A Framework for Stereoscopic 3D Vision: Active Bayesian Inference with Nested Oculomotor Control

Working draft — prepared by GPT for L. Velho, VISGRAF / IMPA

August 8, 2026

Abstract

This document develops a coherent, biologically grounded and computationally explicit framework for stereoscopic three-dimensional vision. The organizing principle is that 3D perception is *active Bayesian inference* of a latent scene, in which eye movements are actions that shape the observations. The framework is organized as a stack of six layers — generative geometry, neural encoding, disparity-field inference, metric scaling and cue fusion, oculomotor control, and an active-sampling outer loop — coupled by three "closures" that force perception and motor control to share variables. Each layer is stated with its governing equations and, where instructive, full derivations. The intent is a scaffold to iterate on: every section is modular and can be deepened, replaced, or instantiated independently. Section 9 sketches two runnable instantiations (a cooperative/belief-propagation network and an amortized variational network), and Section 10 lists open problems.

Contents

1 Organizing principle and notation **2**
1.1 The generative–inferential–active loop 2
1.2 Notation 2

2 Layer 0 — Generative geometry (the forward model) **3**
2.1 Two eyes as calibrated cameras; the epipolar constraint 3
2.2 Vergence and absolute distance 3
2.3 Disparity and relative depth: the $1/D^2$ scaling 3
2.4 The horopter: the Vieth–Müller circle 4
2.5 Vertical disparities: a second distance estimator 4
2.6 Torsion, Listing's law, and the moving epipolar geometry 4

3 Layer 1 — Neural encoding: the disparity likelihood **5**
3.1 The binocular energy model 5
3.2 Why (9) is tuned to disparity 5
3.3 From a population to a likelihood 5

4 Layer 2 — Disparity-field inference: correspondence **6**
4.1 Ill-posedness and regularization 6
4.2 MRF / energy-minimization formulation 6

4.3 Min-sum belief propagation . . . 6
4.4 The Marr–Poggio cooperative network (biological form) . . . 6
4.5 Occlusion and half-matches . . . 7

5 Layer 3 — Metric scaling and cue integration 7
5.1 Scaling the disparity field . . . 7
5.2 Reliability-weighted (MLE) cue combination . . . 7
5.3 The vergence–accommodation conflict, quantified . . . 8

6 Layer 4 — Oculomotor control (the fast inner loop) 8
6.1 The oculomotor plant and the saccadic pulse–step . . . 8
6.2 Disparity-driven vergence as delayed feedback . . . 8
6.3 Accommodation–vergence cross-coupling . . . 9
6.4 Combining saccades and vergence . . . 9

7 Layer 5 — The active-sampling outer loop 9
7.1 Foveated, posture-conditioned observation model . . . 9
7.2 Recursive Bayesian estimation across fixations . . . 9
7.3 Trans-saccadic remapping via efference copy . . . 10
7.4 Gaze as active inference / information maximization . . . 10

8 Integration: three closures and a unified state space 10

9 Computational instantiation 11

10 Open problems and directions to iterate on 11

1 Organizing principle and notation

1.1 The generative–inferential–active loop

We model the visual system as maintaining a posterior over a latent scene $\mathbf{z}$ (a depth/structure field in a world- or head-centred frame) given a history of retinal observations $\mathbf{o}_{1:t}$ and eye postures $\boldsymbol{\phi}_{1:t}$:

$$p(\mathbf{z} \mid \mathbf{o}_{1:t}, \boldsymbol{\phi}_{1:t}) \;\propto\; p(\mathbf{o}_t \mid \mathbf{z}, \boldsymbol{\phi}_t) \underbrace{\int p(\mathbf{z} \mid \mathbf{z}')\, p(\mathbf{z}' \mid \mathbf{o}_{1:t-1}, \boldsymbol{\phi}_{1:t-1})\, d\mathbf{z}'}_{\text{prediction}}. \tag{1}$$

The *likelihood* $p(\mathbf{o}_t \mid \mathbf{z}, \boldsymbol{\phi}_t)$ is fixed by the imaging geometry (Layer 0) and read out by disparity-selective neurons (Layer 1); its inversion into a disparity field is Layer 2; metric interpretation and cue fusion are Layer 3; the eye posture $\boldsymbol{\phi}_t$ and its dynamics are set by oculomotor control (Layer 4); and the choice of *where and at what depth to sample next* is the active outer loop (Layer 5). Everything is one recursive estimator with a control policy.

1.2 Notation

- I — interocular (baseline) distance, $\approx 6.3\,\text{cm}$.
- D — fixation distance along the visual axis; r — distance to an arbitrary point.
- μ — symmetric (horizontal) vergence angle; $\gamma = \mu/2$ half-vergence.
- η — horizontal retinal disparity (radians), relative to the current fixation.

- ΔZ — depth of a point relative to the fixation plane.
- $\mathbf{z}$ — latent scene/depth field; $d(\cdot)$ — retinal disparity field.
- $\boldsymbol{\phi}$ — binocular eye posture (version, vergence, torsion).
- A — accommodative state (dioptres); V — vergence state (metre angles / prism dioptres).

Small-angle approximations ($\tan\theta \approx \theta$, $\sin\theta \approx \theta$) are used wherever a first-order geometric cue is derived; they are accurate to a few percent within Panum's fusional range and the central $\pm 10^\circ$ of the field.

2 Layer 0 — Generative geometry (the forward model)

2.1 Two eyes as calibrated cameras; the epipolar constraint

Treat each eye as a central-projection camera. Let the two eyes be related by rotation $\mathbf{R}$ and translation $\mathbf{t}$ (the baseline). For normalized image rays $\mathbf{x}_L, \mathbf{x}_R$ of the same world point,

$$\mathbf{x}_R^\top \mathbf{E}\, \mathbf{x}_L = 0, \qquad \mathbf{E} = [\mathbf{t}]_\times \mathbf{R}, \tag{2}$$

where $[\mathbf{t}]_\times$ is the skew-symmetric matrix of $\mathbf{t}$. Equation (2) confines the match of a left-image feature to a one-dimensional *epipolar line* in the right image. For the human eyes in primary position the baseline is (nearly) horizontal, so epipolar lines are (nearly) horizontal and horizontal disparity carries the primary depth signal. Crucially, $\mathbf{E}$ depends on eye posture $\boldsymbol{\phi}$ through $\mathbf{R}$: convergence and torsion (Section 2.6) rotate the epipolar geometry, so (2) is a function of the motor state, not a constant.

2.2 Vergence and absolute distance

Place the eye nodal points at $(\pm I/2, 0)$ and a symmetric fixation point at $(0, D)$. Each visual axis makes a half-vergence angle γ with the median plane,

$$\tan\gamma = \frac{I/2}{D}, \qquad \mu = 2\gamma = 2\arctan\frac{I}{2D} \xrightarrow{I \ll D} \frac{I}{D}. \tag{3}$$

Hence *vergence is an (extraretinal) absolute-distance signal*, $D \approx I/\mu$, available from efference copy and extraocular proprioception independently of image content.

2.3 Disparity and relative depth: the $1/D^2$ scaling

Define the angle subtended by the baseline at a point at distance r as $\theta(r) \approx I/r$. The retinal disparity between the fixation point (distance D) and a target at distance $r = D + \Delta Z$ is the difference of these convergence angles:

$$\eta = \theta(D) - \theta(D + \Delta Z) = I\left(\frac{1}{D} - \frac{1}{D + \Delta Z}\right) = \frac{I\,\Delta Z}{D\,(D + \Delta Z)} \xrightarrow{\Delta Z \ll D} \boxed{\eta \approx \frac{I\,\Delta Z}{D^2}}. \tag{4}$$

Result 1 (Disparity is metrically ambiguous without distance)**.** *Inverting* (4) *gives* $\Delta Z \approx (D^2/I)\,\eta$. *Substituting* $D \approx I/\mu$ *from* (3),

$$\Delta Z \;\approx\; \frac{D^2}{I}\,\eta \;=\; \frac{I}{\mu^2}\,\eta. \tag{5}$$

The same retinal disparity maps to depths that differ by orders of magnitude across viewing distances; a distance estimate (from vergence and/or vertical disparity) is required *to scale it.*

Result 2 (Quadratic error amplification)**.** *Differentiating* (5) *at fixed* η *and using* $D = I/\mu$*,*

$$\frac{\partial\,\Delta Z}{\partial D} = \frac{2\,\eta\,D}{I} \quad\Longrightarrow\quad \frac{\delta(\Delta Z)}{\Delta Z} = 2\,\frac{\delta D}{D} = -2\,\frac{\delta\mu}{\mu}.$$

A fractional error in the vergence-derived distance produces twice *that fractional error in recovered depth. This is the geometric root of systematic stereo depth biases and a principal driver of discomfort when display geometry and vergence disagree (Section 5.3).*

2.4 The horopter: the Vieth–Müller circle

The zero-disparity locus consists of points imaged on corresponding retinal locations, i.e. points that subtend the *same* vergence angle as the fixation point. By the inscribed-angle theorem, the set of points from which a fixed segment (here the interocular baseline) subtends a constant angle is an arc of a circle through the segment's endpoints. Therefore the geometric horopter is a circle passing through the two eyes' nodal points and the fixation point — the *Vieth–Müller circle*. Points inside it carry crossed (near) disparity, points outside carry uncrossed (far) disparity. The empirical horopter deviates from the circle (the Hering–Hillebrand deviation), and single (fused) vision is possible only within *Panum's fusional area*, a disparity band of width $\sim 0.1^\circ$ foveally, widening in the periphery; beyond it the percept splits into diplopia, itself a coarse depth signal.

2.5 Vertical disparities: a second distance estimator

Horizontal disparity gives *relative* depth; vertical disparities furnish *absolute* distance and gaze angle without extraretinal signals (Mayhew & Longuet-Higgins). Consider a point at horizontal-plane distance D and azimuth α (so $X = D\sin\alpha$), eyes at $(\pm I/2, 0, 0)$. The distance from eye e to the point is $R_e = \sqrt{(X \mp I/2)^2 + Z^2}$; to first order in I/D,

$$R_R \approx D - \tfrac{I}{2}\sin\alpha, \qquad R_L \approx D + \tfrac{I}{2}\sin\alpha.$$

A small vertical extent subtends a visual angle $\propto 1/R_e$, so the vertical-size ratio (right/left) is

$$\mathrm{VSR} = \frac{R_L}{R_R} \approx \frac{D + \frac{I}{2}\sin\alpha}{D - \frac{I}{2}\sin\alpha} \approx 1 + \frac{I\,\sin\alpha}{D}. \tag{6}$$

Result 3 (Distance from the vertical-disparity field)**.** *The azimuthal gradient of the VSR recovers distance directly:*

$$\frac{\partial\,\mathrm{VSR}}{\partial(\sin\alpha)} \approx \frac{I}{D} \quad\Longrightarrow\quad D \approx I\left(\frac{\partial\,\mathrm{VSR}}{\partial(\sin\alpha)}\right)^{-1}.$$

This provides a purely retinal estimate of D *that can be fused with the vergence estimate in Layer 3, and it is what makes disparity-defined slant and curvature perception veridical over a range of distances.*

2.6 Torsion, Listing's law, and the moving epipolar geometry

Eye orientation is not a free 3-DoF variable: during fixations, torsion is tied to gaze direction by *Listing's law* (orientation vectors lie in Listing's plane). Its binocular extension (L2) rotates each eye's Listing plane outward in proportion to vergence, producing systematic cyclovergence. Because $\mathbf{R}$ in (2)–(2) carries this torsion, the epipolar lines tilt with gaze and vergence, and the empirical pattern of vertical disparities in (6) acquires a vergence-dependent correction. A complete forward model therefore parametrizes $\mathbf{E} = \mathbf{E}(\boldsymbol{\phi})$ with $\boldsymbol{\phi}$ constrained to the Listing/L2 manifold.

3 Layer 1 — Neural encoding: the disparity likelihood

3.1 The binocular energy model

The canonical front end (Ohzawa, DeAngelis & Freeman; Qian; Fleet, Wagner & Heeger) models a V1 binocular *simple* cell as a linear combination of Gabor-filtered left and right images. Write each monocular receptive field (1D, along the epipolar direction) as

$$g_e(x) = \exp\Big(-\frac{x^2}{2\sigma^2}\Big)\cos(\omega x + \phi_e), \qquad e \in \{L, R\}. \tag{7}$$

Disparity preference is encoded either by a *position shift* $\Delta x = x_L - x_R$ or by a *phase shift* $\Delta\phi = \phi_L - \phi_R$ between the two RFs. The simple-cell linear response to a stereo pair (I_L, I_R) is

$$L = \int g_L(x)\, I_L(x)\, dx + \int g_R(x)\, I_R(x)\, dx. \tag{8}$$

A binocular *complex* cell sums the squared outputs of a quadrature pair (L_1, L_2) (RFs identical but in 90° spatial-phase quadrature):

$$C = L_1^2 + L_2^2. \tag{9}$$

3.2 Why (9) is tuned to disparity

Consider a stimulus with pure horizontal disparity δ: $I_R(x) = I_L(x - \delta)$. The monocular Gabor outputs behave like local Fourier components at frequency ω; the binocular sum in a phase-shift cell produces, after squaring and averaging over stimulus phase, a mean response of the form

$$\langle C \rangle \;=\; C_0\Big[\,1 + m\,\cos\big(\omega\,\delta - \Delta\phi\big)\Big], \tag{10}$$

with modulation depth m set by the monocular energies. Thus the cell is maximally excited when the stimulus disparity matches the built-in preferred disparity

$$d^* = \frac{\Delta\phi}{\omega} \quad \text{(phase model)}, \qquad d^* = \Delta x \quad \text{(position model)}. \tag{11}$$

Equivalently, (9) computes a *local cross-correlation* of the two images through a band-pass window: the complex cell is a soft, phase-invariant match detector. The phase model predicts the empirically observed attenuation of tuning at large disparities and the coupling between preferred disparity and spatial frequency; real neurons appear to use a hybrid of position and phase shifts.

3.3 From a population to a likelihood

A bank of energy units at a retinal location, tuned to disparities $\{d_i\}$ across spatial scales $\{\sigma_k = c/\omega_k\}$, yields a population response $\{C_i\}$ that (under a Poisson-like noise model) is read as an unnormalized log-likelihood over local disparity,

$$\log p\big(\mathbf{o}(\mathbf{x}) \mid d(\mathbf{x})\big) \;\approx\; \sum_i \big[\, C_i \log \lambda_i(d) - \lambda_i(d) \,\big] + \text{const}, \tag{12}$$

where $\lambda_i(d)$ is the tuning curve (10) of unit i. Coarse scales give unambiguous but low-resolution estimates; fine scales are precise but suffer $2\pi/\omega$ phase-wrap aliasing. *Coarse-to-fine* propagation (initialize fine-scale search from coarse-scale disparity) resolves the aliasing and is the interface to Layer 2. This is exactly the multiresolution strategy used in computational stereo.

4 Layer 2 — Disparity-field inference: correspondence

4.1 Ill-posedness and regularization

Recovering a dense disparity field $d(\cdot)$ from local match scores is ill-posed: local evidence is ambiguous (the aperture and correspondence problems). Following Marr's computational theory, one restores well-posedness with priors expressing (i) *uniqueness* — at most one match per line of sight — and (ii) *continuity* — disparity varies smoothly except at object boundaries.

4.2 MRF / energy-minimization formulation

Cast inference as MAP estimation over a Markov random field on the image lattice:

$$\hat{d} = \arg\min_{d} \underbrace{\sum_{\mathbf{x}} \rho\big(I_L(\mathbf{x}) - I_R(\mathbf{x} - d(\mathbf{x}))\big)}_{\text{data / matching cost}} + \lambda \sum_{\langle \mathbf{x}, \mathbf{x}' \rangle} V\big(d(\mathbf{x}), d(\mathbf{x}')\big), \tag{13}$$

where ρ is a robust photometric cost (or the negative log-likelihood from Layer 1), V a robust pairwise potential encoding piecewise smoothness (a truncated quadratic or Potts term implements depth discontinuities), and λ trades data against prior. The posterior is $p(d \mid I_L, I_R) \propto \exp(-E(d))$; (13) is its MAP estimate, obtained by graph cuts or belief propagation.

4.3 Min-sum belief propagation

On the MRF (13), min-sum (max-product) messages from node $\mathbf{x}$ to neighbour $\mathbf{x}'$ iterate

$$m^{(t+1)}_{\mathbf{x}\to\mathbf{x}'}(d') = \min_{d} \Big[\rho_{\mathbf{x}}(d) + \lambda V(d, d') + \sum_{\mathbf{x}'' \in \mathcal{N}(\mathbf{x}) \setminus \mathbf{x}'} m^{(t)}_{\mathbf{x}''\to\mathbf{x}}(d) \Big], \tag{14}$$

and the belief (approximate marginal) at $\mathbf{x}$ is $b_{\mathbf{x}}(d) = \rho_{\mathbf{x}}(d) + \sum_{\mathbf{x}'' \in \mathcal{N}(\mathbf{x})} m_{\mathbf{x}''\to\mathbf{x}}(d)$. This is the modern, convergent gloss on the recurrent cortical computation posited below.

4.4 The Marr–Poggio cooperative network (biological form)

The same priors are realizable as a recurrent network of disparity units $C(\mathbf{x}, d) \in [0, 1]$ with local excitation along constant-disparity neighbourhoods (continuity) and inhibition along each line of sight (uniqueness):

$$C^{(t+1)}(\mathbf{x}, d) = \sigma\Big(\underbrace{\sum_{(\mathbf{x}', d') \in S_{\text{exc}}(\mathbf{x}, d)} C^{(t)}(\mathbf{x}', d')}_{\text{continuity}} - \varepsilon \underbrace{\sum_{(\mathbf{x}', d') \in S_{\text{inh}}(\mathbf{x}, d)} C^{(t)}(\mathbf{x}', d')}_{\text{uniqueness}} + C^{(0)}(\mathbf{x}, d) \Big), \tag{15}$$

with σ a threshold nonlinearity and $C^{(0)}$ the initial local match support from Layer 1. Under symmetric non-negative weights the dynamics (15) descend a Lyapunov (energy) function and converge to a fixed point that satisfies the two constraints; it can be read as coordinate relaxation on (13). Julesz's random-dot stereograms establish that this matching runs on raw image structure, prior to and independent of monocular object recognition (*global stereopsis*).

4.5 Occlusion and half-matches

A complete model augments the state with a visibility/occlusion label $v(\mathbf{x}) \in \{\text{binocular}, \text{monocular}\}$; uniqueness then permits unmatched (half-occluded) pixels, and the ordering constraint (matches preserve left-to-right order for opaque surfaces) prunes the search. The da Vinci stereopsis phenomena — depth assigned to monocularly visible regions near occluders — are diagnostic that the biological solver represents occlusion explicitly.

5 Layer 3 — Metric scaling and cue integration

5.1 Scaling the disparity field

Layer 2 returns a *relative* field $d(\cdot)$. Result 1 converts it to metric depth using a distance estimate $\hat{D}$ formed from vergence (3) and vertical disparity (6):

$$\widehat{\Delta Z}(\mathbf{x}) = \frac{\hat{D}^2}{I}\, d(\mathbf{x}). \tag{16}$$

Because Result 2 shows depth error scales as $2\,\delta\hat{D}/\hat{D}$, the reliability of $\hat{D}$ dominates the reliability of metric depth — motivating the fusion below.

5.2 Reliability-weighted (MLE) cue combination

Model each depth cue i (disparity-from-vergence, disparity-from-vertical-disparity, motion parallax, texture, accommodation, familiar size) as an unbiased Gaussian estimate $\hat{Z}_i$ with variance σ_i^2. Under conditional independence the maximum-likelihood (and, with a flat prior, Bayes) combination is the precision-weighted mean:

$$\hat{Z} = \sum_i w_i\, \hat{Z}_i, \qquad w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2}, \qquad \sigma_{\hat{Z}}^2 = \Big(\textstyle\sum_i 1/\sigma_i^2\Big)^{-1}. \tag{17}$$

Two empirical signatures follow and are observed: the fused estimate is *more reliable than any single cue* ($\sigma_{\hat{Z}}^2 \le \min_i \sigma_i^2$), and the weights *reweight with viewing geometry* because the σ_i are distance-dependent (disparity precise near, vergence/vertical-disparity dominant far). *Modified weak fusion* (Landy et al.) extends (17) by allowing cues to disambiguate one another ("promotion") before linear combination, and robust variants down-weight cues in large conflict.

Cue	Primary regime	Distance dependence of reliability
Binocular disparity	near / mid	degrades as $1/D^2$ (Eq. 4)
Vergence	near	degrades with D; poor beyond $\sim 2\,\text{m}$
Vertical disparity	mid, large field	needs eccentricity; weak foveally
Motion parallax	near / mid	needs self- or object motion
Accommodation	very near	weak, saturates by $\sim 2\,\text{m}$
Texture / perspective	far	improves with structured surfaces

Table 1: Qualitative reliability regimes entering the weights w_i of Eq. (17).

5.3 The vergence–accommodation conflict, quantified

In natural viewing, the accommodative demand (dioptres, $= 1/D$) and the vergence demand (metre angles, $\approx I/D$ up to units) are geometrically coupled: both are set by the same D. A stereoscopic display fixes the focal distance at the screen D_s while disparity drives vergence to a simulated distance D_v. The conflict is the mismatch

$$\Delta_{\mathrm{VAC}} = \frac{1}{D_v} - \frac{1}{D_s} \quad [\text{dioptres}], \tag{18}$$

which enters the fusion (17) as a persistent inter-cue disagreement and, through the cross-coupled control loop of Section 6.3, forces the oculomotor system off its natural operating line — the mechanistic account of stereo-display fatigue and the target that varifocal / light-field displays aim to null.

6 Layer 4 — Oculomotor control (the fast inner loop)

6.1 The oculomotor plant and the saccadic pulse–step

The eyeball-in-orbit behaves as an overdamped second-order plant (Robinson). With eye position θ and motoneuron drive u,

$$P(s) = \frac{\Theta(s)}{U(s)} = \frac{K}{(1+sT_1)(1+sT_2)}, \qquad T_1 \gg T_2, \tag{19}$$

with representative time constants $T_1 \sim 150$–$250\,\mathrm{ms}$ and $T_2 \sim 10$–$20\,\mathrm{ms}$. To move the eye quickly and hold it, the controller issues a *pulse–step*: a phasic burst u_p (a velocity command that overcomes plant viscosity) followed by a tonic step u_s (a position command that holds against elasticity). A neural integrator generates the step from the pulse, $u_s(t) = \frac{1}{T_N}\int u_p\,dt$; pulse–step mismatch produces post-saccadic drift. Saccades are *ballistic / open loop*: the command is pre-computed and not corrected mid-flight, and their kinematics obey the saturating *main sequence*

$$\dot{\theta}_{\max}(A) = V_{\max}\big(1 - e^{-A/A_0}\big), \tag{20}$$

relating peak velocity to saccade amplitude A.

6.2 Disparity-driven vergence as delayed feedback

Vergence is, by contrast, *closed loop*: it drives the foveal disparity error to zero. A first-order model with loop delay Δ ($\sim$160 ms) is

$$\tau\,\dot{V}(t) = -V(t) + G\,\big(V^* - V(t-\Delta)\big), \tag{21}$$

where V^* is the demanded vergence and $\eta = V^* - V$ the disparity error. Empirically vergence is *dual-mode*: a fast preprogrammed transient initiates the movement, and a slow disparity-driven (fusional) component nulls residual error.

Result 4 (Delay-stability of the vergence loop)**.** *The characteristic equation of* (21) *is* $\tau s + 1 + Ge^{-s\Delta} = 0$. *Marginal stability* ($s = j\omega_c$) *requires simultaneously*

$$1 + G\cos(\omega_c\Delta) = 0, \qquad \tau\omega_c = G\sin(\omega_c\Delta).$$

These fix the critical gain G_c and oscillation frequency ω_c; loop gains above G_c (or delays above the corresponding Δ_c) produce sustained oscillation. This bounds how aggressively vergence can be tuned and predicts the instability seen with artificially increased loop delay.

6.3 Accommodation–vergence cross-coupling

Accommodation (blur-driven, output A) and vergence (disparity-driven, output V) are two negative-feedback loops joined by cross-links (Hung–Semmlow; Schor). At steady state, with blur stimulus A_s, disparity stimulus V_s, forward gains a, v and cross-gains AC/A (accommodative convergence) and CA/C (convergence accommodation),

$$A = a\,(A_s - A) + \tfrac{\mathrm{CA}}{\mathrm{C}}\,V, \qquad V = v\,(V_s - V) + \tfrac{\mathrm{AC}}{\mathrm{A}}\,A, \tag{22}$$

i.e. the coupled linear system

$$\begin{pmatrix} 1+a & -\mathrm{CA/C} \\ -\mathrm{AC/A} & 1+v \end{pmatrix} \begin{pmatrix} A \\ V \end{pmatrix} = \begin{pmatrix} a\,A_s \\ v\,V_s \end{pmatrix}. \tag{23}$$

The determinant $(1+a)(1+v) - \frac{\mathrm{AC}}{\mathrm{A}}\frac{\mathrm{CA}}{\mathrm{C}}$ governs solvability and, in the dynamic version, stability; a display-induced conflict Δ_{VAC} displaces the (A_s, V_s) operating point off the natural $A_s{=}V_s$ line, and (23) shows how the cross-links then drag accommodation and vergence into mutual error — the quantitative substrate of Section 5.3.

6.4 Combining saccades and vergence

For a target differing in both direction and depth, the classical Hering decomposition writes the two eye positions as a common version plus an opposed vergence,

$$\theta_L = \theta_{\mathrm{ver}} - \tfrac{1}{2}\theta_{\mathrm{vrg}}, \qquad \theta_R = \theta_{\mathrm{ver}} + \tfrac{1}{2}\theta_{\mathrm{vrg}}, \tag{24}$$

with a nonlinear interaction: vergence executed *during* a saccade is markedly accelerated (saccade–vergence facilitation), and the two eyes may make unequal (disconjugate) saccades. An alternative reading (Zhou & King) posits near-independent monocular premotor commands; for control design both reduce to a 2-input plant whose cross-terms must be identified from data.

7 Layer 5 — The active-sampling outer loop

7.1 Foveated, posture-conditioned observation model

Each fixation yields an observation that is high-precision at the fovea and rapidly degraded in the periphery. Writing the scene state as a depth-map vector $\mathbf{z}_t$ in a head/world frame, the observation at fixation $\boldsymbol{\phi}_t$ is

$$\mathbf{o}_t = \mathbf{H}(\boldsymbol{\phi}_t)\,\mathbf{z}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}\big(\mathbf{0}, \mathbf{R}(\boldsymbol{\phi}_t)\big), \tag{25}$$

where $\mathbf{H}(\boldsymbol{\phi}_t)$ selects and warps the currently foveated region (via the geometry of Section 2) and the noise covariance $\mathbf{R}(\boldsymbol{\phi}_t)$ encodes the acuity falloff (small at the fovea, large in the periphery).

7.2 Recursive Bayesian estimation across fixations

With a slow scene dynamics prior $\mathbf{z}_t = \mathbf{F}\mathbf{z}_{t-1} + \mathbf{w}_t$, $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, the master recursion (1) becomes a Kalman filter (linear-Gaussian case; a particle or variational filter otherwise):

$$\text{Predict:} \quad \boldsymbol{\mu}_{t|t-1} = \mathbf{F}\boldsymbol{\mu}_{t-1}, \qquad \boldsymbol{\Sigma}_{t|t-1} = \mathbf{F}\boldsymbol{\Sigma}_{t-1}\mathbf{F}^{\top} + \mathbf{Q}, \tag{26}$$

$$\text{Update:} \quad \mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\top}\big(\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\top} + \mathbf{R}_t\big)^{-1}, \tag{27}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\big(\mathbf{o}_t - \mathbf{H}_t\boldsymbol{\mu}_{t|t-1}\big), \qquad \boldsymbol{\Sigma}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\,\boldsymbol{\Sigma}_{t|t-1}. \tag{28}$$

Depth perception is thus *integrated over many fixations*, not read from a single glance, with the peripheral prior filled in and progressively refined where the fovea lands.

7.3 Trans-saccadic remapping via efference copy

Between fixations the retinal frame changes. An efference copy of the impending eye movement $\Delta\boldsymbol{\phi}_t$ supplies the coordinate transform $\mathbf{T}(\Delta\boldsymbol{\phi}_t)$ that re-registers the running estimate before the next update,

$$\boldsymbol{\mu}_{t|t-1} \leftarrow \mathbf{T}(\Delta\boldsymbol{\phi}_t)\,\boldsymbol{\mu}_{t|t-1}, \qquad \boldsymbol{\Sigma}_{t|t-1} \leftarrow \mathbf{T}\,\boldsymbol{\Sigma}_{t|t-1}\mathbf{T}^\top + \boldsymbol{\Sigma}_{\text{remap}}, \tag{29}$$

with $\boldsymbol{\Sigma}_{\text{remap}}$ absorbing the (imperfect) proprioceptive/efference-copy transform noise. This is the estimator-level counterpart of the perceptual stability we experience across saccades.

7.4 Gaze as active inference / information maximization

The policy closes the loop: choose the next fixation to make the observation most informative about $\mathbf{z}$. In information terms, maximize the expected reduction in posterior entropy (mutual information):

$$\boldsymbol{\phi}^*_{t+1} = \arg\max_{\boldsymbol{\phi}} \Big[\mathcal{H}\big(\mathbf{z}_t \mid \mathbf{o}_{1:t}\big) - \mathbb{E}_{\mathbf{o}_{\boldsymbol{\phi}}}\mathcal{H}\big(\mathbf{z}_t \mid \mathbf{o}_{1:t}, \mathbf{o}_{\boldsymbol{\phi}}\big)\Big] = \arg\max_{\boldsymbol{\phi}} I\big(\mathbf{z}_t; \mathbf{o}_{\boldsymbol{\phi}} \mid \mathbf{o}_{1:t}\big). \tag{30}$$

For the linear-Gaussian model this has the closed form

$$I\big(\mathbf{z}_t; \mathbf{o}_{\boldsymbol{\phi}} \mid \mathbf{o}_{1:t}\big) = \tfrac{1}{2}\log\det\Big(\mathbf{I} + \mathbf{R}(\boldsymbol{\phi})^{-1}\mathbf{H}(\boldsymbol{\phi})\,\boldsymbol{\Sigma}_t\,\mathbf{H}(\boldsymbol{\phi})^\top\Big), \tag{31}$$

so gaze is pulled toward regions of high current uncertainty $\boldsymbol{\Sigma}_t$ that the fovea can measure precisely (small $\mathbf{R}$). Equation (30) is the epistemic term of *expected free energy*; adding a task/goal (pragmatic) term $-\log p(\mathbf{o}_{\boldsymbol{\phi}} \mid \text{goal})$ recovers the full active-inference policy and links stereopsis to fixation-selection models of visual search (Najemnik & Geisler; Friston).

8 Integration: three closures and a unified state space

The framework's coherence rests on three couplings that most treatments leave separate.

Result 5 (Scaling closure)**.** *Disparity is metrically inert without a distance estimate (Result 1); that estimate comes from the vergence* action (3) *(and vertical disparity). Perception and motor control therefore* share *the variable D, and neither can be modelled correctly in isolation.*

Result 6 (Control-regime closure)**.** *The fast loop is closed-loop feedback on a disparity error* (21)*; the outer loop is a policy over a posterior* (30)*. Both are expressible in one estimation-and-control language (LQG / active inference), rather than as unrelated subsystems.*

Result 7 (Reference-frame closure)**.** *Eye posture $\boldsymbol{\phi}$ enters both the forward model* (2) *(tilting epipolar geometry) and the sequential estimator (trans-saccadic remapping). Efference copy of $\boldsymbol{\phi}$ is the connective tissue that keeps the two consistent.*

Collecting the pieces, define the augmented state $\mathbf{s} = (\mathbf{z},\ \boldsymbol{\phi},\ A,\ V)$ evolving as a *hybrid dynamical system*: continuous flow (plant + vergence/accommodation ODEs of Section 6, plus the KF prediction of Section 7) punctuated by discrete saccadic resets that trigger remapping and a fresh foveated update. The perception–action loop is then a single partially observed optimal-control problem whose *estimator* is Layers 1–3+5 and whose *controller* is Layer 4 driven by the policy (30).

9 Computational instantiation

Two runnable realizations, not rivals but different points on an exactness/speed trade-off:

1. **Cooperative / belief-propagation stereo.** Instantiate Layer 1 as a bank of Gabor energy units, Layer 2 as min-sum BP or the relaxation (15) on an MRF, Layer 3 as (17), and Layer 5 as a Kalman filter with the greedy info-max policy (31). Biologically faithful and interpretable; per-frame cost dominated by BP.
2. **Amortized variational stereo.** Replace exact inference in Layers 1–2 by a recognition network $q_\psi(d \mid I_L, I_R)$ trained to approximate the posterior $p(d \mid I_L, I_R)$ of (13) (the deep-stereo family). BP is then the target the network amortizes. Layers 3–5 are unchanged. Fast at run time; the energy model and MRF define the objective it is trained against.

In both, the *same* generative geometry of Layer 0 supplies the likelihood and the metric scaling, so the two instantiations are interchangeable back ends under one front-to-back specification.

10 Open problems and directions to iterate on

- *Vertical-disparity fusion weights.* Derive the distance-dependent reliabilities that set w_i in (17) for the vergence vs. vertical-disparity estimators from (3) and (6), and predict the crossover distance.
- *Nonlinear stability of the coupled A–V loop.* Extend (23) to its dynamic, delayed, nonlinear form and characterize the operating region violated by Δ_{VAC}.
- *Policy beyond greedy info-max.* Replace (30) with a finite-horizon (non-myopic) expected-free-energy policy and compare fixation sequences to human scanpaths on 3D scenes.
- *Listing/L2 in the forward model.* Fold the vergence-dependent torsion of Section 2.6 into $\mathbf{E}(\boldsymbol{\phi})$ and quantify its effect on the vertical-disparity distance cue.
- *XR objective.* Formalize a display-design loss that minimizes Δ_{VAC} and the fusion conflict in (17) subject to hardware constraints (varifocal / light-field).

References

[1] D. Marr and T. Poggio (1976). Cooperative computation of stereo disparity. *Science* 194, 283–287.

[2] D. Marr and T. Poggio (1979). A computational theory of human stereo vision. *Proc. R. Soc. Lond. B* 204, 301–328.

[3] D. Marr (1982). *Vision*. Freeman, San Francisco.

[4] B. Julesz (1971). *Foundations of Cyclopean Perception*. Univ. Chicago Press.

[5] G. F. Poggio and B. Fischer (1977). Binocular interaction and depth sensitivity in striate and prestriate cortex of the behaving rhesus monkey. *J. Neurophysiol.* 40, 1392–1405.

[6] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman (1990). Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science* 249, 1037–1041.

[7] N. Qian (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation* 6, 390–404.

[8] D. J. Fleet, H. Wagner, and D. J. Heeger (1996). Neural encoding of binocular disparity: energy models, position shifts and phase shifts. *Vision Research* 36, 1839–1857.

[9] J. C. A. Read (2005). Early computational processing in binocular vision and depth perception. *Prog. Biophys. Mol. Biol.* 87, 77–108.

[10] J. E. W. Mayhew and H. C. Longuet-Higgins (1982). A computational model of binocular depth perception. *Nature* 297, 376–378.

[11] B. T. Backus, M. S. Banks, R. van Ee, and J. A. Crowell (1999). Horizontal and vertical disparity, eye position, and stereoscopic slant perception. *Vision Research* 39, 1143–1170.

[12] B. J. Rogers and M. F. Bradshaw (1993). Vertical disparities, differential perspective and binocular stereopsis. *Nature* 361, 253–255.

[13] M. O. Ernst and M. S. Banks (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433.

[14] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research* 35, 389–412.

[15] D. A. Robinson (1964). The mechanics of human saccadic eye movements. *J. Physiol.* 174, 245–264.

[16] C. Rashbass and G. Westheimer (1961). Disjunctive eye movements. *J. Physiol.* 159, 339–360.

[17] G. K. Hung and J. L. Semmlow (1980). Static behavior of accommodation and vergence: computer simulation of an interactive dual-feedback system. *IEEE Trans. Biomed. Eng.* 27, 439–447.

[18] C. M. Schor (1992). A dynamic model of cross-coupling between accommodation and convergence. *Optom. Vis. Sci.* 69, 258–269.

[19] J. A. M. van Gisbergen, D. A. Robinson, and S. Gielen (1981). A quantitative analysis of generation of saccadic eye movements by burst neurons. *J. Neurophysiol.* 45, 417–442.

[20] W. Zhou and W. M. King (1998). Premotor commands encode monocular eye movements. *Nature* 393, 692–695.

[21] D. Tweed and T. Vilis (1990). Geometric relations of eye position and velocity vectors during saccades. *Vision Research* 30, 111–127.

[22] L. J. van Rijn and A. V. van den Berg (1993). Binocular eye orientation during fixations: Listing's law extended to include eye vergence. *Vision Research* 33, 691–708.

[23] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks (2008). Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vision* 8(3):33.

[24] J. Najemnik and W. S. Geisler (2005). Optimal eye movement strategies in visual search. *Nature* 434, 387–391.

[25] K. Friston, F. Rigoli, D. Ognibene, C. Mathys, T. FitzGerald, and G. Pezzulo (2015). Active inference and epistemic value. *Cognitive Neuroscience* 6, 187–214.

Iteration notes. Sections are deliberately modular: the geometry (§2), the inference back end (§4–§5), and the control/active layers (§6–§7) can each be swapped or deepened without disturbing the others, because they communicate only through the likelihood, the distance estimate $\hat{D}$, and the eye posture $\boldsymbol{\phi}$. The bibliography is a starting set of canonical sources; verify editions and page ranges before citing formally.