Active Retinal 3D Perception

Foveated Binocular Sensing, Vergence, and Saccadic Scene Understanding

Luiz Velho GPT, OpenAI (collaborator)

September 2026

Abstract

This technical note proposes an active computer-vision architecture inspired by the human visual system. A continuous radiance field is measured through a variable-resolution neural retina composed of Gaussian receptive fields. Two such retinas form a binocular sensor whose fixation, vergence, and focus are actively controlled. A vergence sweep supplies coarse depth hypotheses through left–right correlation; subsequent saccades collect high-resolution evidence over the surface of an object of interest. The accumulated measurements update a persistent, object-centered representation containing geometry, appearance, segmentation, and semantic belief. The central hypothesis is that depth and recognition can emerge from a closed perception–action loop instead of a single, uniformly sampled image.

1 Motivation

Conventional vision systems normally process every image location at a fixed resolution and estimate depth or semantics in a single forward pass. Human vision follows a different strategy: spatial resolution is highest near the fovea, the eyes actively fixate and verge, and rapid saccades gather detailed evidence only where it is useful. Perception is therefore inseparable from sensor control and temporal integration.

The proposed system uses this principle to recover depth, segment objects, and identify their semantic classes. It continuously chooses

(i) where to look,

(ii) at which spatial resolution to sample,

(iii) at which depth to verge and focus, and

(iv) which surface region to inspect next.

The resulting research direction may be called *active foveated binocular perception for semantic 3D scene understanding.*

2 Continuous Variable-Resolution Retina

Let $L^e(\boldsymbol{x}, t)$ denote the continuous image irradiance or radiance observed by eye $e \in \{L, R\}$ at retinal coordinate $\boldsymbol{x} \in \Omega \subset \mathbb{R}^2$ and time t. Retinal unit i produces the measurement

$$r_i^e(t) = \int_\Omega L^e(\boldsymbol{x}, t)\, \phi_i(\boldsymbol{x}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i)\, d\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{c}_i$ is the receptive-field center and $\boldsymbol{\Sigma}_i$ controls its size, shape, and orientation. A normalized Gaussian basis is a convenient choice:

$$\phi_i(\boldsymbol{x}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{c}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{c}_i)\right]}{2\pi\sqrt{\det \boldsymbol{\Sigma}_i}}. \tag{2}$$

This formulation views the retina as adaptive Gaussian quadrature over a continuous visual field. It is differentiable with respect to receptive-field location and scale, which permits both active fixation and end-to-end learning.

2.1 Foveated layout

Let $\boldsymbol{f}$ be the fixation point and $e_i = \|\boldsymbol{c}_i - \boldsymbol{f}\|_2$ the eccentricity of retinal unit i. Variable resolution can be imposed by increasing the receptive-field scale with eccentricity, for example

$$\sigma_i = \sigma_{\min}(1 + \kappa e_i), \qquad \kappa > 0. \tag{3}$$

The receptive-field centers may be distributed on a log-polar lattice using

$$\rho_i = \log(1 + \alpha e_i), \tag{4}$$

so that the fovea contains many narrow receptive fields while the periphery is represented by fewer, broader fields. Consequently, the same number of neural measurements covers a large field of view while preserving high acuity around fixation.

2.2 Center–surround and anisotropic responses

A Difference-of-Gaussians (DoG) unit represents local contrast rather than raw intensity:

$$\phi_i^{\mathrm{DoG}}(\boldsymbol{x}) = G(\boldsymbol{x}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i^c) - \lambda G(\boldsymbol{x}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i^s), \qquad \boldsymbol{\Sigma}_i^s \succ \boldsymbol{\Sigma}_i^c. \tag{5}$$

Peripheral fields may also be anisotropic. If $\boldsymbol{R}_i$ aligns the local axes with the radial and tangential directions relative to the fovea, then

$$\boldsymbol{\Sigma}_i = \boldsymbol{R}_i \begin{bmatrix} \sigma_{r,i}^2 & 0 \\ 0 & \sigma_{\theta,i}^2 \end{bmatrix} \boldsymbol{R}_i^\top. \tag{6}$$

The centers, covariances, channel weights, and center–surround ratios may be fixed from a biological prior or learned with regularization that preserves coverage, smooth scale progression, and limited overlap.

3 Binocular Depth from Retinal Correlation

The left and right retinas observe the scene from viewpoints separated by a baseline B. For rectified pinhole cameras, a point at depth Z has approximately the disparity

$$\delta(Z) = \frac{fB}{Z}, \tag{7}$$

where f is the focal length in compatible units. More general camera models replace this scalar shift by a calibrated warp $\mathcal{W}(\boldsymbol{x}; Z, \boldsymbol{q}_t)$, where $\boldsymbol{q}_t$ contains the current gaze and vergence state.

For candidate depth Z, the right retinal measurement aligned with left unit i is

$$\tilde{r}_i^R(Z, t) = \int_\Omega L^R(\boldsymbol{x}, t)\, \phi_i(\mathcal{W}(\boldsymbol{x}; Z, \boldsymbol{q}_t); \boldsymbol{c}_i, \boldsymbol{\Sigma}_i)\, d\boldsymbol{x}. \tag{8}$$

A simple binocular matching energy is

$$E_i(Z) = \left\| r_i^L - \tilde{r}_i^R(Z) \right\|_2^2, \qquad Z_i^* = \arg\min_Z E_i(Z). \tag{9}$$

In practice, r_i^e can be a learned feature vector rather than a scalar. A normalized correlation score is less sensitive to gain changes:

$$C_i(Z) = \frac{\left\langle \boldsymbol{r}_i^L, \tilde{\boldsymbol{r}}_i^R(Z) \right\rangle}{\|\boldsymbol{r}_i^L\|_2 \|\tilde{\boldsymbol{r}}_i^R(Z)\|_2 + \varepsilon}, \qquad Z_i^* = \arg\max_Z C_i(Z). \tag{10}$$

The collection $E_i(Z)$ forms a *foveated stereo cost volume*. Narrow foveal receptive fields support precise disparity estimates; broader peripheral fields provide coarse, robust depth estimates. A probabilistic depth estimate can be obtained with

$$p_i(Z \mid \boldsymbol{r}^L, \boldsymbol{r}^R) = \frac{\exp[-\beta E_i(Z)]}{\sum_{Z'} \exp[-\beta E_i(Z')]}, \tag{11}$$

whose entropy represents local depth uncertainty.

4 Active Vergence and Focus Sweep

For a chosen viewing direction, the binocular sensor varies vergence over a range of candidate depths. At the correct setting, image structures near the fixated object become registered between the eyes and binocular agreement increases. The sweep therefore implements an active depth measurement:

1. select a fixation direction from the peripheral representation;
2. acquire foveated measurements from both eyes;
3. sweep a discrete or continuous set of vergence/depth hypotheses;
4. align the right and left receptive fields using calibration;
5. compute the cost or posterior in equations (9) and (11);
6. select the most likely depth and estimate its confidence.

Vergence and optical focus need not be identical variables. Vergence supplies a geometric binocular cue, while accommodation/defocus supplies an additional monocular cue. Their joint likelihood may be written as

$$p(Z \mid \mathcal{D}_t) \propto p(\mathcal{D}_{\text{stereo}} \mid Z) p(\mathcal{D}_{\text{defocus}} \mid Z) p(Z \mid \mathcal{D}_{1:t-1}), \tag{12}$$

where temporal evidence supplies the prior. The motor state itself is a useful measurement and must be included among the network inputs.

5 Saccadic Object Inspection

The initial vergence stage yields a coarse object hypothesis and approximate depth. The system then moves the fovea over informative surface regions, including occluding contours, depth discontinuities, high-curvature areas, texture changes, uncertain locations, and candidate semantic parts.

At time t, the observation token set can be represented as

$$\mathcal{Z}_t = \left\{ \boldsymbol{r}_{i,t}^L, \boldsymbol{r}_{i,t}^R, \boldsymbol{c}_{i,t}, \boldsymbol{\Sigma}_{i,t}, e_{i,t}, \boldsymbol{f}_t, \boldsymbol{v}_t, a_t \right\}_{i=1}^N, \tag{13}$$

where $\boldsymbol{f}_t$ is fixation, $\boldsymbol{v}_t$ is vergence, and a_t is the focus or accommodation state. A transformer, graph neural network, or continuous neural field can encode these irregular, multiscale tokens.

The next action u_{t+1} may be selected to maximize expected information gain about an object model M:

$$u_{t+1}^{*} = \arg\max_{u} I(M; \mathcal{Z}_{t+1} \mid u, \mathcal{Z}_{1:t}) \,, \tag{14}$$

or, equivalently, to minimize expected posterior entropy plus movement and computation costs:

$$u_{t+1}^{*} = \arg\min_{u} \mathbb{E}[H(M \mid \mathcal{Z}_{1:t+1}, u)] + \lambda_s C_{\text{saccade}}(u) + \lambda_c C_{\text{compute}}(u). \tag{15}$$

Thus, a saccade is selected because it is expected to clarify geometry, segmentation, or identity, rather than merely because it points toward high image contrast.

6 Persistent 3D Scene and Object Memory

Each fixation updates a persistent representation

$$M_t = \mathcal{U}(M_{t-1}, \mathcal{Z}_t, \boldsymbol{q}_t), \tag{16}$$

where $\mathcal{U}$ is a learned or probabilistic fusion operator. The memory should retain at least

$$\begin{aligned} M_t = \{&\text{depth},\ \text{occupancy},\ \text{surface normals},\ \text{texture},\\ &\text{boundaries},\ \text{parts},\ \text{semantic belief},\ \text{uncertainty}\}. \end{aligned} \tag{17}$$

Possible realizations include a voxel or surfel map, an object-centric latent representation, a neural radiance field, or a Gaussian-splatting model. The essential requirement is that measurements from different gaze states can be registered and fused while their uncertainty is retained.

7 System Architecture

The complete perception–action loop consists of the following modules:

Table 1: Principal modules of the proposed system.

Module	**Function**
Binocular renderer/sensor	Produces calibrated left and right radiance fields, including focus and defocus effects.
Foveated retina	Converts continuous images into multiscale Gaussian or DoG tokens.
Stereo matcher	Builds a retinal cost volume and estimates a depth posterior.
Object proposer	Uses peripheral and coarse-depth evidence to select an object of interest.
Active controller	Chooses fixation, vergence, focus, and retinal scale.
Scene memory	Fuses temporal evidence into a persistent geometric and semantic model.
Prediction heads	Produce depth, instance/part segmentation, surface properties, and semantic identity.

The high-level processing sequence is

$$\begin{aligned}\text{scene} &\longrightarrow \text{binocular radiance} \longrightarrow \text{foveated retinal tokens} \\ &\longrightarrow \text{coarse stereo and object hypothesis} \\ &\longrightarrow \text{saccadic inspection} \longrightarrow \text{semantic 3D object model.}\end{aligned} \tag{18}$$

8 Learning Objectives

The perception modules can be optimized jointly using a weighted objective

$$\mathcal{L} = \lambda_d \mathcal{L}_{\text{depth}} + \lambda_s \mathcal{L}_{\text{seg}} + \lambda_m \mathcal{L}_{\text{sem}} + \lambda_r \mathcal{L}_{\text{recon}} + \lambda_t \mathcal{L}_{\text{temporal}} + \lambda_u \mathcal{L}_{\text{uncertainty}} + \lambda_a \mathcal{L}_{\text{action}}. \tag{19}$$

Here, $\mathcal{L}_{\text{depth}}$ supervises metric depth or disparity; $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{sem}}$ supervise object masks and labels; $\mathcal{L}_{\text{recon}}$ encourages appearance or feature consistency; and $\mathcal{L}_{\text{temporal}}$ enforces agreement after compensating for gaze and scene motion. The uncertainty term calibrates the posterior in equation (11). The action term may be trained by reinforcement learning, imitation of a privileged oracle, or differentiable expected information gain.

Useful regularizers for a learned retina include coverage of the visual field, limited redundancy among neighboring receptive fields, monotonic scale growth with eccentricity, and temporal smoothness except during saccades.

9 Experimental Program

9.1 Phase I: differentiable binocular retina

Implement a simulated stereo camera in Blender, Unity, or Isaac Sim. Render ground-truth depth, instance masks, semantic labels, surface normals, and optical defocus. Compare regular raster sampling, fixed log-polar sampling, and learned Gaussian receptive fields.

9.2 Phase II: vergence-based depth

Construct retinal cost volumes over discrete depth hypotheses. Evaluate depth accuracy as a function of eccentricity, receptive-field size, texture, depth discontinuity, baseline, and vergence-step size. Compare scalar intensity, DoG, and learned feature correlations.

9.3 Phase III: active saccadic inspection

Given an object proposal, train policies that choose successive fixation points. Compare random, saliency-based, boundary-based, uncertainty-based, and learned information-gain policies under a fixed budget of retinal measurements.

9.4 Phase IV: semantic 3D integration

Fuse observations into an object-centered 3D representation and predict the complete mask, geometry, and class. Evaluate the number of fixations and total measurements required to reach a target accuracy, in addition to conventional depth, segmentation, and recognition metrics.

10 Research Questions and Expected Contributions

The proposal raises several concrete research questions:

- Can Gaussian retinal measurements recover depth as accurately as dense stereo while using substantially fewer samples?

- Which receptive-field arrangement best balances peripheral detection and foveal metric accuracy?
- Does physical or simulated vergence improve matching relative to a purely computational disparity sweep?
- How should stereo, defocus, motor state, and temporal parallax be fused?
- Which saccade policy best reduces joint uncertainty in geometry, segmentation, and semantics?
- Can a learned retinal layout outperform a biologically prescribed one without losing stability or interpretability?

The anticipated contribution is not merely a foveated image encoder or a new stereo matcher. It is a closed-loop visual system in which sensing resolution, fixation, vergence, focus, and semantic inference are optimized together. In this view, segmentation and recognition emerge through active accumulation of binocular evidence.

11 Conclusion

Gaussian receptive fields provide a mathematically clean bridge between a continuous radiance field and a variable-resolution neural representation. Applied binocularly, they support depth inference through correlation under candidate geometric warps. Active vergence converts the depth search into a controlled sensing operation, while saccades progressively refine the geometry, appearance, segmentation, and identity of an object. A persistent 3D memory integrates these sparse but strategically selected measurements. This framework defines a concrete route toward efficient, interpretable, and biologically inspired active 3D perception.

References

[1] D. H. Ballard, "Animate vision," *Artificial Intelligence*, vol. 48, no. 1, pp. 57–86, 1991. https://doi.org/10.1016/0004-3702(91)90080-4

[2] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 996–1005, 1988. https://doi.org/10.1109/5.5968

[3] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, pp. 194–203, 2001. https://doi.org/10.1038/35058500

[4] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, 1995.