1 Introduction

When you are walking from your bedroom to your kitchen, processing of your environment includes the layout of the furniture, recognizing items in each of those rooms, and even the types of tasks we do in each of those places (Castelhano & Heaven, 2011; Castelhano & Henderson, 2007, 2008b; Castelhano & Witherspoon, 2016; Henderson, 2003; Torralba et al., 2006; Williams & Castelhano, 2019). Scene perception seems seamless and effortless, belying many of the underlying processes that occur so that we can understand, interact with, and navigate everyday environments that vary in their content and scope (see Figure 1). We can also see how all these processes interact when we go grocery shopping and have a list of items to collect around that space. Not only do you need to navigate to the correct area of the supermarket, but you also need to distinguish among different types of items when you get there and choose an item based on any number of discriminating factors (Castelhano & Heaven, 2010; Castelhano & Henderson, 2003; Castelhano & Krzyś, 2020; Fernandes & Castelhano, 2021; Man et al., 2019). Another way to think about how we interact with scenes is not just within a single indoor space, but spaces as we explore different places. For instance, navigating in a new city is markedly different from navigating to a location that is highly familiar, such as a daily commute to work or class (Barhorst-Cates et al., 2016; Castelhano, Pollatsek, et al., 2009; Castelhano & Krzyś, 2020; Castelhano & Pollatsek, 2010; Epstein & Baker, 2019; Maguire et al., 2016). The arrays of different types of information available from the visual environment and how these are used across tasks demonstrate the complexity of scene perception and are as varied as the properties of scenes themselves (see Figure 1).

Even from a few examples, such as thinking about navigating a supermarket, we can easily see many ways in which scene perception as an area of research is a misnomer. Beyond the initial perceptual processing, scene perception research encompasses different types of processing of real-world environments, including attention, eye-movement guidance, memory, effects on other types of processing (i.e., context effects on object recognition), and spatial processing. Because of the complexity and enormity of the scene-processing literature, it is helpful to divide up the work into various sections. Here, we use the information-processing timeline and start with early perceptual processes before moving on to more complex and elaborate processing of scenes and space. We describe each of these, along with a summary of the six main sections of the Element (Initial Scene Understanding; Online Scene Representations; Long-Term Memory for Scenes; Eye Movements and Scenes; Searching through Scenes; and Spatial Representations and Navigation).

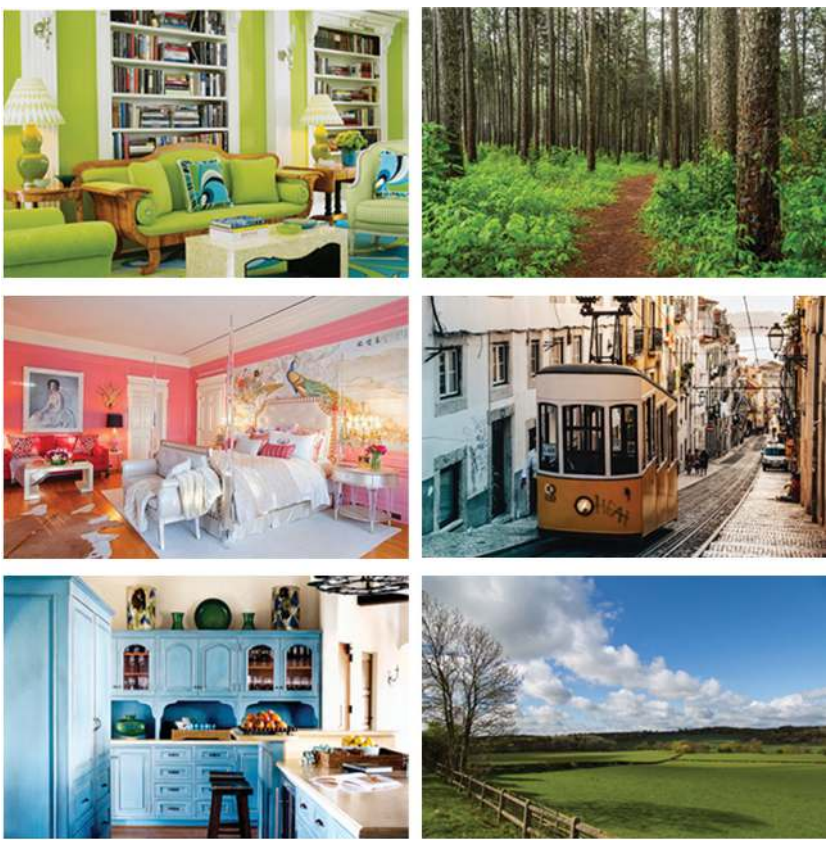

Figure 1 Different types of scenes are illustrated across example images. On the left-hand side are indoor scenes and on the right are outdoor scenes, which can be natural landscapes, cityscapes, or a combination of human-made and natural.

Think back to when you have been quickly scrolling through a streaming video service, where you only have a snapshot of each show in the form of an image. To choose, you have to quickly identify the image to determine if something looks interesting. In Section 2 we will examine how scenes are initially perceived and identified. If we use the processing timeline as a guide, scene perception begins with questions about how scenes are initially processed. In fact, when we think of scene perception, the first thing that comes to mind is how we are able to initially assess and understand the world around us. This is perhaps why the term is such a misnomer as these questions were first asked in the literature and the name persisted. This section will review what seemed at first like instantaneous processing, but is now understood as very fast processing that is thought to occur in a fraction of a second (Biederman, 1972;

Castelhano & Henderson, 2008a; Castelhano & Pollatsek, 2010; Goffaux et al., 2005; Greene & Oliva, 2009a, 2009b; Oliva & Schyns, 1997; Potter, 1976; Schyns & Oliva, 1994).

Going about the day, as we move around and do different tasks, the environment, and thus our representation of the environment, changes – in discreet as well as significant ways. In Section 3, we will examine the ways in which online representation was originally conceptualized and how this theoretical framework and tool is viewed today in light of more recent findings. While the notion that different types of information are prioritized across space and time when examining a scene is not new, it has implications for the ongoing representation of visual information as we view a scene.

Based on how information is prioritized for further scrutiny, researchers have also examined how information is represented from moment to moment (i.e., online representation). Online representation arises from a basic information-processing model of cognition, where information from the real world is acquired and reconstructed in the mind. Initially, the reconstruction of the world was thought to be quite veridical, such that we had an accurate and complete portrayal and understanding of the visual world around us. It is interesting that the importance of the internal representation is highlighted also in reference to eye movements. Because of the structure and mechanics of the eye, we see with high acuity only at the location to which the eyes are directly pointed; and yet, our perception of the world is that it is stable, ever present, and highly detailed across the whole of the visual field. This juxtaposition is solved with the notion of a veridical or, at the very least, highly accurate internal presentation of the world. While we are focused on one region, the representation supports the perception that the whole of the visual world is present in high detail. While intuitively appealing, this view was deemed to be impossibly difficult to compute (Gilman, 1994; Warren, 2012). We consider these and other approaches to how to think about scene representations that support perception and tasks in the moment in this third section.

After being in a space for some time and then leaving it, there is a question of how much of the information of the space and objects remains. Section 4 will examine the various points of view on how information is stored in memory. Building on the information prioritized in the moment and then held onto as a person explores a scene, researchers have also examined how information from scenes is stored in long-term memory. Much like online representations, researchers initially assumed that long-term representations (information kept in memory once a person has left a room or stopped viewing a room) were quite rich. This was supported by a number of studies showing how briefly viewed scenes could be recognized with high accuracy for some time (Potter, 1976;

Potter & Levy, 1969), even when the number of scenes held in memory was in the thousands (Mandler & Johnson, 1976; Shepard, 1967; Standing, 1973). This amazing feat of memory was thought to be possible only with highly detailed memory. Interestingly, while the notion did come under fire for some time, new research has once again swung the pendulum back – and shown that the information retained in long-term memory is more detailed (Konkle et al., 2010; Konkle & Brady, 2010). The nuance is now in how different details of the scene are more memorable, and how these different details lead to differences in retention of information over time. This fourth section will address how recent studies have shed further light on the nuances of different types of information represented in memory.

When walking down a city street, ads are looking to attract your attention from billboards, bus shelters, and posters stuck on walls and poles. Distractedly, you can feel yourself drawn to the images and words, or when you are focused on a task at hand (e.g., checking a text or driving), they can be utterly ignored. Section 5 will examine the influence of various sources of information on the allocation of attention and eye movements. Research into how attention is allocated is governed by the notion that attention was either pulled to attractive/distinctive regions (bottom-up influences) or pushed to useful or task-relevant regions (top-down influences). When examining the role of these different influences, eye movements indicate where and when a person is paying attention to different aspects of the scene and are an important tool in this research. This fifth section will review attentional processes and how eye movements can tell us something about them.

A simple task like making a cheese and cucumber sandwich requires us to locate and assemble the different components. This searching, whether for ingredients or tools, lies at the heart of so many tasks, and yet there are still a lot of questions about how we go about doing this successfully. Section 6 will examine how the various aspects of a scene affect performance, and how traditional notions of context can be broken down into different types of influences from the larger context. The interaction of information prioritized for further scrutiny and the information contained within the online representation is best encapsulated by the visual search task. In a visual search task, participants are given a target and then asked to locate it within the scene as quickly as possible. Traditionally, visual search tasks were investigated using arrays of shapes, but when searching in scenes, visual search performance is also affected by other factors (Castelhano & Heaven, 2011; Castelhano & Pereira, 2018; Loftus & Mackworth, 1978; Malcolm & Henderson, 2010; Võ & Henderson, 2011). This sixth section will examine these different factors influencing visual search and how it has evolved over the last decade.

Although most research to date has examined scene perception while participants are viewing scenes, we know that in our daily lives we process scene information while standing in them (Castelhano & Krzyś, 2020; Castelhano & Witherspoon, 2016; Gibson, 1979). Section 7 will examine how scene processing is influenced by spatial aspects of information. Research examining these aspects of scenes faces new and interesting problems and constraints. For instance, the spatial arrangement of structures and objects has to be kept in mind even when not in full view, as some of the information is behind the viewer. For this reason, many researchers have examined scene processing across viewpoints (Castelhano et al., 2008; Castelhano & Pollatsek, 2010; Epstein et al., 2003, 2005; Garsoffky et al., 2002; Li et al., 2016) and across different views of panoramic images (Garsoffky et al., 2002; Park et al., 2010). This seventh section will cover different aspects of spatial scene processing, from across different views and when incorporating information into a larger representation that extends beyond the current view.

Overall, in addition to examining a number of theoretical research domains in which the study of scene processing has been led, the current review will also look at how this research is applied to real-world examples. The "Application in the Real World" presented at the end of each section will highlight one example of how these fundamental questions about processing influence our understanding of other tasks and events. For instance, we will explore real-world problems such as how an advertisement is looked at, the veracity of eyewitness memory of a scene, the performance of radiologists in detecting problems in an x-ray, and the impact of pictures on the acceptance of fake news. The extensive research in scene processing can give insights into how these tasks operate, as well as the limitations of human performance under those task demands.

2 Initial Scene Understanding

Early studies showed how quickly information from real-world scenes could be understood. In a now seminal study, Potter (1976) showed participants a rapid sequence of briefly presented, unrelated images (referred to as Rapid Serial Visual Presentation or RSVP) and asked them before or after the sequence whether an image was present in the stream. Images were shown for as little as 113 ms each, mimicking a brief fixation on the image. The results revealed that when given a label for an image prior to viewing the stream of images, participants could easily identify the target image; however, when given the label afterwards, they could not. Together, the findings led to the conclusion that although images could quickly be processed to the point of interpretation and

understanding, the memory of that information is fleeting without additional time to consolidate it.

Potter's (1976) study was in line with other studies from that time that explored not only how quickly images were understood but also to what extent different types of information drove this rapid understanding (Biederman, 1972; Friedman, 1979; Loftus & Mackworth, 1978; Shepard, 1967). Based on research from the memory literature, Friedman (1979) proposed that scene representations are initially formed by an inference made based on the perceived objects, which were largely held to be the basic semantic unit of the scene across a number of studies (Biederman, 1972; Friedman, 1979; Loftus & Mackworth, 1978; Shepard, 1967). The rapid understanding of scenes found in Potter (1976), however, spurred researchers to move away from the notion of the object as the basic unit of understanding and to examine how different visual features contribute.

To examine how different aspects of the scene were perceived over time, Fei Fei Li and colleagues (Li et al., 2007) had participants view images for various durations (from an extremely brief 27 ms to a longer exposure of 500 ms). Based on this view, participants would then write a description of what they saw in the image. These open-ended responses allowed the researchers to extract and organize different descriptors into a hierarchical tree of attributes. Analysis of the responses showed that with longer durations, specific objects were included in the description as well as whole narrations as to what event the image may have captured. In contrast, with very brief exposures to the image, many of the descriptors centered on perceptual features, colors, and shapes. This focus on visual features is how many researchers have approached the initial processing of scenes and how they lead to the identification of the image. We turn to these studies next.

While examining basic visual features and their contribution to scene understanding, one notion that has been debated is which feature provides crucial information for identification: color vs. spatial frequencies. Both color and spatial frequency information are processed in the early visual cortex and are thought to be the basic components used to derive a visual representation of the environment (Castelhano & Henderson, 2008b; Greene & Oliva, 2009b; Larson & Loschky, 2009; Nijboer et al., 2008; Oliva & Schyns, 2000; Oliva & Torralba, 2001). Spatial frequency information is thought to convey scene structure information, with larger shapes conveyed by lower frequency and details and edges conveyed in high-frequency information (see Figure 2). Thus, either component (or both) could be used to derive identifying information when a scene image is first viewed.

Researchers have debated for some time whether the rapid understanding of scene images is driven by the color in scene images or edge-based contours

Figure 2 The same image shown with low spatial frequencies only (left side) and high spatial frequencies only (right side). Different spatial frequency bands are thought to convey different aspects of the scene, but all convey some information about the scene structure. See text for more details.

(Bacon-Mace et al., 2005; Biederman, 1988; Biederman & Ju, 1988; Castelhano & Henderson, 2008b; Delorme et al., 2000; Goffaux et al., 2005; Macé et al., 2010; Oliva & Schyns, 1997, 2000; Schyns & Oliva, 1994). Biederman (Biederman, 1988; Biederman & Ju, 1988) argued that because the information processing occurred so quickly, only contours and edges had an influence on initial scene processing, while color contributed minimally.

Further research into the contribution of edges and contours has shown that they are sufficient to support scene understanding. For instance, Schyns and Oliva (Oliva & Schyns, 1997; Schyns & Oliva, 1994) had participants identify hybrid images. Hybrid images were composites of two photographic images – one that occupied a low spatial frequency range (seen only as blurry contours) and another that occupied a high spatial frequency range (seen only as sharp edges and detailed textures). In an initial study, Schyns and Oliva (1994) found that when the hybrid images were briefly presented (~50 ms), participants tended to categorize the image in the low-frequency more readily than the high-frequency range. With longer exposures (~100 ms), this pattern flipped such that the high-frequency image was categorized more readily than the low-frequency image. Based on this pattern of results, the researchers concluded that image processing proceeds from blurred to more detailed image properties. However, in a follow-up study they used an implicit training method, such that images were presented rapidly either at high or low frequency (with the complementary frequencies presenting as white noise). This training encouraged participants to attend to either the high- or low-frequency image information exclusively, as the white noise did not offer any useful information. After training, participants were then shown the true hybrid images (with an image presented at each of the high- and low-frequency ranges) for both short and longer exposure durations. They found that participants tended to report the category of the image at the frequency range in which they were trained – in other words, both high and low image information were available at short exposure durations. This finding threw new light on previous results, as it was not the case that high-frequency information was not available at shorter exposure durations or took longer to process; instead, the progress of processing information from blurred to more detailed image properties was a mere preference or default of the system. Thus, the progression in how visual features from real-world images are processed and have an influence is not fixed, but rather subject to influences and changes due to tasks and circumstances.

Although many researchers have concentrated on how spatial frequencies and edge information contribute to the initial understanding of scene images, others have been interested in the possible contributions of color. As mentioned earlier, many early studies showed no evidence that color contributed to the

understanding of scenes – or rather that there was no discernible cost to presenting images without color (Codispoti et al., 2012; Delorme et al., 2000; Yao & Einhauser, 2008). For instance, Delorme et al. (2000) had participants classify images as to whether they contained fruit or an animal presented briefly (~32 ms). They showed images in both full color and grayscale (black-and-white images). They found that classification of images was only mildly impaired when color information was removed and concluded that color was not used to make these classification decisions.

The findings with scene images seemed in complete contradiction to many studies that examined the recognition of individual objects, which did find a benefit from color (Joseph & Proffitt, 1996; Mapelli & Behrmann, 1997; Price & Humphreys, 1989; Tanaka & Presnell, 1999). For instance, in one study, Tanaka and Presnell (1999) had participants categorize pictures of objects that were either in expected or unexpected colors. Importantly, they made a distinction between high-diagnostic objects, which are those that are highly associated with a specific color (e.g., a banana), and low-diagnostic objects, which are not associated with a specific color (e.g., a lamp). They found that when high-diagnostic objects were shown in colors other than the expected ones (monochrome or incongruent colors), performance was negatively affected. However, there were no effects on performance when low-diagnostic objects were shown in the different color conditions. They concluded that color does have an effect on the initial processing of objects, but only in certain cases – where there is an association between the object and its semantic category.

The results from the object recognition literature seem to contradict what was found in the scene literature, especially because scenes were largely conceptualized as collections of spatially arranged objects. One important difference between the studies examining contributions of color in scene images and in objects was in how and whether the color was linked to the semantic representation of the visual information being depicted. One approach to examining the contribution of color was to examine how it affected scene categories that were associated with specific colors (Castelhano & Henderson, 2008b; Castelhano & Rayner, 2008; Gegenfurtner & Rieger, 2000; Oliva & Schyns, 2000; Spence et al., 2006; Wichmann et al., 2002). In one study, Oliva and Schyns (2000) examined whether color had an important influence on scene perception when the scene colors were diagnostic of the scene category (see Figure 3). They had participants categorize scene images that were presented either in full color, no color, or abnormal colors. The abnormal colors were defined as those on the opposite side of the color space, such that each hue was swapped with its opposite (e.g., blues for yellows, etc.). Importantly, scene categories were distinct in the color space they occupied, such that

Figure 3 The images based on the different color conditions used in Oliva and Schyns (2000). They include (A) normal color, (B) abnormal color, and (C) monochrome images.

none of the color-diagnostic categories overlapped (e.g., coast, canyon, desert, forest). They found that not only were the colored images categorized more quickly than the no color images, but there was also a cost for images presented with abnormal colors. Oliva and Schyns conclude that for certain scene categories that are associated with a specific color space, color does contribute to the initial understanding of those images.

In another study, Castelhano and Henderson (2008) also investigated contributions of color to initial scene perceptions by examining whether the structure of the scene modulated the effectiveness of color contributions. They compared colored and grayscale images that were presented either with a full range of spatial frequencies (normal images) or with high spatial frequency removed (blurred images; see Figure 4). They found that when presented normally, there was no additional improvement in performance when images were presented in color over grayscale. However, when images were blurred, there was a significant improvement in performance for colored images over grayscale ones. Further experiments examined whether the color in blurred images was merely helping to define structure in the blurred images. When blurred images were also shown in abnormal colors there was no corresponding benefit in