

MIT Press • Open Encyclopedia of Cognitive Science

Scene Perception

Melissa Lê-Hoa Võ¹

¹Ludwig-Maximilians-Universität München

MIT Press

Published on: May 26, 2026

URL: <https://oecs.mit.edu/pub/mhjvjq8o>

License: [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License \(CC-BY-NC-ND 4.0\)](#)

As soon as you open your eyes, you perform something that is referred to as *scene perception*, a cognitive process by which the brain interprets and understands its visual environment. It involves recognizing, organizing, and extracting meaningful information from complex scenes, enabling humans to identify objects, their relationships, spatial layout, and relevant features within a given environment. This cognitive process is key to navigating and interacting effectively with our surroundings. A hallmark of scene perception is that it takes only a brief glimpse to extract rich meaning from complex visual input. The remarkable speed and ease with which we process scenes is the product of swiftly integrating bottom-up visual information with top-down scene knowledge. It is this fascinating cognitive feat that will be covered here.

What can you glimpse from these scenes?

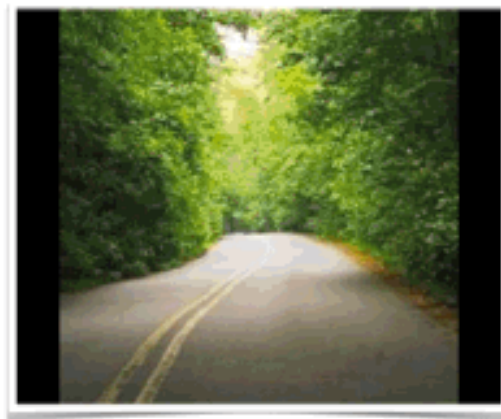


Figure 1

Demonstration of rich information extraction from only a glimpse of a scene.

History

Although formal research on “scene perception” as a defined concept emerged in the 20th century (see [Figure 1](#)), earlier philosophical and observational writings already touched on related ideas about visual perception and the human ability to interpret environments. Hermann [von Helmholtz's \(1867/1962\)](#) theory of unconscious inference, for instance, suggests that perception is an active process in which the brain makes implicit assumptions or inferences to interpret sensory information, particularly in complex or ambiguous scenes. This idea posits that our perception of scenes is not a direct reflection of the sensory

input but results from the brain's unconscious guessing and hypothesis testing based on prior knowledge and context.

In the early 1910s, Max Wertheimer and his colleagues Wolfgang Köhler and Kurt Koffka famously founded the Gestalt School (“Gestaltschule”) in Germany, which significantly contributed to understanding perceptual organization and gestalt principles (e.g., that “the whole is perceived as more than just the sum of its parts”). However, it was not until the 1970s and 80s that researchers like David Marr, Molly Potter, and Irv Biederman systematically put the feats of scene perception to a test. Their seminal works provided some core insights on scene perception, for example, that visual perception involves transforming retinal input into a series of increasingly abstract, computational representations that enable scene understanding ([Marr, 1982](#); see also [Groen et al., 2022](#)); that the briefest of glances at a complex scene is enough to process its meaning ([Potter, 1975](#); see also [Greene & Oliva, 2009](#)); and that *scene semantics* and *syntax* play a key role in the processing of objects therein ([Biederman, 1976](#); see also [Vö, 2021](#)).

Core concepts

The scene grammar framework

The scene grammar framework (see [Figure 2](#); [Vö et al., 2019](#); [Vö, 2021](#)) proposes that objects within scenes are hierarchically organized and constrained by *scene grammar*—internalized rules that predict what objects tend to be where within a scene, how objects are positioned relative to one another, and how objects relate to the global scene context. Within a scene, so-called *anchor objects* (e.g., the shower or bathtub) guide the placement, identification, and use of associated objects (e.g., the shampoo inside the shower or the towels next to it), forming meaningful and functionally optimized clusters or *phrases*. This highly structured composition of objects in scenes helps the brain efficiently process complex visual environments, boosting efficient object recognition and search as well as scene understanding, all of which enable us to efficiently adapt and function in a diverse range of settings ([Biederman, 1976](#); [Draschkow & Vö, 2017](#); [Henderson & Ferreira, 2004](#); [Vö & Henderson, 2009](#); [Vö & Wolfe, 2013](#)).

Scene Grammar

Hierarchical Predictions

Figure 2

The scene grammar framework.

What guides attention in real-world scenes?

A red poppy along a lush green hiking path might grab your attention, or a ball that your kid is launching your way. The view that attention is guided by low-level features like color contrast or movements was prominently featured in [Itti & Koch's \(2000\)](#) saliency model, but later accounts favored a greater role for top-down guidance of attention. For instance, [Nuthmann & Henderson \(2010\)](#) found preferred viewing location (PVL) close to the center of objects within naturalistic scenes, arguing that attentional selection in scenes is inherently object-based rather than saliency driven. Keeping visual saliency controlled, [Vö and Henderson \(2009\)](#), for instance, showed that objects violating scene grammar (e.g., a fire hydrant in a kitchen) keep attracting eye movements once fixated (it has been debated whether semantic or syntactic violations actually attract attention parafoveally, i.e., from the corner of your eye, or not; [Vö & Henderson, 2009](#)). And, more recently, deep neural networks like DeepGaze III ([Kümmerer et al., 2022](#))—which combines image information with the history of previous fixations—have been shown to predict where observers might look next while freely viewing a scene. Most of the time, however, we do not just passively view scenes but instead search for objects and interact with them for the purpose of goal-directed actions. Try searching for the shampoo in [Figure 3](#).



Figure 3
Search for the shampoo in this bathroom.

Most likely, scene grammar allowed you to ignore the left two-thirds of the scene (i.e., the toilet and sink phrases) and to limit your search to the shower phrase, likely to contain the target. And before deciding that the detergent is absent from this scene, you would likely open the cupboard and find it inside.

Questions, controversies, and new developments

As in many other disciplines, the rise of deep neural networks has opened up new possibilities for answering long-standing questions. A particular type of deep neural network—so-called generative adversarial networks, or GANs ([Goodfellow et al., 2014](#)), for instance—are becoming increasingly good at

“hallucinating” images of scenes that do not actually exist. What exactly is the scene grammar that GANs have learned ([see Kallmayer & Vö, 2024](#))? Lesioning GANs could serve as a useful testbed to investigate what actually makes a scene.

Lately, the classic notion of object affordances à la Gibson has been reintroduced to scene perception [see [Affordances](#)]. How do functions affect scene categorization ([Greene et al., 2016](#)), and which parts of a scene drive function understanding ([Müller Karoza et al., 2025](#))? Making use of more interactable three-dimensional virtual environments will likely provide new answers to some old questions.

Although often merely used as an analogy, processing objects in scenes and words in sentences might actually share common, domain-general cognitive mechanisms. Do, for example, children with developmental language delay also show impediments in the processing of pictorial information ([Bahn et al., 2025](#); [Lindfors et al., 2025](#))? Investigating the commonalities and differences between scene perception and language processing has the potential to improve the understanding of both of these core cognitive abilities (for some initial thoughts on this matter, see also [Henderson & Ferreira, 2004](#))

Broader connections

The study of scene perception is inherently interdisciplinary. As briefly outlined above, knowing a scene’s grammar will strongly affect visual search in naturalistic environments providing a highly efficient, top-down guidance of attention that allows detecting objects in scenes even if they are occluded or even hidden from view [see [Visual Search](#); [Attention](#)]. Also, people interested in memory have been intrigued by what features of a scene make it more or less memorable ([Guo & Bainbridge, 2025](#)), and recently deep convolutional neural networks trained on large datasets have been shown to successfully predict memorability scores for new images (for a review, see [Rust & Mehrpour, 2020](#)).

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under Germany’s Excellence Strategy (EXC 3066/1 “The Adaptive Mind,” Project No. 533717223) as well as by the DFG Research Unit FOR 5368.

Further reading

- Bartnik, C. G., & Groen, I. I. A. (2023). Visual perception in the human brain: How the brain perceives and understands real-world scenes. *Oxford Research Encyclopedia of Neuroscience*. <https://doi.org/10.1093/acrefore/9780190264086.013.437>
- Hansen, B. C., Greene, M. R., Lewinsohn, H. A. S., Kris, A. E., Smyth, S., & Tang, B. (2025). Brain-guided convolutional neural networks reveal task-specific representations in scene processing. *Scientific Reports*, 15, 13025. <https://doi.org/10.1038/s41598-025-96307-w>
- Wiesmann, S. L., & Vö, M. L.-H. (2025). Flexible usage of object and global scene information during human scene categorization. *Journal of Experimental Psychology: Human Perception and Performance*.

<https://doi.org/10.1037/xhp0001342>

References

- Bahn, D., Deniz Türk, D., Tsenkova, N., Schwarzer, G., Vö, M. L.-H., & Kauschke, C. (2025). Processing of scene-grammar inconsistencies in children with developmental language disorder—Insights from implicit and explicit measures. *Brain Sciences*, 15(2), 139. <http://doi.org/10.3390/brainsci15020139>

↩

- Biederman, I. (1976). On processing information from a glance at a scene: Some implications for a syntax and semantics of visual processing. *UDIGS '76: Proceedings of the ACM/SIGGRAPH Workshop on User-oriented Design of Interactive Graphics Systems*, 75–88. <https://doi.org/10.1145/1024273.1024283>

↩

- Draschkow, D., & Vö, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, 7, 16471. <https://doi.org/10.1038/s41598-017-16739-x>

↩

- Goodfellow, I, Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *arXiv*. <https://doi.org/10.48550/arXiv.1406.2661>

↩

- Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4), 464–472. <https://doi.org/10.1111/j.1467-9280.2009.02316.x>

↩

- Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M. & Fei-Fei, L. (2016) Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, 145(1), 82–94. <https://doi.org/10.1037/xge0000129>

↩

- Groen, I. I. A., Dekker, T. A, Knapen, T., & Silson, E. H. (2022) Visuospatial coding as ubiquitous scaffolding for human cognition. *Trends in Cognitive Sciences*, 26(1), 81-96. <https://doi.org/10.1016/j.tics.2021.10.011>

↩

- Guo, X. & Bainbridge, W.A. (2025). Visual memory for natural scenes. In J. Wixted, T. Abel, S. Fusi, M. Rugg, L. Mickes (Eds.), *Learning and Memory: A Comprehensive Reference*, 3rd edition. Academic Press.

↩

- Henderson, J. M., & Ferreira, F. (Eds.). (2004). *The interface of language, vision, and action: Eye movements and the visual world*. Psychology Press.

↩

- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10–12), 1489–1506. [https://doi.org/10.1016/S0042-6989\(99\)00163-7](https://doi.org/10.1016/S0042-6989(99)00163-7)

↩

- Kallmayer, A., & Vö, M. L.-H. (2024). Anchor objects drive realism while diagnostic objects drive categorization in GAN generated scenes. *Communications Psychology*, 2(1), 68. <https://doi.org/10.1038/s44271-024-00119-z>

↩

- Kümmerer, M., Bethge, M., & Wallis, T. S. A. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5), 7. <https://doi.org/10.1167/jov.22.5.7>

↩

- Lindfors, H., Hansson, K., Cohn, N., & Andersson, A. (2025). Similarities in semantic processing across verbal and pictorial domains in school children with developmental language disorder. *Frontiers in Psychology*, 16, 1548289. <https://doi.org/10.3389/fpsyg.2025.1548289>

↩

- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company.

↩

- Müller Karoza, L. A., Wiesmann, S. L., & Vö, M. L.-H. (2025). The role of anchor objects in scene function understanding. *Scientific Reports*, 15(1), 20247. <https://doi.org/10.1038/s41598-025-04122-0>

↩

- Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8), 20 1–19. <http://doi.org/10.1167/10.8.20>

↩

- Potter, M. C. (1975). Meaning in visual search. *Science*, 187(4180), 965–966. <https://doi.org/10.1126/science.1145183>

↩

- Rust, N. C. & Mehrpour, V. (2020). Understanding image memorability. *Trends in Cognitive Sciences*, 24(7), 557–568. <https://doi.org/10.1016/j.tics.2020.04.001>

↩

- Vő, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, 181, 10-20. <https://doi.org/10.1016/j.visres.2020.11.003>

↩

- Vő, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3), 24. <https://doi.org/10.1167/9.3.24>

↩

- Vő, M. L.-H., & Wolfe, J. M. (2013). Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science*, 24(9), 1816–1823. <http://doi.org/10.1177/0956797613476955>

↩

- Vő, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205-210. <https://doi.org/10.1016/j.copsyc.2019.03.009>

↩

- von Helmholtz, H. (1962). *Treatise on physiological optics* (J. P. C. Southall, Ed.). Dover Publications. (Original work published 1867)

↩