Controlled Synthetic Environments for Studying Mid-Level Vision

Joshua M. Martin (joshua.martin@tu-darmstadt.de)
Technical University of Darmstadt, Germany

Thomas S. A. Wallis (thomas.wallis@tu-darmstad.de)
Technical University of Darmstadt, Germany

Abstract

While artificial neural networks have advanced image-computable object recognition models, their ability to model mid-level vision remains limited. A key bottleneck is the scarcity of datasets with dense, high-quality annotations necessary for probing these intermediate computations. Here, we introduce a flexible synthetic image generation pipeline built in Blender that produces richly structured scenes with automatic pixel-level annotations, such as surface normals and segmentation masks. Drawing inspiration from digital embryos and dead leaves stimuli, the pipeline enables controlled manipulation of scene statistics and object properties. This provides targeted inputs for training and evaluating artificial neural networks, enabling detailed analysis of how specific visual features contribute to mid-level perceptual processes.

Keywords: computer vision; mid-level vision; segmentation; physical rendering

Introduction

Artificial neural networks (ANNs) have become valuable tools in vision science (Doerig et al., 2023), revealing both parallels and discrepancies with human perception (Wichmann & Geirhos, 2023). So far, ANNs have mainly been used to study object classification, aided by abundant, easily annotated object-centered datasets (Schrimpf et al., 2018; Rajalingham et al., 2018). Yet visual perception involves more than high-level recognition—mid-level processes like contour integration, figure-ground segmentation, and depth inference are critical for forming coherent scenes from low-level features (Anderson, 2020).

Despite its importance, mid-level vision remains underexplored in comparisons between artificial and biological systems (but see Doerig et al., 2020; Lonnqvist et al., 2025). One key challenge is the lack of suitable datasets. While datasets suitable for object recognition tasks are widely available, they often lack the necessary annotations—such as pixel-level segmentations or surface normals—that are required for studying mid-level processes. Moreover, the statistical variability of natural images makes it difficult to control for specific scene properties, limiting the ability to isolate and study certain visual phenomena in a systematic way.

A Rendering Pipeline for Mid-Level Vision

To address these challenges, we present a novel synthetic image generation pipeline built in Blender, a powerful open-source 3D rendering engine (see Figure 1). Our pipeline is designed to create large-scale, highly controlled datasets for studying mid-level visual processes. Unlike existing rendering pipelines (e.g. Gan et al., 2020) focused on realism or interactivity, our design prioritizes control of scene composition and object properties.

Our pipeline draws inspiration from two main paradigms. First, digital embryos, which are procedurally generated 3D objects that simulate embryonic development (Hauffen et al., 2012; Huber et al., 2024). These shapes exhibit complex geometry but lack semantic associations, making them ideal targets for generalization and shape analysis without relying on prior beliefs from standard object categories. Second, the "dead leaves" model, which generates images by sequentially layering random shapes and textures. This approach replicates key statistical properties of natural scenes (Lee, Mumford & Huang, 2001; Madhusudana et al., 2021) and produces occlusions and depth cues, making it well-suited for studying processes such as segmentation and grouping.

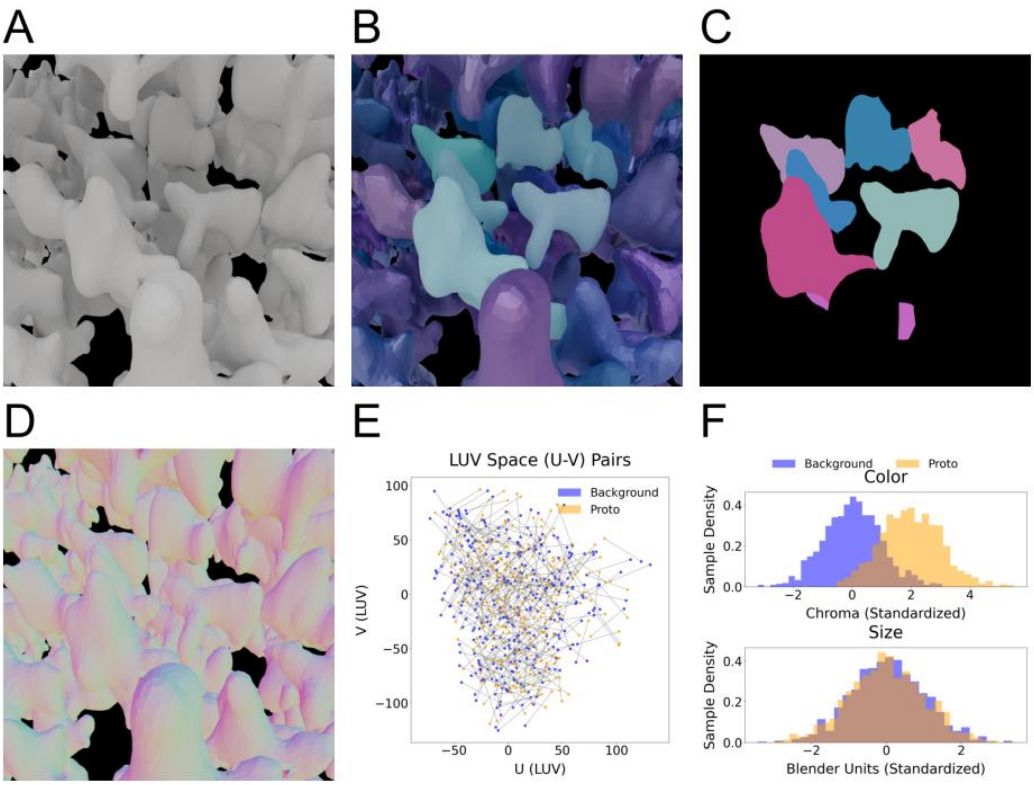

Figure 1. Digital embryos arranged in a dead leaves configuration (A), with color variation between proto and background objects (B), corresponding proto-object segmentation mask (C), surface normals (D), chroma pair sampling in LUV space (E), and feature variation histograms (F).

Quantifying Feature Contributions

By combining these two approaches, we generate richly varied scenes that are closer to real-world complexity than traditional laboratory stimuli, while remaining fully controllable and reproducible. The flexibility of the pipeline supports a wide range of experimental manipulations, such as varying object texture, illumination, material, and orientation in 3D space. For example, researchers can define sets of 'proto-objects' (i.e., targets of visual computations) that systematically vary along a single visual feature dimension (e.g., color) while holding others constant (e.g., size or shape, see Figure 1). Comparing model performance across such datasets allows one to evaluate the individual and combined contribution of different cues to processes, such as object segmentation, enabling hypothesis-driven experimentation in a tightly controlled setting.

Another major advantage of our synthetic approach is the automatic generation of rich, structured annotations for every rendered image. These include pixel-perfect segmentation masks, surface normals, depth maps, and comprehensive object- and scene-level metadata (see Figure 1, for examples). This ensures perfect alignment between inputs and ground truth, facilitating the creation of large-scale datasets at minimal cost. These annotations are critical for supervised learning, benchmarking model performance, and drawing meaningful comparisons between performance in artificial and biological visual systems.

Taken together, the features of our pipeline offer a powerful tool for enabling a more focused application of ANNs to study mid-level perceptual processes. Its precise control and rich annotations enable targeted hypothesis testing across a range of visual tasks.

References

Anderson, B. L. (2020). Mid-level vision. *Current Biology, 30*(3), R105–R109.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., ... & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience, 24*(7), 431–450.

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLoS Computational Biology, 16*(7), e1008017.

Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., ... & Yamins, D. L. (2020). Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint* arXiv:2007.04954.

Hauffen, K., Bart, E., Brady, M., Kersten, D., & Hegdé, J. (2012). Creating objects and object categories for studying perception and perceptual learning. *Journal of Visualized Experiments, (69)*, e3358.

Huber, L. S., Mast, F. W., & Wichmann, F. A. (2024). Comparing supervised learning dynamics: Deep neural networks match human data efficiency but show a generalisation lag. In *ICLR 2024 Workshop on Representational Alignment*.

Lee, A. B., Mumford, D., & Huang, J. (2001). Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision, 41*, 35–59.

Lonnqvist, B., Scialom, E., Gokce, A., Merchant, Z., Herzog, M. H., & Schrimpf, M. (2025). Contour Integration Underlies Human-Like Vision. *arXiv preprint arXiv:2504.05253*.

Madhusudana, P. C., Lee, S. J., & Sheikh, H. R. (2021). Revisiting dead leaves model: Training with synthetic data. *IEEE Signal Processing Letters, 29*, 209–213.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience, 38*(33), 7255–7269.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007.

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science, 9*(1), 501–524.