adjacent series of questions. The sustained block was preceded by two practice blocks of 32 questions each.

Finally, in the uncued block, pictures were presented before the categorization question. The uncued block was pseudorandomized, with constraints that were identical to the randomized block, and it was also preceded by 32 practice trials.

Trial Procedure

In the randomized and sustained block, each trial began with the presentation of a question that remained on screen for 1500 msec. A fixation cross was then presented in the center of the screen for a variable interval of between 500 and 750 msec. After this interval, a picture was presented and remained on screen for 20 msec. After picture offset participants answered the question while the screen was blank, and, following their response and an additional 750 msec of blank screen, feedback was presented and stayed on screen for 1500 msec. In particular, participants were asked to decide whether the picture they had just seen matched the question that had been asked before the picture and indicate their choice by pressing one of two alternative keys (v or b) on a computer keyboard. Participants were required to respond in all trials, and the response key was counter-balanced across participants. After an intertrial interval of 500 msec, the next trial began. In the uncued block, an initial fixation cross (500 msec) was followed by a 20-msec picture and a 1000-msec blank interval. The question was then presented and stayed on screen until the participant responded. Subsequently, feedback was presented for 1500 msec.

EEG Recording and ERP Analysis

EEG was recorded at a sampling rate of 512 Hz from 62 active sites using an ActiveTwo Biosemi system. Three additional sensors were placed below the participant’s

left eye and laterally to the outer canthus of each eye to allow for detection of blinks and eye movements. The EEG was referenced to an additional reference electrode located near Cz during the recording. A hardware fifth-order low-pass filter with a -3 dB attenuation factor at 50 Hz was applied online. Off-line analysis was performed using Emegs (Peyk, De Cesarei, & Junghöfer, 2011). EEG data were initially filtered (0.1 Hz high pass and 40 Hz low pass), and eye movements were corrected by means of an automated regressive method (Schlögl et al., 2007). After this correction, the three additional sensors for eye movement correction were discarded. Trials and sensors containing artifactual data were detected through a semiautomatic procedure (Junghöfer, Elbert, Tucker, & Rockstroh, 2000). Trials containing a high number of neighboring bad sensors were discarded; for the rest of the trials, sensors containing artifactual data were replaced by interpolating the nearest good sensors. Finally, data were rereferenced to the average of all sensors, and a baseline correction was performed, based on the 100 msec before stimulus onset.

Data Analysis

General Strategy

The experimental design contained an orthogonal association between picture content, scenario, and number and the question that was asked. Specifically, this design allowed us to independently investigate bottom–up category differences or top–down decisional modulations. The analysis was organized along the main questions of the paper, which are summarized in Table 1.

RQ1. What Is the Role of Top–Down Attention, Independent of Bottom–Up Features?

This analysis focused on the modulation, which is related with target relevance, and therefore included a Decision factor (target vs. nontarget). In addition, to examine whether target-related differences depend on the question that is being asked and whether they vary according to the block structure (RQ4), we added Domain (content, scenario, or number) and Block (randomized, sustained or uncued) as additional factors.

RQ2. What Is the Role of Bottom–Up Features, Independent of Attention?

To examine bottom–up driven differences between scenes, we averaged ERPs depending on the contents, scenario, or number of elements that were present in the picture. Therefore, this analysis involved the comparison of animals versus vehicles, indoors versus outdoors, and one versus two elements. This analysis was carried out in the uncued condition to examine scene processing in a condition in which a categorization

question had not been asked and in which participants did not yet have a categorization template that could modulate scene processing.

RQ2.1. Are Bottom-Up Driven ERP Category Differences Related to a Categorization Based on Scene Statistics?

A further analysis was carried out to functionally characterize the ERP differences observed in RQ2. Specifically, we investigated the extent to which bottom-up ERP modulations correlate with the performance of a classification algorithm, which operates based solely on scene statistics of energy and clutter.

RQ3. Do Categorization Goals Modulate Bottom-Up Analysis?

A main objective of this study was to investigate whether an active categorization goal is able to modulate bottom-up processing. To this end, we replicated the analysis carried out for RQ2 on all blocks (both cued and uncued) and added an additional factor, Probed (probed, not probed or uncued), which indicated whether each bottom-up category (e.g., animal or vehicle) was probed by the categorization question (e.g., Is it an animal?) or not (e.g., Is it indoors?). This factor is exemplified in Table 1 and in Figure 1.

Q4. What Is the Role of Memory Systems in Maintaining the Categorization Goal?

Categorization questions were shown before the image in the randomized block, where they varied on a trial-by-trial basis, and in the sustained block, where they remained the same throughout 32 trials. Therefore, in the analysis of RQ1, we included an additional factor, Block (sustained, randomized, uncued), to investigate modulations, which are related to the maintenance of a categorization template through several trials. In RQ3, we compared the cued (randomized and sustained) blocks to the uncued block using the Probed factor (probed, not probed, uncued). In a preliminary analysis, we used a Probed factor with five levels (randomized probed, randomized not probed, sustained probed, sustained not probed, uncued) and observed similar patterns in the sustained and randomized blocks. In the reported analysis, the sustained and randomized blocks were therefore collapsed together.

Behavioral Responses

RTs were collected after each picture in the cued conditions (randomized and sustained) and after each question in the uncued condition. RTs were only analyzed for correct responses, and responses that were faster or slower than 2.5 *SD* from the individual mean were excluded from the analysis. Error rates and RTs were

analyzed according to RQ1 and therefore were averaged according to the question domain, the block in which each trial was presented, and whether the picture and the question matched or not (Decision factor: target vs. nontarget).

Statistical Analysis

In all analyses, data were analyzed using repeated-measure univariate ANOVAs with Huynh–Feldt correction. In all cases in which a significant main effect of a factor with more than two levels was observed, we proceeded with post hoc tests using paired-sample *t* tests. For all ANOVA effects, we calculated and have reported the partial eta squared, which reflects the proportion of variance that is accounted for by experimental manipulations. For data visualization, condition averages are accompanied by within-participant *SEMs* (Loftus & Masson, 1994), calculated following the procedure suggested by O'Brien and Cousineau (2014).

To collapse multidimensional ERP data for statistical analysis, the region and time intervals of interest were selected based on previous studies in the field of scene and object categorization (Groen et al., 2016; De Cesarei et al., 2013; Groen, Ghebreab, Prins, Lamme, & Scholte, 2013; Scholte et al., 2009; Rousselet et al., 2007; VanRullen & Thorpe, 2001). For the analysis of top-down modulations (RQ1), ERPs were averaged over lateral temporal areas in the 250–400 msec time interval, and hemispheric differences were analyzed by adding a Hemifield factor (left vs. right) to the ANOVA design. In the analysis on the effects of bottom-up features (Questions RQ2–RQ3), two time intervals were selected, namely from 80 to 180 msec and from 180 to 250 msec over central posterior sensors.

RESULTS

Behavioral Responses

Error rates and RTs are reported in Figure 3. In terms of categorization accuracy, we observed a main effect of Block, $F(2, 70) = 4.172, p = .021, \eta_p^2 = .107$, with fewer errors in the sustained block compared with both the randomized block, $t(35) = 2.37, p = .023$, and the uncued block, $t(35) = 2.979, p = .005$, and no difference between the randomized and the uncued block, $t(35) = -0.037, p = .971$. We also observed significant differences between Question domains, $F(2, 70) = 77.590, p < .001, \eta_p^2 = .689$, with fewer errors when the question concerned picture content and lowest accuracy when it concerned scenario, and significant differences between all question domains, all $t_s(35) > 6.497, p_s < .001$. No significant differences involving the Decision factor (target vs. nontarget) were observed.

Analysis on RTs revealed main effects of Block, $F(2, 70) = 34.758, p < .001, \eta_p^2 = .498$, with faster responses in the sustained block compared with the randomized and

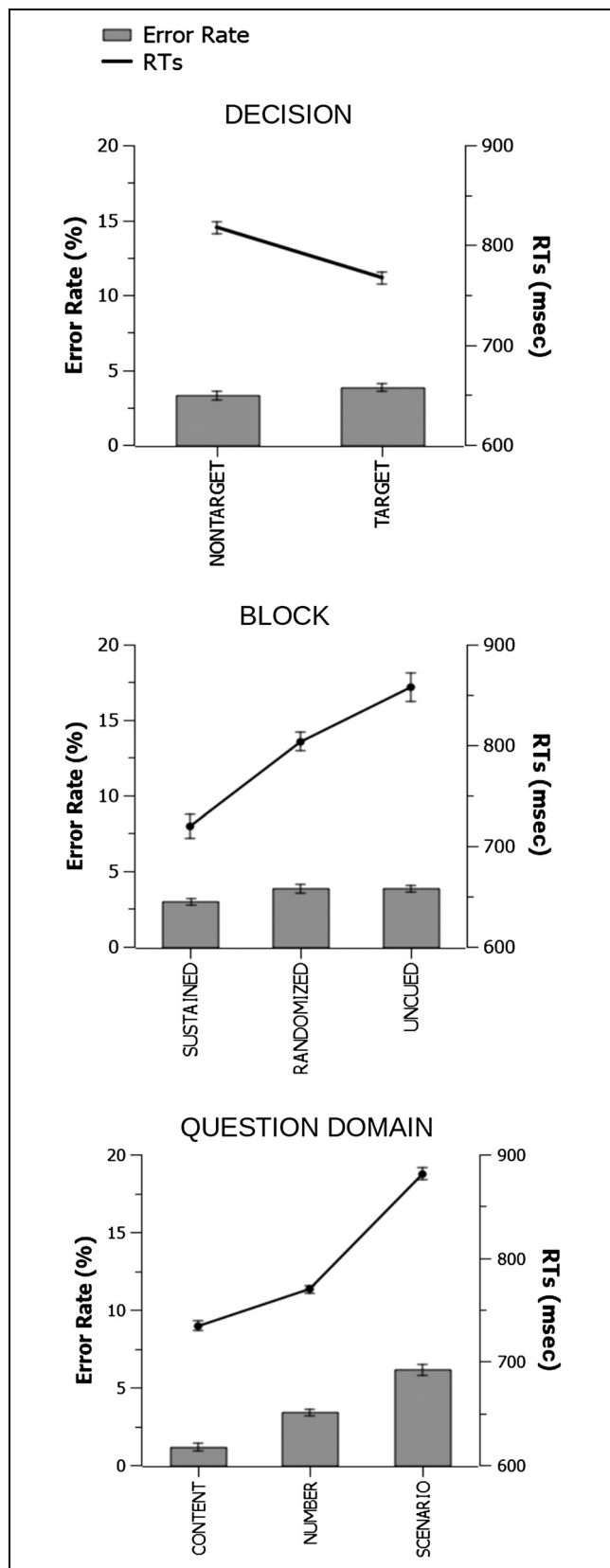


Figure 3. Behavioral results for RTs (lines) and error rate (bars), showing the effects of Decision (target vs. nontarget), Block (sustained vs. randomized vs. uncued), and Question domain (content vs. scenario vs. number). In each plot, error bars represent the SEM for within-participant designs.

uncued blocks, $t(35) > 6.307$, $ps < .001$, and more rapid responses in the randomized compared with the uncued block, $t(35) = 3.184$, $p = .003$. We also observed a Decision effect, $F(1, 35) = 35.140$, $p < .001$, $\eta_p^2 = .501$, with faster responses to target scenes compared with nontarget scenes, $t(35) = 5.678$, $p < .001$, and a Question domain effect, $F(2, 70) = 229.241$, $p < .001$, $\eta_p^2 = .868$, with faster responses to questions concerning scene content compared with those containing number of elements or scenario, $ts(35) > 7.367$, $ps < .001$, and more rapid responses to number of elements compared with scenario questions, $t(35) = 15.25$, $ps < .001$.

Finally, we observed an interaction between Block and Question domain, $F(4, 140) = 6.511$, $p < .001$, $\eta_p^2 = .157$. In all blocks, we observed a Question domain effect, $F(2, 70) > 110.256$, $p < .001$, $\eta_p^2 > .759$, with faster responses for questions concerning number than those regarding scenario and faster responses for content rather than scenario, $ts(35) > 4.73$, $ps < .001$. In the uncued block, no difference was observed between questions concerning content and number, $t(35) = .846$, $p = .403$.

ERP Results

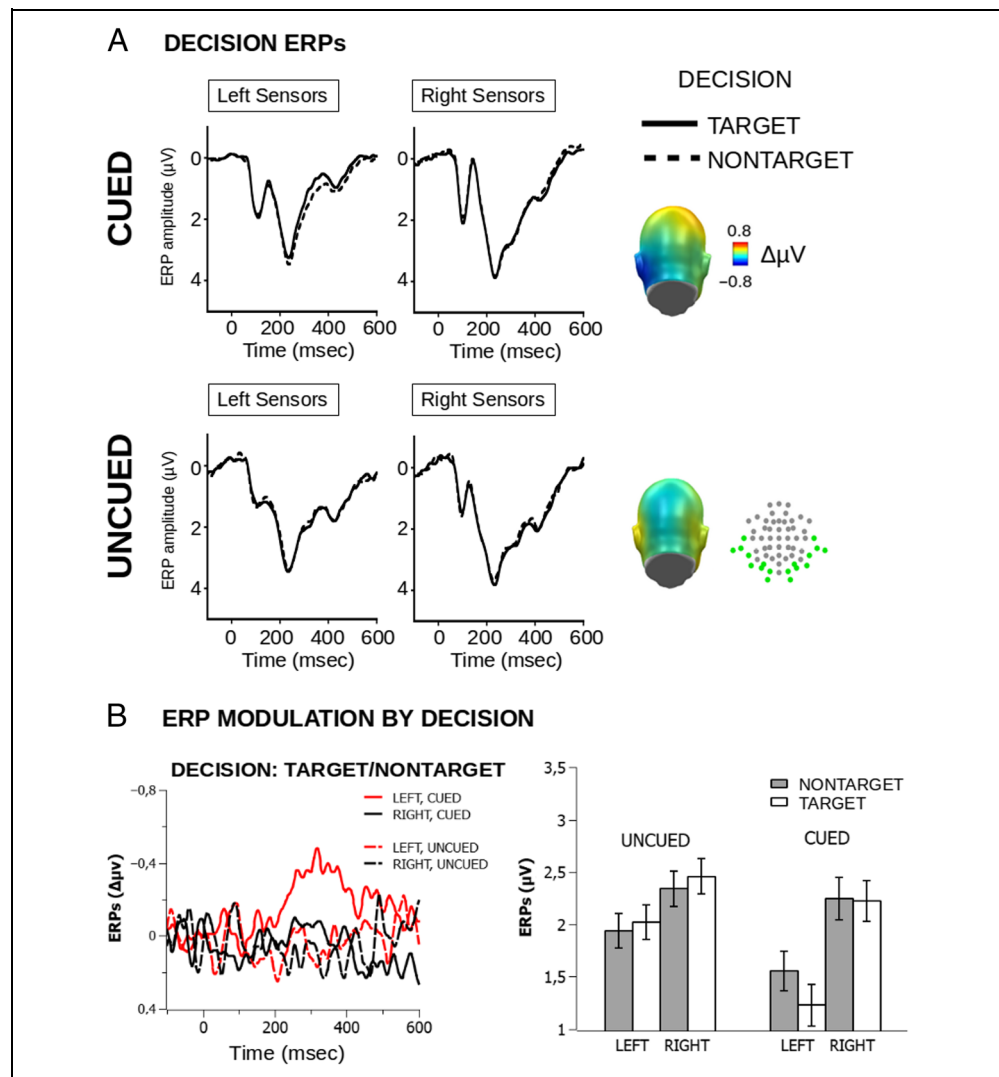
RQ1: What Is the Role of Top-Down Attention, Independent of Bottom-Up Features?

We analyzed the data set regarding Decision (targets vs. nontargets) to investigate the effects of the categorization task, independent of bottom-up differences between scenes (Figure 4A). The main result of this analysis was that, on left sensors, targets elicited less positive ERPs compared with nontargets in the sustained and randomized, $Fs(1, 35) > 5.793$, $ps = .021$, $\eta_p^2 > .142$, but not in the uncued, blocks (Figure 4B). The ANOVA results, analysis, and post hoc for the significant Hemisphere \times Block \times Decision interaction are reported in Table 3. In addition to the three-way interaction, a significant main effect of Block was observed, $F(2, 70) = 4.53$, $p = .017$, $\eta_p^2 = .115$, indicating more positive ERPs in the uncued compared with the randomized block, $t(35) = -2.624$, $p = .013$. A significant effect of Hemisphere, $F(1, 35) = 5.89$, $p = .021$, $\eta_p^2 = .144$, indicated more positive ERPs on right compared with left sensors, and significant two-way interactions between Hemisphere and Block, $F(2, 70) = 6.499$, $p = .003$, $\eta_p^2 = .157$, and Hemisphere and Decision, $F(1, 35) = 9.288$, $p = .004$, $\eta_p^2 = .210$, were also observed as a result of the three-way interaction described in Table 3. In this analysis, no significant interactions simultaneously involving factors Domain and Decision were observed.

RQ2: What Is the Role of Bottom-Up Features, Independent of Attention?

The effects of category were analyzed separately for each picture domain in the uncued block (Figure 5A).

Figure 4. The effects of Decision on ERPs. (A) ERP waveforms for target and nontarget scenes, separately in cued and uncued blocks, are represented. Sensors used for computing these average waveforms are reported in green on the sensor map. Topographic plots show the ERP difference between targets and nontargets, from a back of head view. (B) ERP waveforms for the target/nontarget difference are reported, along with a bar graph showing the mean and within-participant SEM of the scored 250–400 msec time interval, separately for left and right sensors and for the cued and uncued blocks.



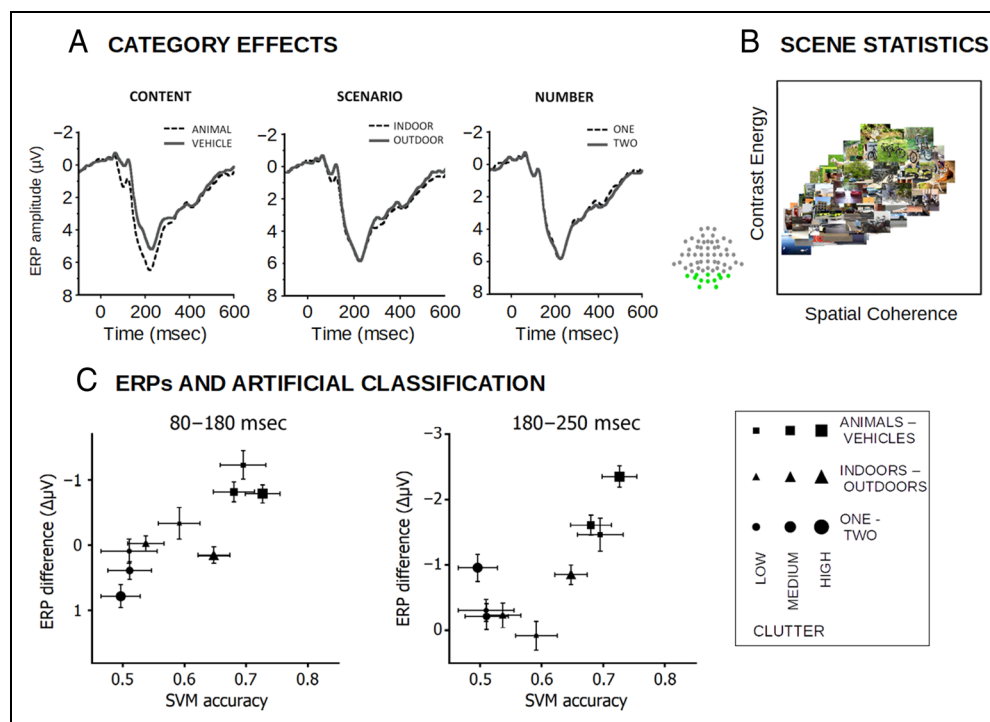
Pictures of animals elicited significantly more positive ERPs compared with vehicles in both time intervals. Indoor scenes also elicited more positive ERPs compared with outdoor scenes, but only in the 80–180 msec time

interval. For the number domain, no significant or close to significant main effect or interaction was observed. Statistical results for each comparison in the uncued condition are reported in Table 4.

Table 3. ANOVA Results of the Hemisphere \times Block \times Decision Analysis

	F	df Effect	df Error	p	η_p^2
Hemisphere \times Block \times Decision	4.189	2	70	.021	.107
Left					
Block \times Decision	3.599	2	70	.033	.093
Sustained, decision effect:	5.839	1	35	.021	.143
Randomized, decision effect:	5.793	1	35	.021	.142
Uncued, decision effect:	0.535	1	35	.469	.015
Decision	6.133	1	35	.002	.252
Right					
Block \times Decision	1.855	2	70	.164	.050
Decision	0.227	1	35	.636	.006

Figure 5. The effects of Category on ERPs. (A) ERPs in the uncued block are shown for all categorical differences. Sensors used for computing these average waveforms are reported in green on the sensor map. (B) Distribution of pictures in the scene statistic space defined by contrast energy and spatial coherence. (C) The relationship between early ERP category modulation and performance of a SVM, as a function of scene clutter. Symbol shape represents the category of each comparison, whereas symbol size represents the percentile of scene clutter based on spatial coherence. Each of the 9 points corresponds to the ERP modulation and SVM classification of 192 pictures (e.g., 96 animals and 96 vehicles). Vertical error bars represent within-participant SEM, whereas horizontal error bars show the SEM calculated across the 100 permutations on which SVM accuracy is computed.



RQ2.1: Are Bottom-Up Driven ERP Category Differences Related to a Categorization Based on Scene Statistics?

An SVM was trained to perform the same classification tasks as human participants based on scene statistics of energy and clutter (Figure 5B). The SVM performed above chance when categorizing according to the Content and Scenario dimension (content accuracy $M = 65.69\%$, $SEM = 1.90\%$; scenario accuracy $M = 58.35\%$, $SEM = 2.01\%$) but not for the number dimension (accuracy $M = 51.13\%$, $SEM = 1.85\%$).

SVM accuracy values were correlated with the category-specific ERP modulation in the 80–180 and 180–250 msec time intervals in the uncued condition (Figure 5C). To increase the points in this correlation while keeping enough trials and pictures in the ERP and machine learning analysis, the data set was further split into three percentiles based on spatial coherence. Ranking was conducted within each of the eight conditions defined by the combination of content, scene, and number, and each rank was equally present in the training and test set. Confidence intervals for ERP data are calculated using the suggested corrections for within-participant designs (O'Brien & Cousineau, 2014; Loftus & Masson, 1994). For both time intervals, a linear relationship between artificial algorithm performance and ERP modulation was observed, Pearson $r(9) = -.85$, $p = .004$, for the 80–180 msec time interval, and Pearson $r(9) = -.781$, $p = .012$, for the 180–250 msec time interval.

RQ3: Do Categorization Goals Modulate Bottom-Up Analysis?

The effects of goals on the bottom-up processing of scenes are reported in Figure 6. The main result of this analysis is that, in the 180–250 msec time interval, bottom-up differences for the content dimension (animals vs. vehicles) were more pronounced in the cued blocks when the categorization question did not probe picture content, compared with both the uncued, $F(1, 35) = 7.666$, $p = .009$, $\eta_p^2 = .18$, and the nonprobed condition, $F(1, 35) = 3.667$, marginally significant $p = .064$, $\eta_p^2 = .095$ (Figure 6A). The ANOVA results supporting this interaction are reported in Table 4. Moreover, differences between indoor and outdoor scenes were observed in the 180–250 msec time interval for the cued blocks, $F(1, 35) = 5.924$, $p = .020$, $\eta_p^2 = .145$, but not for the uncued block, $F(1, 35) = 0.554$, $p = .462$, $\eta_p^2 = .016$ (Figure 6B; Table 4). However, no interaction simultaneously involving factors Category and Probe, or Category and Block, was observed in this analysis. Finally, no significant effects or interactions involving the Probe or Category factors were observed for the number domain in either time interval.

DISCUSSION

This study investigated the categorization of natural scenes and focused on the role of task goals. Specifically, the same scenes were categorized under different task

Table 4. ANOVA Results of the Analysis of Category Effects in the Uncued and Cued Blocks, Separately for the Content Category (Animal vs. Vehicle), Scenario (Indoors vs. Outdoors), and Number (One vs. Two)

Category	Research Question	Time of Interest	ANOVA Effect	Post hoc	F	df Effect	df Error	p	η_p^2
Content category: animals vs. vehicles	RQ2. Uncued	80–180 msec	Category		94.721	1	35	<.001	.73
			Probe × Time × Category		7.691	2	70	<.001	.18
	RQ3	180–250 msec	Category		174.203	1	35	<.001	.833
			Probe × Category		4.438	2	70	.015	.113
				Probed vs. uncued, Category effect	1.053	1	35	.312	.029
				Not probed vs. uncued, Category effect	7.666	1	35	.009	.18
				Not probed vs. probed, category effect	3.667	1	35	.064	.095
				Category (overall)	81.108	1	35	<.001	.699
				Category (probed)	46.901	1	35	<.001	.573
				Category (not probed)	96.476	1	35	<.001	.734
			Category (uncued)	50.955	1	35	<.001	.593	
Scenario category: indoors vs. outdoors	RQ2. Uncued	80–180 msec	Time × Category		8.919	1	35	.005	.203
			Category		15.337	1	35	<.001	.305
	RQ3	180–250 msec	Category		0.554	1	35	.462	.016
			Time × Category		7.072	1	35	.012	.168
				Category (cued blocks only)	33.01	1	35	<.001	.485
				Category (cued blocks only)	5.92	1	35	.02	.145

Table 4. (continued)

Category	Research Question	Time of Interest	ANOVA Effect	Post hoc	F	df Effect	df Error	p	η_p^2
Number category: (one vs. two elements)	RQ2: Uncued		No main effects or interactions involving Category factor		<0.98	1	35	>.329	<.027
	RQ3		No main effects or interactions involving Category factor		<2.33	1 ^a	35 ^a	>.136	<.062

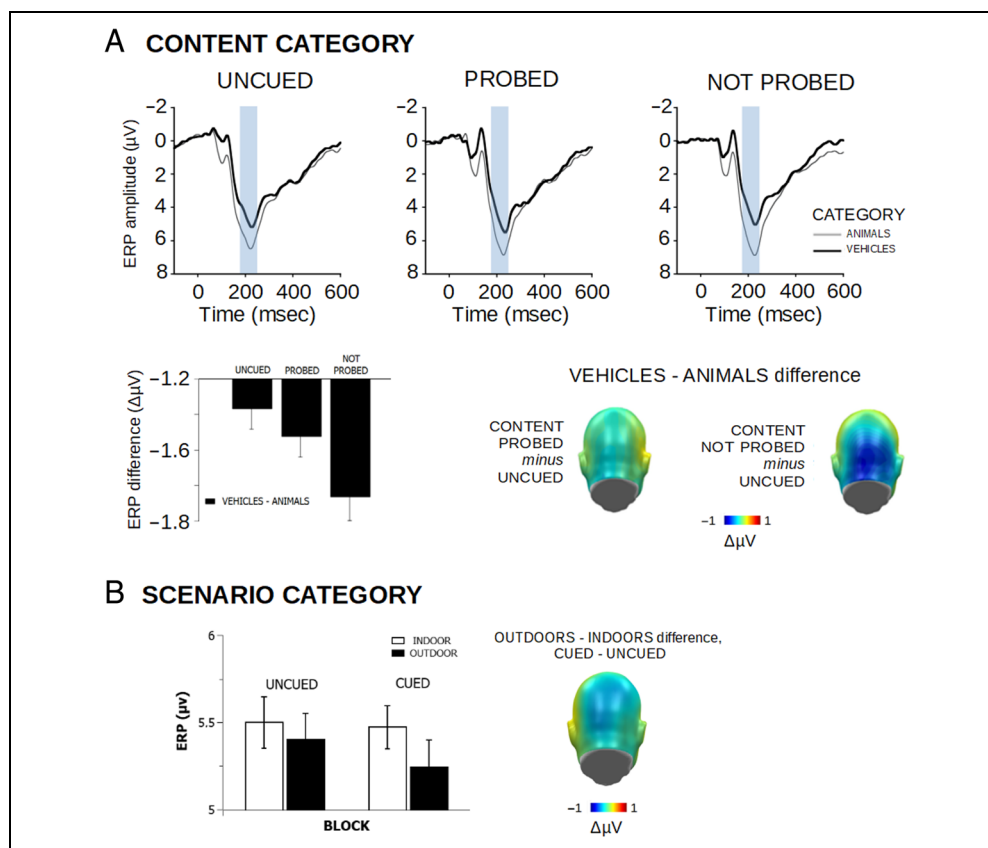
^aDegrees of freedom are shown for the ANOVA interaction Time × Category.

goals that were operationalized as different categorization questions. Initial visual processing was driven by scene statistics, and later stages reflected target processing. However, even early stages reflecting category-specific differences were modulated by categorization goals.

Decision-related and domain-independent effects were first observed after the peak of the P2, from 250 to 400 msec (RQ1). In this time interval, we did not observe any effect related to question or category domain, and less positive ERPs for targets compared with nontargets were observed on left temporal sensors. Therefore, this ERP modulation reflects processes related to the decision task that is being carried out and is category-general rather than category-specific. The direction of this ERP effect is reminiscent of the classic rapid categorization effect (De Cesare et al., 2015; Rousselet et al., 2007; Codispoti et al., 2006; VanRullen & Thorpe, 2001; Thorpe et al., 1996); in this study, however, it was considerably delayed and lateralized. A delayed task-related effect compared with sensory-driven ERP modulations was observed in a previous study, in which the target stimuli were animals or humans in different blocks (Rousselet et al., 2007). It has been suggested that the switch of top-down biases, which is required when multiple categorization goals are pursued throughout an experiment, is responsible for the late effects that were observed (Rousselet et al., 2007). However, in Rousselet et al.'s study, each block contained 192 trials, whereas in this study, the target category changed every 32 trials (sustained block) or every other trial (randomized block). Therefore, it is likely that a more flexible strategy was used by the system to provide optimal top-down bias in rapid scene categorization. This categorization strategy may have prevented top-down modulations from influencing earlier visual areas and restricted top-down decisional presetting to a relatively later stage of processing, which is reflected in the activity over temporal scalp areas (Rousselet et al., 2007). In this study, the decision-related ERP modulation was left-lateralized, which may suggest that verbal processes, such as object naming, are involved in this type of categorization task (e.g., Price, Moore, Humphreys, Frackowiak, & Friston, 1996). An intermediate stimulus representation, which is independent from visual attributes, may be adaptive when frequent changes in target stimulus require the system to change top-down bias (Rousselet et al., 2007). However, left lateralization on temporal or posterior regions was not reported in two previous studies, which used a multiple-category design (Harel et al., 2014; VanRullen & Thorpe, 2001), although more left than right sensors showed significant effects in Rousselet et al.'s (2007) study. Therefore, future studies are needed to further understand the role of lateralized processing in scene categorization.

Scene statistics and, in particular, statistics of contrast energy and spatial coherence modulated ERPs in the time range from 80 to 250 msec from stimulus onset

Figure 6. Modulation of Category effects, separately for the content and the scenario category. (A) Average ERPs for the animal and vehicle pictures are shown, separately for trials in which the picture was not preceded by any question (uncued), in which it was preceded by a question probing the “content” domain (i.e., “animal” or “vehicle”: probed) or by another question (content not probed). The bar graph shows the category differences in each of the three conditions. Scalp topographies report the ERP difference between the Category modulation in the probed and unprobed, compared with the uncued, conditions. (B) ERP modulation in the cued and uncued conditions for the Scenario category is reported from the same sensor group that is shown in Figure 5. The bar graph shows ERP averages and within-participant SEM for indoor and outdoor scenes in the cued and uncued blocks. The scalp topography shows the difference between the ERP modulation for scenario category (outdoors minus indoors) in the cued, compared with the uncued, condition.



(RQ2). This modulation is consistent with previous data (Groen et al., 2012; Scholte et al., 2009), which indicated that ERP activity at about 100 msec after stimulus onset can be predicted based on these scene statistics. Moreover, categorical differences were observed within some, but not all, picture domains. More specifically, we observed that, within each domain, different categories were associated with differently pronounced ERP modulations, which were maximal for content (animals vs. vehicles), intermediate for scenario (indoors vs. outdoors), and absent for number (one vs. two). What does this ERP modulation reflect (RQ2.1)? Here, we observed a clear correlation between these ERP modulations and the accuracy of an artificial categorization algorithm, which categorized scenes solely on the basis of contrast statistics. In this time interval, ERP category modulations were associated with the processing of category-specific visual regularities in contrast profile (Groen et al., 2012; Scholte et al., 2009). Importantly, contrast statistics are by no means the only factor that modulates early ERPs, and a number of sensory and perceptual factors, including overall spectral power (De Cesarei et al., 2013), figure completion (Hazenbergh & Van Lier, 2015), and symmetry (Bertamini & Makin, 2014), modulate early ERPs. The present results indicate

that ERP modulation in the 80–250 msec time interval reflects the processing of contrast statistics not only as a main modulatory effect (Scholte et al., 2009) but also in terms of categorical differences.

The analysis of bottom-up regularities and the guidance by top-down goals interacted in the categorization of natural scenes, and this interaction began early in time (180 msec; RQ3). Similarly, in a previous study, participants viewed natural scenes, and the bottom-up driven difference between manmade and natural scenes was suppressed when participants performed an orthogonal letter discrimination task or a 2-back memory task (Groen et al., 2016); however, no task-related difference in the bottom-up driven difference between manmade and natural scenes was observed before 250 msec. Although the exact latency for the observation of task effects may depend on differences in task design or demands, these results are consistent in indicating that top-down presetting is able to modulate bottom-up driven processing. Moreover, the current results indicate that not only the presence of a picture task but also the specific categorization goal may modulate the use of scene statistics as reflected by ERPs as early as 180 msec after scene onset.

Scenes are rich sources of information, and goal-related attention can be directed to different features of a scene (Malcolm et al., 2016). In particular, a distinction has been made between vision at a glance, which rapidly captures the categorical belonging or gist of a scene, and vision with scrutiny, which happens after initial perception and allows observers to focus on nondefault scene characteristics (Hochstein & Ahissar, 2002). In vision at a glance, goal-directed attention is not required, and default categorization is carried out even in the absence of specific requirements (Rosch et al., 1976). The present results showed that the categorization, which was carried out in the uncued block, which was not guided by task goals, did not differ from the categorization of the foreground object; in other words, participants in the uncued block categorized pictures according to the category (animal/vehicle) of the foreground object. During vision with scrutiny, however, when task goals required participants to focus on scene properties such as object number or scene location, recurrent processing may take place and require participants to keep processing the same scenes, resulting in more pronounced, content-specific modulations. Similarly, scenario-specific ERP differences in the 180–250 msec time interval were only evident when the system was actively engaged in cued categorization, and recurrent processing may have directed attention to a nondefault characteristic such as scenario (Kadar & Ben-Shahar, 2012). Recurrent processing refers to the activity of reentrant connections from higher-order visual areas, such as parietal or frontal areas, which interact with early visual areas. It has been suggested that recurrent processing plays a critical role in scene categorization (Koivisto, Railo, Revonsuo, Vanni, & Salminen-Vaparanta, 2011; Lupyán, Thompson-Schill, & Swingley, 2010; Hochstein & Ahissar, 2002), as well as in visual awareness (Lamme & Roelfsema, 2000). In particular, the Reverse Hierarchy Theory of visual perception (Hochstein & Ahissar, 2002) suggests that the outcomes of the processing of low visual areas are, by default, nonaccessible to explicit perception. When task demands require further scrutiny, attention can turn back to low visual areas, until task-relevant information is accessed (Hochstein & Ahissar, 2002).

Here, we did not observe significant differences between the sustained and randomized categorization conditions, in terms of ERP modulations (RQ4). In terms of behavior, better performance was observed in the sustained compared with the randomized context and in the cued compared with the uncued blocks (Evans, Horowitz, & Wolfe, 2011). In a previous study examining the effects of varied versus blocked target change in a visual search task (Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004), behavioral facilitation was first observed 200 msec after the presentation of the category cue. Moreover, extensive practice over a 3-week period did not modulate early ERP differences (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001). Taking these results

together, this early categorization seems to reflect an attentional set, which is aimed at attaining actual goals, whereas long-term goals, which are supported by memory processes, come to play later on.

Categorization of scene content (animals vs. vehicles) was faster and more accurate compared with decision-making regarding both the scenario (indoors vs. outdoors) and the number of foreground elements (one vs. two). Previous studies varied categorization demands to examine the speed and efficiency of the visual system. The tasks carried out included those comparing basic and superordinate categorization (Mack & Palmeri, 2015; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009), tasks based on the processing of global scene properties (Hansen et al., 2018; Greene & Oliva, 2009a), and tasks probing the role of perceptual and functional features (Groen et al., 2018; Greene & Hansen, 2017). Global scene properties (including openness, naturalness, and others; Oliva & Torralba, 2001) and, in general, scene statistics (De Cesare et al., 2017; Simoncelli & Olshausen, 2001) allow the visual system to rapidly understand the gist of a scene (Greene & Oliva, 2009b); for instance, similarly low exposure times were required to make accurate judgments about these properties, including naturalness or openness (Greene & Oliva, 2009a). Our results are apparently at odds with the previous observations of efficient categorization of the “openness” property, as we observed a better performance for content categorization (animal vs. vehicles) than for scenario (indoors vs. outdoors) categorization. However, the present categorization questions do not fully coincide with global scene properties. Specifically, although the animal/vehicle task can be carried out based on the naturalness dimension, the same is not true for the indoor/outdoor task; an “outdoor” scene may be high in openness, as in the case of a landscape image of a beach, or low in openness, as in the case of a courtyard with trees blocking the horizon. Moreover, another study reported slower responses for scenario decision (Is it indoors/outdoors?) compared with content (Is it an animal/vehicle?) decision (Kadar & Ben-Shahar, 2012). Future studies may more specifically target the role of global scene properties, in addition to perceptual and functional characteristics (Groen et al., 2018; Greene & Hansen, 2017; Fei-Fei, Iyer, Koch, & Perona, 2007), to determine the timing of extraction of several types of information from a visual scene.

Limitations and Future Directions

Although the present results highlight the interaction between bottom-up analysis and top-down modulation, further studies are necessary to gain more insight into how these processes interact in the categorization of natural scenes. For instance, as far as top-down modulations are concerned, one avenue of future research could compare goals that are based on global scene properties (e.g.,

high/low in openness) with goals based on other classifications (e.g., indoors/outdoors; Kadar & Ben-Shahar, 2012; Fei-Fei et al., 2007), as has already been suggested above. On the bottom–up end of the interaction, perceptual features are central in determining the efficiency of processing of a natural scene. For instance, the relative area, which is subtended by the foreground and the background elements, may be important in determining the ease or difficulty of tasks that require participants to decide on the object content (foreground element) or on the scene location (background). In the endeavor of investigating top–down and bottom–up contributions to scene categorization, recent methods such as machine learning and EEG/MEG decoding approaches may provide important information. For instance, deep convolutional neural networks rival human participants in accuracy and show functional properties, which are reminiscent of the organization of the visual system (VanRullen, 2017; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016). Therefore, deep convolutional neural networks might help us to understand which scene features are most diagnostic for the task at hand (Schyns, 1998). At the same time, decoding EEG differences during categorization tasks may help to understand the informational value of EEG differences across the whole scalp during natural scene categorization (Fahrenfort, Van Driel, Van Gaal, & Olivers, 2018; Haxby et al., 2001).

The role of inferences in vision has been an object of study since the 19th century (Von Helmholtz, 1867) and is represented in the current debate on the principles that govern brain functioning (Friston & Kiebel, 2009; Bar, 2003). In vision, top–down inferences are most evident when the bottom–up information is present but degraded (Schendan & Ganis, 2015; Gregory, 1970). Categorization goals may provide the system with a template of the search item or feature, which may guide perception. Therefore, effects of top–down goals might be more evident when participants are confronted with degraded (e.g., with alterations in the spatial frequency domain such as low-pass filtering, high-pass filtering, or phase scrambling), compared with intact, scenes.

The present results were observed with a relatively brief exposure time of 20 msec. It has been suggested that the processing of contextual information and the preference for a basic or a superordinate level of categorization may depend on the exposure time of a scene (Vanmarcke, Calders, & Wagemans, 2016; Mack & Palmeri, 2015). If exposure time determines different perceptual goals, then these different goals may engage the system in the selection and analysis of different features of a visual scene (Malcolm et al., 2016) and ultimately result in modulation of early scene processing. It is possible that, with a longer exposure time, figures are better separated from the background (Lamme, Zipser, & Spekreijse, 2002) and context effects are observed. However, several studies indicate that the processing that can be carried out with a short exposure time, when the

picture is not backward masked, does not differ greatly from that carried out with a longer exposure time (Busey & Loftus, 1994; Loftus, Duncan, & Gehrig, 1992; Loftus & Ginn, 1984). Therefore, future studies could investigate whether exposure time and backward masking modulate early scene categorization or whether the iconic scene representation that is obtained from a brief exposure suffices for scene categorization.

Conclusions

Categorization goals are central in active vision, and here, they modulated the use of scene statistics beginning from an early stage of visual processing. We discussed the direction of this effect in relation to a shift from a default “content” categorization goal to a non-default and possibly more complex goal, such as the indoor/outdoor distinction. Goals may involve either default processing modes, which are aimed at quickly making sense of the world, or task-specific settings, which hijack existing processing stages to accomplish current aims. Our comprehension of scene understanding, including early stages of vision, can be greatly improved when not only the presence or absence of goals but also the type of goal is taken into account.

Reprint requests should be sent to Andrea De Cesarei, Department of Psychology, University of Bologna, Viale Berti Pichat 5, 40127 Bologna, Italy, or via e-mail: andrea.decesarei@unibo.it.

REFERENCES

- Bar, M. (2003). A cortical mechanism for triggering top–down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*, 600–609.
- Bertamini, M., & Makin, A. D. J. (2014). Brain activity in response to visual symmetry. *Symmetry*, *6*, 975–996.
- Busey, T. A., & Loftus, G. R. (1994). Sensory and cognitive components of visual information acquisition. *Psychological Review*, *101*, 446–469.
- Cauchoix, M., & Crouzet, S. M. (2013). How plausible is a subcortical account of rapid visual recognition? *Frontiers in Human Neuroscience*, *7*, 39.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.
- Codispoti, M., Ferrari, V., Junghöfer, M., & Schupp, H. T. (2006). The categorization of natural scenes: Brain attention networks revealed by dense sensor ERPs. *Neuroimage*, *32*, 583–591.
- Crouzet, S. M., & Serre, T. (2011). What are the visual features underlying rapid object recognition? *Frontiers in Psychology*, *2*, 326.
- De Cesarei, A., & Codispoti, M. (2015). Can the outputs of LGN Y-cells support emotion recognition? A computational study. *Computational Intelligence and Neuroscience*, *2015*, 695921.
- De Cesarei, A., Loftus, G. R., Mastria, S., & Codispoti, M. (2017). Understanding natural scenes: The contribution of image statistics. *Neurosciences & Biobehavioral Reviews*, *74*, 44–57.

- De Cesarei, A., Mastria, S., & Codispoti, M. (2013). Early spatial frequency processing of natural images: An ERP study. *PLoS One*, *8*, e65103.
- De Cesarei, A., Peverato, I. A., Mastria, S., & Codispoti, M. (2015). Modulation of early ERPs by accurate categorization of objects in scenes. *Journal of Vision*, *15*, 14.
- Enns, J. T. (2004). Object substitution and its relation to other forms of visual masking. *Vision Research*, *44*, 1321–1331.
- Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide accumulation of information about multiple categories in rapid scene perception. *Psychological Science*, *22*, 739–746.
- 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. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience*, *13*, 171–180.
- Fahrenfort, J. J., Van Driel, J., Van Gaal, S., & Olivers, C. N. (2018). From ERPs to MVPA using the Amsterdam Decoding and Modeling toolbox (ADAM). *Frontiers in Neuroscience*, *12*, 368.
- Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*, 10.
- Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America, A, Optics and Image Science*, *4*, 2379–2394.
- Fize, D., Boulanouar, K., Chatel, Y., Ranjeva, J. P., Fabre-Thorpe, M., & Thorpe, S. (2000). Brain areas involved in rapid categorization of natural images: An event-related fMRI study. *Neuroimage*, *11*, 634–643.
- Friston, K. J., & Kiebel, S. J. (2009). Cortical circuits for perceptual inference. *Neural Networks*, *22*, 1093–1104.
- Greene, M. R., & Hansen, B. (2017). Visual, functional, and semantic contributions to scene categorization. *Journal of Vision*, *17*, 552.
- Greene, M. R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*, 464–472.
- Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–176.
- Gregory, R. (1970). *The intelligent eye*. New York: Mac Graw-Hill.
- Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*, 1409–1422.
- Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2012). Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. *PLoS Computational Biology*, *8*, e1002726.
- Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, S. (2016). The time course of natural scene perception with reduced attention. *Journal of Neurophysiology*, *115*, 931–946.
- Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A., & Scholte, H. S. (2013). From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, *33*, 18814–18824.
- Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, *7*, e32962.
- Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *372*, 20160102.
- Hansen, N. E., Noesen, B. T., Nador, J. D., & Harel, A. (2018). The influence of behavioral relevance on the processing of global scene properties: An ERP study. *Neuropsychologia*, *114*, 168–180.
- Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, E962–E971.
- Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*, 40–48.
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430.
- Hazenberg, S. J., & van Lier, R. (2015). Disentangling effects of structure and knowledge in perceiving partly occluded shapes: An ERP study. *Vision Research*, *126*, 109–119.
- Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*, 791–804.
- Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision*, *3*, 499–512.
- Junghöfer, M., Elbert, T., Tucker, D. M., & Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology*, *37*, 523–532.
- Kadar, I., & Ben-Shahar, O. (2012). A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *Journal of Vision*, *12*, 1–17.
- Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *Journal of Neuroscience*, *31*, 2488–2492.
- Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience*, *23*, 571–579.
- Lamme, V. A. F., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *Journal of Cognitive Neuroscience*, *14*, 1044–1053.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. In P. J. Lang, R. F. Simons, & M. Balaban (Eds.), *Attention and orienting* (pp. 97–135). Mahwah, NJ: Erlbaum.
- Loftus, G. R., Duncan, J., & Gehrig, P. (1992). The time course of perceptual information that results from a brief visual presentation. *Journal of Experimental Psychology: Human Perception & Performance*, *18*, 530–549.
- Loftus, G. R., & Ginn, M. (1984). Perceptual and conceptual masking of pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 435–441.
- Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review*, *1*, 476–490.
- Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*, 513–536.
- Lupyan, G., Thompson-Schill, S. L., & Swingle, D. (2010). Conceptual penetration of visual processing. *Psychological Science*, *21*, 682–691.
- Macé, M. J. M., Joubert, O. R., Nespoulous, J. L., & Fabre-Thorpe, M. (2009). The time-course of visual categorizations: You spot the animal faster than the bird. *PLoS One*, *4*, e5927.
- Mack, M. L., & Palmeri, T. J. (2015). The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General*, *144*, 551–569.

- Malcolm, G. L., Groen, I. I., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, *20*, 843–856.
- Miyashita, Y. (1993). Inferior temporal cortex: Where visual perception meets memory. *Annual Review of Neuroscience*, *16*, 245–263.
- Nishijo, H., Ono, T., Tamura, R., & Nakamura, K. (1993). Amygdalar and hippocampal neuron responses related to recognition and memory in monkey. *Progress in Brain Research*, *95*, 339–357.
- O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology*, *10*, 56–67.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.
- Peyk, P., De Cesare, A., & Junghöfer, M. (2011). ElectroMagnetoEncephalography software: Overview and integration with other EEG/MEG toolboxes. *Computational Intelligence and Neuroscience*, *2011*, 861705.
- Philiastides, M. G., & Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex*, *16*, 509–518.
- Pratt, H. (2011). Sensory ERP components. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford handbook of event-related potential components* (pp. 89–114). New York: Oxford University Press.
- Price, C. J., Moore, C. J., Humphreys, G. W., Frackowiak, R. S. J., & Friston, K. J. (1996). The neural regions sustaining object recognition and naming. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, *263*, 1501–1507.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.
- Rossion, B., & Caharel, S. (2011). ERP evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. *Vision Research*, *51*, 1297–1311.
- Rousselet, G. A., Macé, M. J., Thorpe, S. J., & Fabre-Thorpe, M. (2007). Limits of event-related potential differences in tracking object processing speed. *Journal of Cognitive Neuroscience*, *19*, 1241–1258.
- Schendan, H. E., & Ganis, G. (2015). Top-down modulation of visual processing and knowledge after 250 ms supports object constancy of category decisions. *Frontiers in Psychology*, *6*, 1289.
- Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., & Pfurtscheller, G. (2007). A fully automated correction method of EOG artifacts in EEG recordings. *Clinical Neurophysiology*, *118*, 98–104.
- Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W., & Lamme, V. A. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision*, *9*, 29.
- Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition*, *67*, 147–179.
- Seger, C. A., & Peterson, E. J. (2013). Categorization = decision making + generalization. *Neuroscience and Biobehavioral Reviews*, *37*, 1187–1200.
- Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.
- Sokolov, E. N. (1963). *Perception and the conditioned reflex*. New York: Macmillan.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.
- Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.
- Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, *14*, 411–443.
- Vanmarcke, S., Calders, F., & Wagemans, J. (2016). The time-course of ultrarapid categorization: The influence of scene congruency and top-down processing. *i-Perception*, *7*, 2041669516673384.
- VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142.
- VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*, 454–461.
- Von Helmholtz, H. (1867). *Handbook of physiological optics* (Vol. 3). Leipzig: Leopold Voss.
- Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 419.
- Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research*, *44*, 1411–1426.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3485–3492). San Francisco, CA, June.