ACTIVE VISION IN BINOCULAR DEPTH ESTIMATION: A TOP-DOWN PERSPECTIVE

Matteo Priorelli
Institute of Cognitive Sciences and Technologies
National Research Council of Italy
Padova, Italy

Giovanni Pezzulo
Institute of Cognitive Sciences and Technologies
National Research Council of Italy
Rome, Italy

Ivilin Peev Stoianov
Institute of Cognitive Sciences and Technologies
National Research Council of Italy
Padova, Italy
ivilinpeev.stoianov@cnr.it

ABSTRACT

Depth estimation is an ill-posed problem: objects of different shapes or dimensions, even if at different distances, may project to the same image on the retina. Our brain uses several cues for depth estimation, including monocular cues such as motion parallax and binocular cues like diplopia. However, it is still unclear how the computations required for depth estimation are implemented in biologically plausible ways. State-of-the-art approaches to depth estimation based on deep neural networks implicitly describe the brain as a hierarchical feature detector. Instead, we propose an alternative approach that casts depth estimation as a problem of active inference. We show that depth can be inferred by inverting a hierarchical generative model that simultaneously predicts the eyes projections from a 2D belief over an object. Model inversion consists of a series of biologically plausible, homogeneous transformations based on Predictive Coding principles. Under the plausible assumption of a nonuniform fovea resolution, depth estimation favors an active vision strategy that fixates the object with the eyes, rendering the depth belief more accurate. This strategy is not realized by first fixating on a target and then estimating the depth, but by combining the two processes through action-perception cycles, with a similar mechanism of the saccades during object recognition. The proposed approach requires only local (top-down and bottom-up) message passing that can be implemented in biologically plausible neural circuits.

1 Introduction

Depth estimation is a complex process involving continuous activation of every level of the visual cortex and even higher-level regions. Disparity-sensitive cells of a different kind can be found early in the visual cortex [1, 2], and it seems that the resulting signals travel through the dorsal and ventral pathways for different purposes: parietal regions (in particular anterior and lateral intraparietal) have major contributions in depth estimation for visually-guided actions in hand and eye movements [3, 4], while the inferotemporal cortex supports the creation of 3D shapes based on the relative disparity between objects [2, 5]. The brain can rely on several cues to estimate the depth of objects, the most important ones being: (i) the binocular disparity that allows the visual cortex to have access to two different perspectives of the same environment; (ii) the motion parallax effect that happens when objects at a greater distance move slower than nearby objects; (iii) the angular difference between the eyes when fixating the same object (*vergence*). However, the exact contribution of these mechanisms on the overall depth estimation and, critically, where and how the information processing of these signals occurs, is still unclear.

Traditionally, the visual cortex has been associated with a feature detector: as the sensory signals climb the hierarchy, more complex features are detected by increasing cortical levels, such that from lines and contours high-level representations of objects are constructed. This view has inspired the development of Convolutional Neural Networks, which have led to remarkable results in object recognition tasks [6]. Despite its success, this bottom-up approach is not able to capture several top-down mechanisms that affect our everyday perception of the external world, as in the case of

visual illusions [7]. In recent years, a different perspective has emerged, based on Predictive Coding theories, that views these illusions not just as unexpected phenomena but as expressions of the main mechanism through which our brain is able to efficiently predict and act over the environment [8, 9]. Under this view, the biases we perceive are actually hints to better minimize the errors between our sensations and our predictions [10]. Furthermore, vision is increasingly considered to be an active process that constantly tries to reduce the uncertainty of what will happen next.

In this article, we apply this predictive and inferential view of perception to depth estimation. Specifically, we advance an Active Inference model that is able to estimate the depth of an object based on two projected images, through a process of prediction error minimization and active oculomotor behavior. In our model, depth estimation does not consist of a bottom-up process that detects disparities in the images of the two eyes, but of an inference of top-down projective predictions from a high-level representation of the object. In other words, the estimation of the object's depth naturally arises by inverting a visual generative model where the resulting prediction errors flow up the cortical hierarchy, in contrast with the direct process occurring with neural networks.

2 Methods

The theory of Active Inference assumes that an agent is endowed with a generative model that makes predictions over sensory observations [10, 11, 12, 13], as shown in Figure 1. The discrepancy between predictions and observations generates a prediction error that is minimized in order to deal with a dynamical environment and to anticipate what will happen next. This generative model hinges on three components encoded in generalized coordinates of increasing temporal orders (e.g., position, velocity, acceleration, and so on): hidden states $\tilde{\boldsymbol{x}}$, hidden causes $\tilde{\boldsymbol{v}}$, and sensory signals $\tilde{\boldsymbol{s}}$. These components are expressed through a non-linear system that defines the prediction of sensory signals and the evolution of hidden states and causes across time:

$$\begin{aligned} \tilde{\boldsymbol{s}} &= \tilde{\boldsymbol{g}}(\tilde{\boldsymbol{x}}) + \boldsymbol{w}_s \\ \mathcal{D}\tilde{\boldsymbol{x}} &= \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) + \boldsymbol{w}_x \end{aligned} \tag{1}$$

In this context, $\mathcal{D}$ denotes a differential operator that shifts all temporal orders by one, such as $\mathcal{D}\tilde{\boldsymbol{x}} = [\boldsymbol{x}', \boldsymbol{x}'', \boldsymbol{x}''', \dots]$. Furthermore, $\boldsymbol{w}_s$ and $\boldsymbol{w}_x$ stand as noise terms drawn from a Gaussian distribution. The considered joint probability is divided into distinct distributions:

$$p(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) = p(\tilde{\boldsymbol{s}}|\tilde{\boldsymbol{x}})p(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{v}})p(\tilde{\boldsymbol{v}}) \tag{2}$$

Typically, each distribution is approximated with Gaussian functions:

$$\begin{aligned} p(\tilde{\boldsymbol{s}}|\tilde{\boldsymbol{x}}) &= \mathcal{N}(\tilde{\boldsymbol{g}}(\tilde{\boldsymbol{x}}), \tilde{\boldsymbol{\pi}}_s^{-1}) \\ p(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{v}}) &= \mathcal{N}(\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}), \tilde{\boldsymbol{\pi}}_x^{-1}) \\ p(\tilde{\boldsymbol{v}}|\boldsymbol{\eta}) &= \mathcal{N}(\boldsymbol{\eta}, \tilde{\boldsymbol{\pi}}_v^{-1}) \end{aligned} \tag{3}$$

where $\boldsymbol{\eta}$ is a prior, and the distributions are expressed in terms of precisions (or inverse variances) $\tilde{\boldsymbol{\pi}}_s$, $\tilde{\boldsymbol{\pi}}_x$, and $\tilde{\boldsymbol{\pi}}_v$.

Following a variational inference method [14], these distributions are inferred via approximate posteriors $q(\tilde{\boldsymbol{x}})$ and $q(\tilde{\boldsymbol{v}})$. Under appropriate assumptions, minimizing the Variational Free Energy (VFE) $\mathcal{F}$, defined as the disparity between the KL divergence of real and approximate posteriors and the log evidence:

$$\mathcal{F} = \mathop{\mathbb{E}}_{q(\tilde{\boldsymbol{x}})}\left[\ln \frac{q(\tilde{\boldsymbol{x}})}{p(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})}\right] = \mathop{\mathbb{E}}_{q(\tilde{\boldsymbol{x}})}\left[\ln \frac{q(\tilde{\boldsymbol{x}})}{p(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{y}})}\right] - \ln p(\tilde{\boldsymbol{y}}) \tag{4}$$

leads to the minimization of prediction errors. The belief updates $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\nu}}$, concerning hidden states and hidden causes respectively, expand as follows:

$$\begin{aligned} \dot{\tilde{\boldsymbol{\mu}}} - \mathcal{D}\tilde{\boldsymbol{\mu}} &= -\partial_\mu \mathcal{F} = \partial \tilde{\boldsymbol{g}}^T \tilde{\boldsymbol{\pi}}_s \tilde{\boldsymbol{\varepsilon}}_s + \partial_\mu \tilde{\boldsymbol{f}}^T \tilde{\boldsymbol{\pi}}_\mu \tilde{\boldsymbol{\varepsilon}}_\mu - \mathcal{D}^T \tilde{\boldsymbol{\pi}}_\mu \tilde{\boldsymbol{\varepsilon}}_\mu \\ \dot{\tilde{\boldsymbol{\nu}}} - \mathcal{D}\tilde{\boldsymbol{\nu}} &= -\partial_\nu \mathcal{F} = \partial_\nu \tilde{\boldsymbol{f}}^T \tilde{\boldsymbol{\pi}}_\mu \tilde{\boldsymbol{\varepsilon}}_\mu - \tilde{\boldsymbol{\pi}}_\nu \tilde{\boldsymbol{\varepsilon}}_v \end{aligned} \tag{5}$$

Here, $\tilde{\boldsymbol{\varepsilon}}_s$, $\tilde{\boldsymbol{\varepsilon}}_\mu$, and $\tilde{\boldsymbol{\varepsilon}}_\nu$ denote the prediction errors of sensory signals, dynamics, and priors:

$$\begin{aligned} \tilde{\boldsymbol{\varepsilon}}_s &= \tilde{\boldsymbol{s}} - \tilde{\boldsymbol{g}}(\tilde{\boldsymbol{\mu}}) \\ \tilde{\boldsymbol{\varepsilon}}_\mu &= \mathcal{D}\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}}) \\ \tilde{\boldsymbol{\varepsilon}}_\nu &= \tilde{\boldsymbol{\nu}} - \boldsymbol{\eta} \end{aligned} \tag{6}$$

A simple Active Inference scheme can handle various tasks, yet the effectiveness of the theory stems from a hierarchical structure that enables the brain to grasp the hierarchical associations between sensory observations and their causes

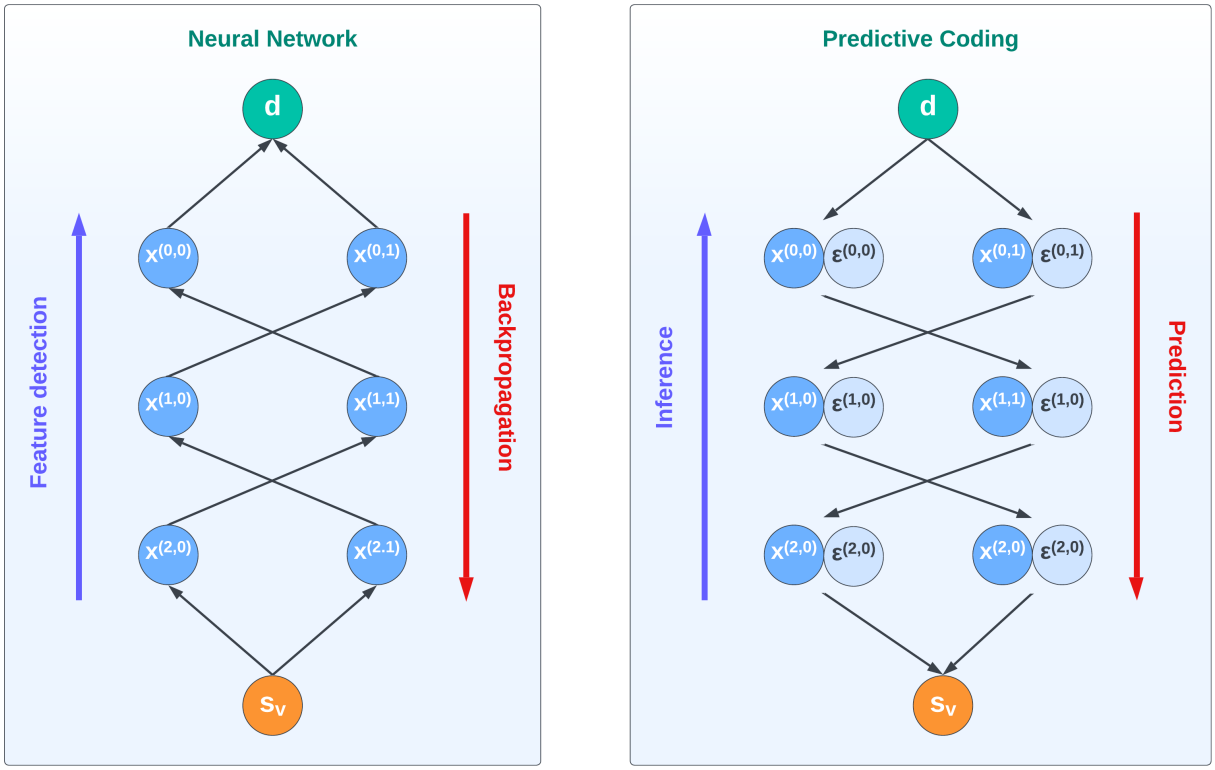

Figure 1: Information processing in neural networks (left) and Predictive Coding (right). In a neural network, the visual observation $\mathbf{s}_v$ travels through the cortical hierarchy in a bottom-up way, detecting increasingly more complex features $\boldsymbol{x}^{(i,j)}$ and eventually estimating the depth of an object $\boldsymbol{d}$. The descending projections are considered here as feedback signals that convey backpropagation errors. In contrast, in a Predictive Coding Network the depth $\boldsymbol{d}$ is a high-level belief generating a visual prediction that is compared with the observation. This process leads to a cascade of prediction errors $\boldsymbol{\varepsilon}^{(i,j)}$ – associated with each intermediate prediction $\boldsymbol{x}^{(i,j)}$ – that are minimized throughout the hierarchy, eventually inferring the correct belief (for details, see the Methods section).

[15]. Specifically, the model delineated above can be expanded by linking each hidden cause with another generative model; as a result, the prior becomes the prediction from the layer above, while the observation becomes the likelihood of the layer below:

$$\begin{aligned}\dot{\tilde{\boldsymbol{\mu}}}^{(j)} &= \mathcal{D}\tilde{\boldsymbol{\mu}}^{(j)} + \partial\tilde{\boldsymbol{g}}^{(j)T}\tilde{\boldsymbol{\pi}}_{\nu}^{(j)}\tilde{\boldsymbol{\varepsilon}}_{\nu}^{(j)} + \partial_{\mu}\tilde{\boldsymbol{f}}^{(j)T}\tilde{\boldsymbol{\pi}}_{\mu}^{(j)}\tilde{\boldsymbol{\varepsilon}}_{\mu}^{(j)} - \mathcal{D}^{T}\tilde{\boldsymbol{\pi}}_{\mu}^{(j)}\tilde{\boldsymbol{\varepsilon}}_{\mu}^{(j)} \\ \dot{\tilde{\boldsymbol{\nu}}}^{(j)} &= \mathcal{D}\tilde{\boldsymbol{\nu}}^{(j)} + \partial_{\nu}\tilde{\boldsymbol{f}}^{(j)T}\tilde{\boldsymbol{\pi}}_{\mu}^{(j)}\tilde{\boldsymbol{\varepsilon}}_{\mu}^{(j)} - \tilde{\boldsymbol{\pi}}_{\nu}^{(j-1)}\tilde{\boldsymbol{\varepsilon}}_{\nu}^{(j-1)}\end{aligned} \tag{7}$$

Here:

$$\begin{aligned}\tilde{\boldsymbol{\varepsilon}}_{\mu}^{(j)} &= \mathcal{D}\tilde{\boldsymbol{\mu}}_{\mu}^{(j)} - \tilde{\boldsymbol{f}}^{(j)}(\tilde{\boldsymbol{\mu}}^{(j)}, \tilde{\boldsymbol{\nu}}^{(j)}) \\ \tilde{\boldsymbol{\varepsilon}}_{\nu}^{(j)} &= \tilde{\boldsymbol{\mu}}_{\nu}^{(j+1)} - \tilde{\boldsymbol{g}}^{(j)}(\tilde{\boldsymbol{\mu}}^{(j)})\end{aligned} \tag{8}$$

In contrast, the execution of action is accomplished by minimizing the proprioceptive component of the VFE concerning the motor control signals $\boldsymbol{a}$:

$$\dot{\boldsymbol{a}} = -\partial_a \mathcal{F}_p = -\partial_a \tilde{\boldsymbol{s}}_p \tilde{\boldsymbol{\pi}}_p \tilde{\boldsymbol{\varepsilon}}_p \tag{9}$$

Here, $\partial_a \boldsymbol{s}_p$ stands for the partial derivative of proprioceptive observations regarding the motor control signals, $\tilde{\boldsymbol{\pi}}_p$ are the precisions of the proprioceptive generative model, and $\tilde{\boldsymbol{\varepsilon}}_p$ are the generalized proprioceptive prediction errors:

$$\tilde{\boldsymbol{\varepsilon}}_p = \tilde{\boldsymbol{s}}_p - \tilde{\boldsymbol{g}}_p(\tilde{\boldsymbol{\mu}}) \tag{10}$$

In summary, in Active Inference goal-directed behavior is generally possible by firstly biasing the belief over the hidden states through a specific cause. The latter acts as a prior that encodes the agent's belief about the state of affairs of the world. In this context, action follows because the hidden states generate a proprioceptive prediction error that is

suppressed through a reflex arc [16]. For instance, suppose that the agent has to rotate the arm by a few degrees. The belief over its arm's angle is subject to two opposing forces:

$$\dot{\boldsymbol{\mu}} = \boldsymbol{\pi}_p \boldsymbol{\varepsilon}_p - \boldsymbol{\pi}_v \boldsymbol{\varepsilon}_v \tag{11}$$

one coming from above (pulling it toward its expectation) and one coming from below (pulling it toward what it is currently perceiving). The tradeoff between the two forces is expressed in terms of the precisions $\boldsymbol{\pi}_p$ and $\boldsymbol{\pi}_v$, which encode the agent's level of confidence about the particular prediction errors. By appropriately tuning the precision parameters, it is possible to smoothly push the belief toward a desired state, eventually driving the real arm through Equation 9.

3 Results

3.1 Homogeneous transformations as Hierarchical Active Inference

Classical Predictive Coding models are passive, in the sense that the model cannot select its visual stimuli [8]. Rather, our Active Inference model can actively control "eyes", to sample the preferred stimuli that reduce prediction errors.

State-of-the-art implementations of oculomotor behavior in Active Inference rely on a latent state (or belief) over the eye angle, and attractors are usually defined directly in the polar domain [17, 18]. While having interesting implications for simulating saccadic and smooth pursuit eye movements, such models do not consider the fact that eyes fixate the target from two different perspectives. A similar limitation can be found in models of reaching, in which the 3D position of the object to reach is directly provided as a visual observation [19, 20]. Furthermore, since there is only a single level specified in polar coordinates, if one wants to fixate or reach a target defined in Cartesian coordinates, a relatively complex dynamics function has to be defined at that level.

Using a *hierarchical kinematic model* – based on Active Inference – that includes both intrinsic (e.g., joint angles and limb lengths) and extrinsic (e.g., Cartesian positions) coordinates affords efficient control of a simulated body [21]. The extrinsic configuration of the motor plant is computed hierarchically, as shown in Figure 2. For the relatively simple kinematic control tasks targeted in [21], these computations only required two simple transformations between reference frames: translations and rotations. However, a hierarchical kinematic model can be easily extended to afford more complex tasks, requiring different transformations.

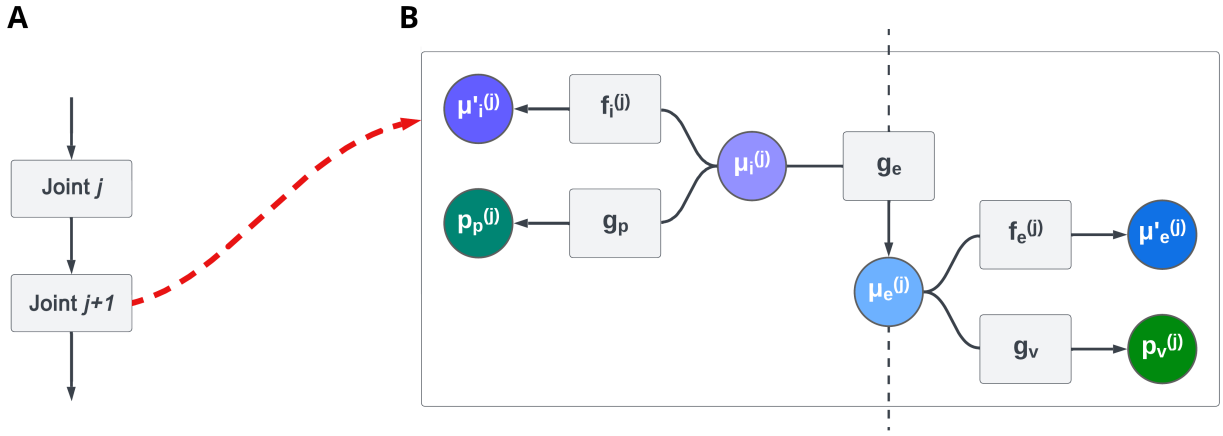

Figure 2: (**A**) An example of a portion of a kinematic plant. (**B**) Factor graph of a single level j of the hierarchical kinematic model composed of intrinsic $\boldsymbol{\mu}_i^{(j)}$ and extrinsic $\boldsymbol{\mu}_e^{(j)}$ beliefs. These beliefs generate proprioceptive and visual predictions $\boldsymbol{p}_p^{(j)}$ and $\boldsymbol{p}_v^{(j)}$ through generative models $\boldsymbol{g}_p$ and $\boldsymbol{g}_v$, respectively. Furthermore, the beliefs predict trajectories – here, only velocities $\boldsymbol{\mu}_i'^{(j)}$ and $\boldsymbol{\mu}_e'^{(j)}$ – through dynamics functions $\boldsymbol{f}_i^{(j)}$ and $\boldsymbol{f}_e^{(j)}$. Note that the extrinsic belief of level $j-1$ acts as a prior for layer j through a kinematic generative model $\boldsymbol{g}_e$.
See [21] for more details.

In robotics, transformations between reference frames are usually realized through the multiplication of a linear transformation matrix. These operations can be decomposed into simpler steps where *homogeneous coordinates* are multiplied once at a time through the chain rule, allowing more efficient computations. Specifically, if the x and y axes

represent a Cartesian plane, a homogeneous representation augments the latter with an additional dimension called *projective space*. In this new system, multiplying the point coordinates by the same factor keeps the mapping unaltered, i.e., $(p_x, p_y, 1) \equiv (p_z p_x, p_z p_y, p_z)$.

Affine transformations preserve parallel lines and have the following form:

$$\boldsymbol{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

where the last row ensures that every point always maps to the same plane. By the chain rule, a point in the plane $\boldsymbol{p}^0$ can be rotated and translated by multiplication of the corresponding transformations:

$$\boldsymbol{p}^f = \begin{bmatrix} p_x^f \\ p_y^f \\ 1 \end{bmatrix} = \begin{bmatrix} c_\theta & -s_\theta & 0 \\ s_\theta & c_\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & l_x \\ 0 & 1 & l_y \\ 0 & 0 & 1 \end{bmatrix} \cdot \boldsymbol{p}^0 = \begin{bmatrix} c_\theta(p_x^0 + l_x) - s_\theta(p_y^0 + l_y) \\ s_\theta(p_x^0 + l_x) + c_\theta(p_y^0 + l_y) \\ 1 \end{bmatrix} \tag{13}$$

where s_θ and c_θ are the sine and cosine of the rotation θ, while l_x and l_y are the coordinates of the translation.

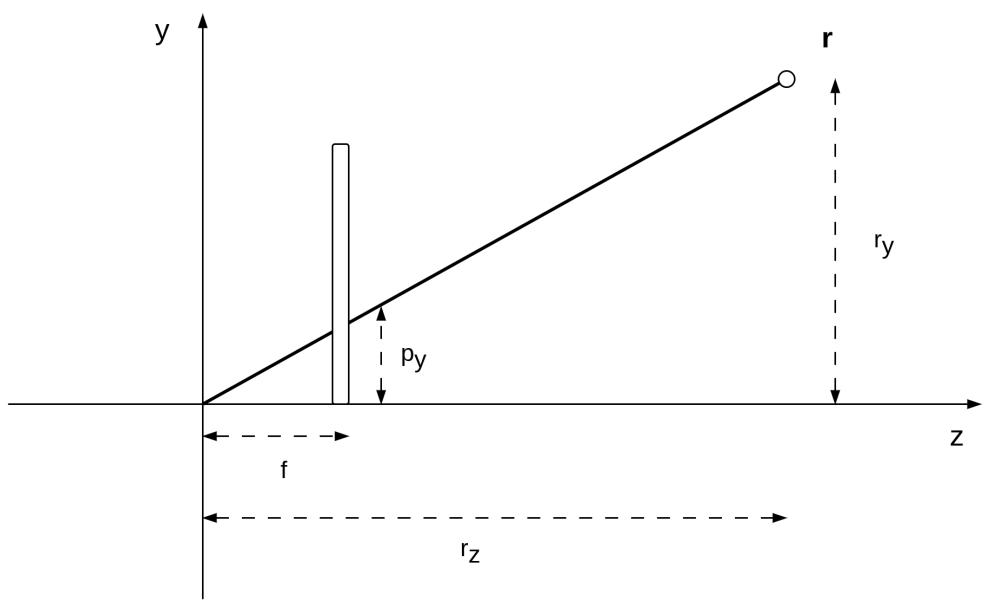

Figure 3: Projection of a 3D point in the camera plane (only two dimensions are shown). The y coordinate of the real point r_y and of the projected point p_y are related by the ratio between the focal length f and the real point depth r_z.

By appropriately changing the values of the matrix, more affine transformations can be obtained, such as shearing or scaling. Critically, if the last row is modified, it is also possible to realize perspective projections:

$$\boldsymbol{p}^f = \begin{bmatrix} p_x^f \\ p_y^f \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ z_x & z_y & 1 \end{bmatrix} \boldsymbol{p}^0 = \begin{bmatrix} p_x^0 \\ p_y^0 \\ z_x p_x^0 + z_y p_y^0 + 1 \end{bmatrix} \tag{14}$$

so that the new point does not map on the same plane $p_z = 1$ as before. Thus, to map it back to the Cartesian plane, we divide the x and y coordinates by the last element:

$$\begin{bmatrix} p_x^0 \\ p_y^0 \\ z_x p_x^0 + z_y p_y^0 + 1 \end{bmatrix} \equiv \begin{bmatrix} \frac{p_x^0}{z_x p_x^0 + z_y p_y^0 + 1} \\ \frac{p_y^0}{z_x p_x^0 + z_y p_y^0 + 1} \\ 1 \end{bmatrix} \tag{15}$$

This special transformation is critical for computer vision in order to project points on an image plane or to estimate the depth of an object. If we have a 3D point $\boldsymbol{a} = (a_x, a_y, a_z, 1)$ expressed in homogeneous coordinates, we can obtain the

corresponding 2D point $\boldsymbol{p}$ projected in the camera plane by first performing a roto-translation – similar to Equation 13 – through a matrix that encodes where the camera is located and oriented (the *extrinsic parameters*):

$$\boldsymbol{r} = \begin{bmatrix} r_x \\ r_y \\ r_z \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}^T & -\boldsymbol{R}^T\boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \boldsymbol{w} \tag{16}$$

and then scaling and converting the point to 2D through the so-called *camera matrix*:

$$\boldsymbol{p}' = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{r} = \begin{bmatrix} fr_x \\ fr_y \\ r_z \end{bmatrix} \tag{17}$$

The projection is performed by multiplying the depth coordinate z by the focal length f, which represents the distance of the image plane from the origin. As before, since the homogeneous representation is up to a scale factor, to transform the point $\boldsymbol{p}'$ into the Cartesian space we divide the camera coordinates by the depth coordinate p'_z (as shown in Figure 3):

$$p_x = \frac{fr_x}{r_z} \qquad\qquad p_y = \frac{fr_y}{r_z} \tag{18}$$

Keeping this in mind, we generalize the deep kinematic model of [21] by assuming that each level sequentially applies a series of homogeneous transformations. Specifically, the first belief, called $\boldsymbol{\mu}_t^{(i)}$, contains information about a particular transformation (e.g., by which angle to rotate or by which length to translate a point).

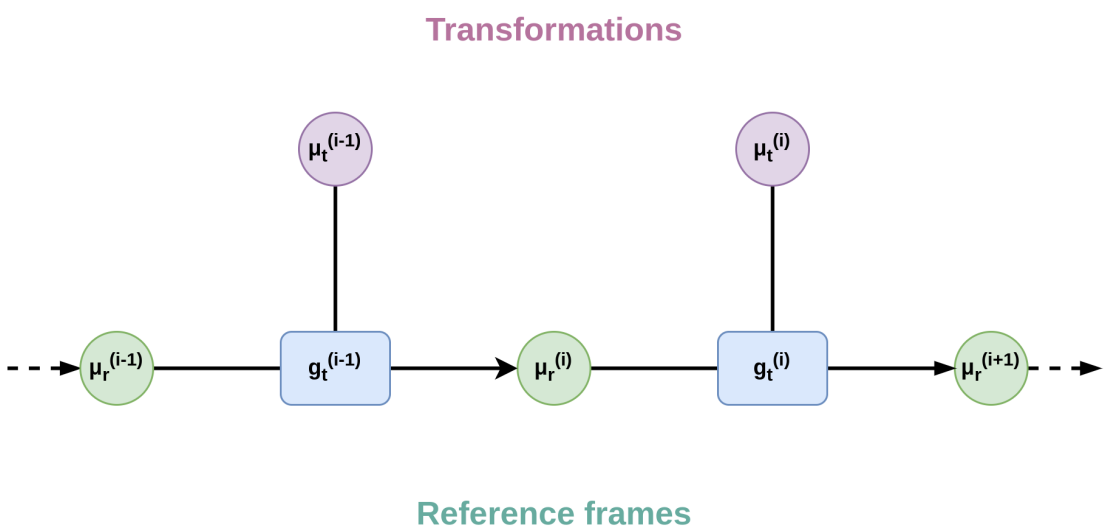

Figure 4: Representation of the hierarchical relationships of a generalized model with homogeneous transformations. The belief over a reference frame $\boldsymbol{\mu}_r$ of level i is passed to a function $\boldsymbol{g}_t$ encoding a homogeneous transformation along with a belief over a particular transform $\boldsymbol{\mu}_t$ (e.g., angle for rotation or length for translation), generating the reference frame of level $i+1$.

This belief then generates a homogeneous transformation relative to that Degree of Freedom (DoF), which is then multiplied by a second belief expressed in a particular reference frame (as exemplified in Figure 4):

$$\boldsymbol{\mu}_r^{(i+1)} = \boldsymbol{g}_t^{(i)}(\boldsymbol{\mu}_t^{(i)}, \boldsymbol{\mu}_r^{(i)}) = \boldsymbol{T}^{(i)}(\boldsymbol{\mu}_t^{(i)}) \cdot \boldsymbol{\mu}_r^{(i)} \tag{19}$$

This equation leads to simple gradient computations through the generated prediction error $\boldsymbol{\varepsilon}_r^{(i+1)}$ which is needed to iteratively update the two beliefs:

$$\begin{aligned} \frac{\partial \boldsymbol{g}_t^{(i)}}{\partial \boldsymbol{\mu}_r^{(i)}}^T \boldsymbol{\varepsilon}_r^{(i+1)} &= \boldsymbol{T}^{(i)T} \cdot \boldsymbol{\varepsilon}_r^{(i+1)} \\ \frac{\partial \boldsymbol{g}_t^{(i)}}{\partial \boldsymbol{\mu}_t^{(i)}}^T \boldsymbol{\varepsilon}_r^{(i+1)} &= \frac{\partial \boldsymbol{T}^{(i)}}{\partial \boldsymbol{\mu}_t^{(i)}} \odot [\boldsymbol{\varepsilon}_r^{(i+1)} \cdot \boldsymbol{\mu}_r^{(i)T}] \end{aligned} \tag{20}$$

where $\odot$ is the element-wise product.

3.2 A hierarchical generative model for binocular depth estimation

In this section, we explain how depth estimation arises by inverting the projective predictions of the two eyes, using a hierarchical generative model. For simplicity, we consider an agent interacting with a 2D world, where the depth is the x coordinate. Nonetheless, the same approach could be used to estimate the depth of a 3D object. We construct the generative model hierarchically, starting from a belief $\boldsymbol{\mu}_a = (\mu_{a,x}, \mu_{a,y}, 1)$ about the absolute 2D position of an object encoded in homogeneous coordinates, where $\mu_{a,x}$ is the depth belief. Then, two parallel pathways generate specular predictions $\boldsymbol{p}_r^{(i)}$ that receive the eye angles encoded in a common vergence-accommodation belief $\boldsymbol{\mu}_\theta = (\theta_a, \theta_v)$ – and transform the absolute coordinates of the object into the two reference frames relative to the eyes:

$$\boldsymbol{p}_r^{(i)} = \boldsymbol{g}_a^{(i)}(\boldsymbol{\mu}_a, \boldsymbol{\mu}_\theta) = \boldsymbol{T}^{(i)}(\boldsymbol{\mu}_\theta) \cdot \boldsymbol{\mu}_a \tag{21}$$

where $\boldsymbol{T}^{(i)}(\boldsymbol{\mu}_\theta)$ is the homogeneous transformation corresponding to the extrinsic parameters of the camera:

$$\boldsymbol{T}^{(i)}(\boldsymbol{\mu}_\theta) = \begin{bmatrix} c_\theta^{(i)} & s_\theta^{(i)} & -l^{(i)} s_\theta^{(i)} \\ -s_\theta^{(i)} & c_\theta^{(i)} & -l^{(i)} c_\theta^{(i)} \\ 0 & 0 & 1 \end{bmatrix} \tag{22}$$

Here, $l^{(i)}$ is the distance between an eye and the origin (i.e., the middle of the eyes), and the absolute eye angles are:

$$\theta^{(0)} = \theta_a - \theta_v \qquad\qquad \theta^{(1)} = \theta_a + \theta_v \tag{23}$$

Each of these beliefs generates a prediction over a point projected to the corresponding camera plane:

$$\boldsymbol{p}_c^{(i)} = \boldsymbol{g}_r(\boldsymbol{\mu}_r^{(i)}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & f & 0 \end{bmatrix} \cdot \frac{\boldsymbol{\mu}_r^{(i)}}{\mu_{r,x}^{(i)}} \tag{24}$$

Figure 5 provides a neural-level illustration of the model, with the two branches originating from the two beliefs at the top. Note that while the eye angles belief $\boldsymbol{\mu}_\theta$ generates separate predictions for the two eyes, proprioceptive predictions directly encode angles in the vergence-accommodation system, which is used for action execution [10].

The absolute point belief (encoded in generalized coordinates up to the 2nd level) is updated as follows:

$$\begin{aligned} \dot{\tilde{\boldsymbol{\mu}}}_a &= \begin{bmatrix} \boldsymbol{\mu}'_a + \sum_i \partial_a \boldsymbol{g}_a^T \boldsymbol{\varepsilon}_r^{(i)} + \partial \boldsymbol{f}_a^T \boldsymbol{\varepsilon}_{\mu,a} \\ -\boldsymbol{\varepsilon}_{\mu,a} \end{bmatrix} \\ \dot{\boldsymbol{\varepsilon}}_r^{(i)} &= \boldsymbol{\mu}_r^{(i)} - \boldsymbol{g}_a^{(i)}(\boldsymbol{\mu}_a, \boldsymbol{\mu}_\theta) - \boldsymbol{\varepsilon}_r^{(i)}/\pi_r^{(i)} \\ \dot{\boldsymbol{\varepsilon}}_{\mu,a} &= \boldsymbol{\mu}'_a - \boldsymbol{f}_a(\boldsymbol{\mu}_a) - \boldsymbol{\varepsilon}_{\mu,a}/\gamma_a \end{aligned} \tag{25}$$

where $\pi_r^{(i)}$ and $\boldsymbol{\varepsilon}_r^{(i)}$ are the precisions and prediction errors of the beliefs below, while $\boldsymbol{f}_a$, γ_a and $\boldsymbol{\varepsilon}_{\mu,a}$ are the function, precision, and prediction error of the dynamics of the same belief.

This belief is thus subject to different prediction errors $\boldsymbol{\varepsilon}_r^{(i)}$ coming from the two eyes, and the depth of the point is estimated by averaging these two pathways. Furthermore, an attractor can be defined in the dynamics function $\boldsymbol{f}_a$ if one wants to control the object encoded in absolute coordinates – e.g., for reaching or grasping tasks.

Similarly, the belief update equation for $\boldsymbol{\mu}_\theta$ is the following:

$$\begin{aligned} \dot{\tilde{\boldsymbol{\mu}}}_\theta &= \begin{bmatrix} \boldsymbol{\mu}'_\theta + \sum_i \partial_\theta \boldsymbol{g}_a^T \boldsymbol{\varepsilon}_r^{(i)} + \partial \boldsymbol{g}_p^T \boldsymbol{\varepsilon}_p + \partial \boldsymbol{f}_\theta^T \boldsymbol{\varepsilon}_{\mu,\theta} \\ -\boldsymbol{\varepsilon}_{\mu,\theta} \end{bmatrix} \\ \dot{\boldsymbol{\varepsilon}}_p &= \boldsymbol{s}_p - \boldsymbol{g}_p(\boldsymbol{\mu}_\theta) - \boldsymbol{\varepsilon}_p/\pi_p \\ \dot{\boldsymbol{\varepsilon}}_{\mu,\theta} &= \boldsymbol{\mu}'_\theta - \boldsymbol{f}_\theta(\boldsymbol{\mu}_\theta) - \boldsymbol{\varepsilon}_{\mu,\theta}/\gamma_\theta \end{aligned} \tag{26}$$

where π_p, $\boldsymbol{\varepsilon}_p$, $\boldsymbol{s}_p$, and $\boldsymbol{g}_p$ are the proprioceptive precision, prediction error, observation, and likelihood function – which is an identity mapping in the following simulations – while $\boldsymbol{f}_\theta$, γ_θ and $\boldsymbol{\varepsilon}_{\mu,\theta}$ are the function, precision, and prediction error of the belief dynamics.

This belief, besides being affected by the proprioceptive contribution, is also subject to the same prediction errors $\boldsymbol{\varepsilon}_r^{(i)}$ in the same way as the absolute belief. The overall free energy can be thus minimized along two different pathways: (i) by changing the belief about the absolute location of the object (including the depth), or (ii) by modifying the angle

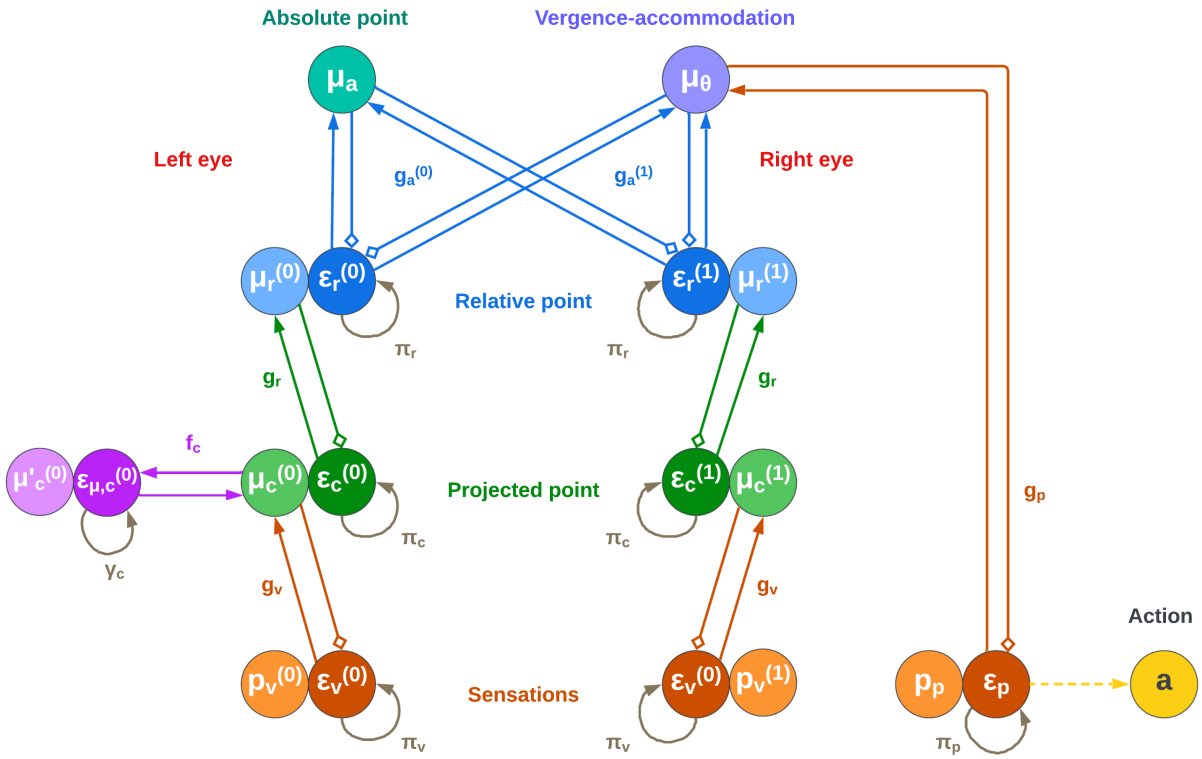

Figure 5: Neural-level implementation of a hierarchical generative model to estimate the depth of a point, through Active Inference. Small squares indicate inhibitory connections. Differently from a neural network, depth is estimated by first generating, from a point in absolute coordinates $\boldsymbol{\mu}_a$ and the vergence-accommodation angles $\boldsymbol{\mu}_\theta$, two predictions $\boldsymbol{p}_r^{(i)}$ of the point relative to each eye. The new belief in turn computes a projection $\boldsymbol{p}_c^{(i)}$, and finally a visual prediction $\boldsymbol{p}_v^{(i)}$. The predictions are then compared with the visual observations, generating prediction errors throughout the hierarchy and eventually driving the beliefs at the top toward the correct values. Note that eye movements are directly triggered to suppress the proprioceptive prediction error $\boldsymbol{\varepsilon}_p$. Intentional eye movements (e.g., for target fixation) can be instead achieved by setting a prior in the dynamics function $\boldsymbol{f}_c$ of the belief over the projected point $\boldsymbol{\mu}_c$ (note that for better readability, the figure only shows the dynamics function $\boldsymbol{f}_c$.

of fixation of the eyes. As will be shown in the next section, the possibility of using both pathways may create some stability issues during goal-directed movements.

Also in this case, an attractor can be specified in the dynamics function $\boldsymbol{f}_\theta$ if one wants to explicitly control the dynamics of the eyes – e.g., not fixating a point on the camera plane, but rotating the eyes along a particular direction or by a particular angle.

Finally, the belief update equation for the projected point $\boldsymbol{\mu}_c^{(i)}$ is the following:

$$
\begin{aligned}
\dot{\tilde{\boldsymbol{\mu}}}_c &= \begin{bmatrix} \boldsymbol{\mu}_c' - \boldsymbol{\varepsilon}_c^{(i)} + \partial \boldsymbol{g}_v^T \boldsymbol{\varepsilon}_v^{(i)} + \partial \boldsymbol{f}_c^{(i)T} \boldsymbol{\varepsilon}_{\mu,c}^{(i)} \\ -\boldsymbol{\varepsilon}_{\mu,c}^{(i)} \end{bmatrix} \\
\dot{\boldsymbol{\varepsilon}}_c^{(i)} &= \boldsymbol{\mu}_c^{(i)} - \boldsymbol{g}_r(\boldsymbol{\mu}_r^{(i)}) - \boldsymbol{\varepsilon}_c^{(i)}/\pi_c^{(i)} \\
\dot{\boldsymbol{\varepsilon}}_v^{(i)} &= \boldsymbol{s}_v^{(i)} - \boldsymbol{g}_v(\boldsymbol{\mu}_c^{(i)}) - \boldsymbol{\varepsilon}_v^{(i)}/\pi_v^{(i)} \\
\dot{\boldsymbol{\varepsilon}}_{\mu,c}^{(i)} &= \boldsymbol{\mu}_c^{(i)\prime} - \boldsymbol{f}_c^{(i)}(\boldsymbol{\mu}_c^{(i)}) - \boldsymbol{\varepsilon}_{\mu,c}^{(i)}/\gamma_c^{(i)}
\end{aligned}
\tag{27}
$$

where $\pi_c^{(i)}$ and $\boldsymbol{\varepsilon}_c^{(i)}$ are the precisions and prediction errors of the beliefs below, $\pi_v^{(i)}$, $\boldsymbol{\varepsilon}_v^{(i)}$, $\boldsymbol{s}_v^{(i)}$, and $\boldsymbol{g}_v$ are the visual precision, prediction error, observation, and likelihood function, while $\boldsymbol{f}_c$, γ_c and $\boldsymbol{\varepsilon}_{\mu,c}$ are the function, precision, and prediction error of the belief dynamics. Note that in the following simulations, we approximated $\boldsymbol{g}_v$ by a simple identity mapping, so that $\boldsymbol{s}_v^{(i)}$ conveys a Cartesian position.

Differently from the belief over the eye angles $\boldsymbol{\mu}_\theta$ which is only biased by the likelihoods of the levels below, this belief is both subject to a prior encoded in $\boldsymbol{\varepsilon}_c^{(i)}$, and a visual likelihood from $\boldsymbol{\varepsilon}_v^{(i)}$.

3.3 Active vision and target fixation with action-perception cycles

The model advanced here can not only infer the depth of a point, but also fixate it using active vision. This is possible by specifying an appropriate attractor in the dynamics function of the last belief $\boldsymbol{\mu}_c^{(i)}$, or in other words, by setting an "intention" in both eyes [22], such that the projected position will be at the center of the camera planes:

$$\begin{aligned} \boldsymbol{e}_c^{(i)} &= \boldsymbol{\mu}_c^{(i)} - (1, 0) \\ \boldsymbol{\mu}_c^{(i)\prime} &= \boldsymbol{f}_c^{(i)}(\boldsymbol{\mu}_c^{(i)}) = \lambda \boldsymbol{e}_c^{(i)} \end{aligned} \tag{28}$$

In short, $\boldsymbol{f}_c^{(i)}$ returns a velocity encoding the difference $\boldsymbol{e}_c^{(i)}$ between the current belief and the center of the camera plane – expressed in homogeneous coordinates – multiplied by an attractor gain λ. Thus, the agent thinks that the projected point will be pulled toward the center, with a velocity proportional to $\boldsymbol{e}_c^{(i)}$. The prediction errors generated then travel back through the hierarchy, affecting both the absolute and the eye angles beliefs. Since what we want in this case is modifying the latter (directly generating proprioceptive predictions for movement), the former pathway can be problematic. In fact, if $\boldsymbol{\mu}_a$ already encodes the correct depth of the object, fixation occurs very rapidly; however, if this is not the case, the prediction errors $\boldsymbol{\varepsilon}_r^{(i)}$ are free to flow through all the open pathways, driving the beliefs toward different directions and leading the free energy minimization process to be stuck with incorrect depth coordinate and eye angles [23].

This abnormal behavior can be avoided if we decompose the task into cyclical phases of action and perception [24]. During an action phase, the absolute belief is kept fixed, so that the relative prediction errors $\boldsymbol{\varepsilon}_r^{(i)}$ can only flow toward the belief over the eye angles, resulting in the eyes moving according to the depth belief. During a perception phase, action is blocked (e.g., either by setting the attractor gain or the proprioceptive precision to zero), while $\boldsymbol{\varepsilon}_r^{(i)}$ is free to flow in any direction, which has the result of pushing the depth belief toward the correct value signaled through the sensory observations. Depth estimation is therefore achieved through multiple action and perception cycles until the overall free energy is minimized.

Figure 6 shows a sequence of time frames of a depth estimation task, in which the (perceptual) process of depth estimation and the (active) process of target fixation cyclically alternate in different phases every 100 time steps. As can be seen from the visualization of the point projections, the distance between the real and estimated target positions slowly decreases, while both positions approach the center of the camera planes, affording efficient depth estimation.

3.4 Model comparison

We tested the model introduced in sections 3.2 and 3.3 in a depth estimation task that consists of inferring the 2D position of the object shown in Figure 6. We compared three different versions of the model. In the first version, the eyes are kept in a fixed position, with the eyes parallel to one another (*infer parallel*). In the second and third versions, the model can actively control the eye angles, but the initial values are set at the correct target position (*infer vergence*) or at a random location (*active vision*, as displayed in Figure 6).

Furthermore, the fovea of the simulated "eye" could have either a uniform or a nonuniform resolution, where in the latter condition the object is represented with greater accuracy when it is near the point of fixation. This reflects the fact that the biological fovea has far more receptors at the center than in the peripheral vision, which has been modeled with an exponential link [25]. Specifically, in our implementation the variability $\boldsymbol{\Sigma}_v$ of the Gaussian error in the visual observations (i.e., in the generative process) exponentially increases with the distance d between the point of fixation and the real target position:

$$\boldsymbol{\Sigma}_v = e^{d/k} \tag{29}$$

where k is a scaling factor (equal to 1.5 in the simulations). In the uniform condition, the visual noise was set to zero.

Figure 7 shows the results of the simulations, including accuracy (the number of trials in which the agent successfully predicts the 2D position of the target, left panel), mean error (distance between the real and estimated target position at the end of every trial, middle panel), and estimation time (number of steps needed to correctly estimate the target, right panel). The number of time steps for each phase has been set to 100, as before. The figure shows that depth estimation with parallel eyes (*infer parallel*) in the nonuniform condition results in a very low accuracy, especially when the target position is far from the fixation point. This is to be expected, since in this condition the fovea has low resolution at the periphery. Rather, if the angle of the eyes is set to fixate on the target (*infer vergence*), the accuracy is much higher and

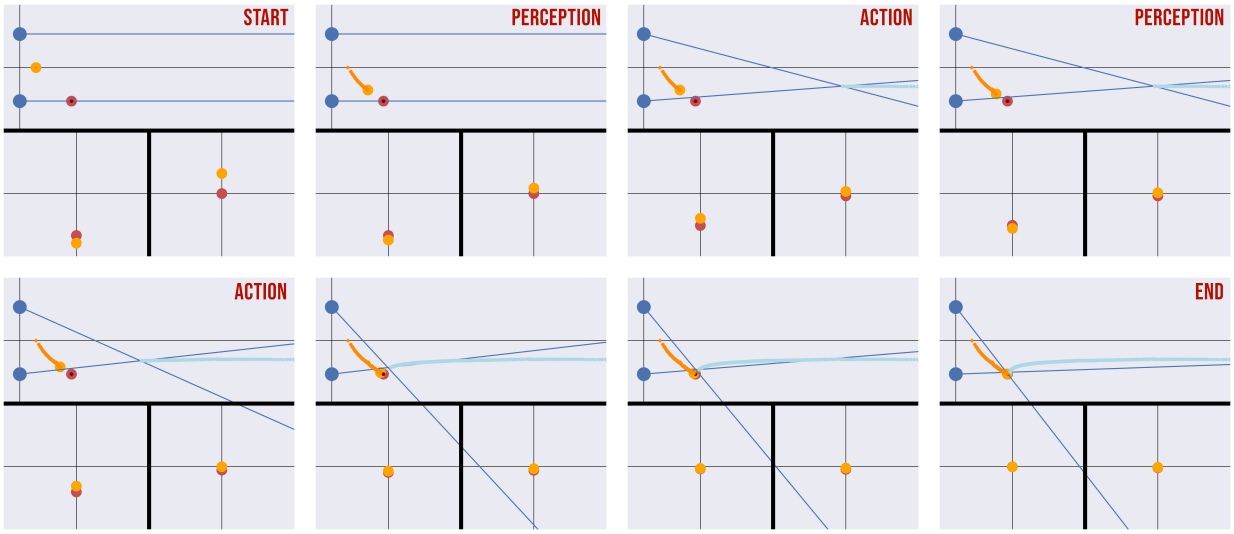

Figure 6: Sequence of time frames of a depth estimation task with simultaneous target fixation. The agent uses alternating action-perception phases to avoid being stuck during the minimization process. Each frame is composed of three images: a third-view perspective of the overall task (top), and a first-view perspective consisting of the projection of the target to the camera planes of both eyes (bottom left and bottom right). In the top panel, the eyes are represented with blue circles, while the real and estimated target positions are in red and orange. The fixation trajectory (when vergence occurs) is represented in cyan. The thin blue lines are the fixation angles of the eyes. In the bottom panel, the real and estimated target positions are in red and orange. The abscissa and the ordinate represent target depth and its projection.

few fluctuations occur. Finally, the *active vision* model that simultaneously implements depth estimation and target fixation achieves a level of performance that is almost on par with the model where the fixation is initialized at the correct position. Indeed, the only appreciable difference between the last two conditions is the slightly greater number of time steps in the *active vision* condition.

This pattern of results shows three main things. First, the hierarchical Active Inference model is able to solve the depth estimation problem, as evident from its perfect accuracy in the task. Second, the model is able not only to infer depth but also to simultaneously select the best way to sample its preferred stimuli (i.e., by fixating on the target). This is possible because, during a trial (as exemplified graphically in Figure 6), the active vision model obtains increasingly more accurate estimates of the depth, as the point of fixation approaches the target. Note that this pattern of results emerges because of the nonuniform resolution of the fovea. In fact, if the fovea resolution is assumed to be uniform (such as in camera models of artificial agents), the best accuracy is achieved by keeping the eyes parallel (Figure 7, *infer parallel* condition). In this case, fixating the target does not help but rather hinders and slows down depth estimation, which is probably caused by the increased effort that the agent needs to make in order to infer the reference frames of the eyes rotated in different directions. This has the consequence of further increasing the time needed to estimate the depth for the active vision model.

Intuitively, the better performance in the uniform condition is due to the lack of noise in the visual input. Although a more realistic scenario should consider noise also in this case, it is reasonable to assume that it would have a much smaller amplitude, due to a uniform distribution. Considering the nonuniform sensory distribution only, the better performance in the *infer vergence* condition relative to *active vision* could be due to the fact that in the former case the agent starts the inference process from a fixation on the correct 3D position. Thus, an active vision strategy in the *infer vergence* condition only needs to estimate the object's depth. If we compare these two scenarios, we can note that *active vision* performs almost optimally, similar to the *infer vergence* condition when the eye angles are set to the correct values for object fixation. However, the latter condition rarely occurs in a realistic setting and a more meaningful depth estimation comparison would be between *active vision* and the more general case in which the agent is fixating another object, or not fixating anything in particular – which we approximated with the *infer parallel* simulation.

Figure 7: Simulation results. Performance of the depth estimation task with nonuniform (top) and uniform (bottom) fovea resolution during inference with the eyes parallel and fixed (*infer parallel*), inference with the eyes fixating on the target (*infer vergence*), and simultaneous inference and target fixation (*active vision*). The accuracy (left panel) measures the number of trials in which the agent successfully predicts the 2D position of the target, the mean error (middle panel) measures the distance between the real and estimated target positions at the end of every trial, and the time (right panel) measures the number of steps needed to correctly estimate the target.

4 Discussion

We advanced a hierarchical Active Inference model for depth estimation and target fixation operating in the state-space of the projected camera planes. Our results show that depth estimation can be solved by inference – that is, by inverting a hierarchical generative model that predicts the projection of the eyes from a 2D belief over an object. Furthermore, our results show that active vision dynamics makes the inference particularly effective – and that fixating the target drastically improves task accuracy (see Figure 7). Crucially, the proposed model can be implemented in biologically plausible neural circuits for Predictive Coding [8, 9, 10], which only require local (top-down and bottom-up) message passing. From a technical perspective, our model shows that inference in any homogeneous transformation can be iteratively realized by combining generative models at different levels, each of which computes a specific transformation that can be, e.g., a roto-translation for kinematic inference [21], or a projection for computer vision.

This proposal has several elements of novelties compared to previous approaches to depth estimation. First, by focusing on inference and local message passing, our proposal departs from the trend of viewing cortical processing from a purely bottom-up perspective. The latter is common in neural network approaches, which start from the image of an object and gradually detect more and more complex features, eventually estimating its depth. Our proposal is also distinct from a direct approach that generates the depth of an object from a top-down perspective, e.g., with vergence cues. The role of vergence has been considered key in facilitating binocular fusion [5] and maximizing coding efficiency in a single environmental representation [26], but recent studies have dramatically reduced the importance of this mechanism in depth estimation. Binocular fusion might not be strictly necessary for this task [27] – 90% performance of depth estimation is attributed to diplopia [28] – and has never actually been tested as an absolute distance cue without eliminating all possible confounds [28]. Moreover, when fixating a target, there is always a vertical fixation disparity of monocular images, with no line precisely intersecting to form a vergence angle [29] – and it has been demonstrated that vergence does not correspond to the exact distance of the object being gazed at [30]. In keeping with this body of evidence, in our model the vergence belief does not play a critical role in depth estimation: it only operates along with a high-level belief over the 2D position of the object in order to predict the projections for the two eyes. These projections

are compared with the visual observations, and the resulting prediction errors that flow back through the hierarchy drive the update of both beliefs (about the eye angles and the 2D object position), in two possible ways: they might change (i) the estimated depth of the object, or (ii) the vergence-accommodation angles of the eyes, ultimately realizing a specific movement. In sum, depth estimation is not purely a top-down process, but rather it is realized through the inversion of a generative projective model and by averaging the information obtained through the two parallel pathways of the eyes. In conclusion, our model supports the direct (from disparity to both vergence and depth) rather than the indirect (from disparity to vergence, and then to depth) hypothesis of depth estimation [27]. Furthermore, this account is in line with the fact that small changes in vergence (*delta theta*) are the consequence and not a direct hint about depth estimation [28], and that reflex-like vergence mechanisms serve only to eliminate small vergence errors and not to actively transfer the gaze to new depth planes [31].

Second, an interesting consequence of this architecture is that – in contrast to standard neural networks – it permits imposing priors over the depth belief in order to drive and speed up the inferential process. Such priors may come through different sensory modalities or other visual cues, e.g., motion parallax or perspective, which we did not consider here. This is in line with the finding that vergence alone is unable to predict depth with ambiguous cues [5], suggesting that depth belief is constantly influenced by top-down mechanisms and higher-level cues, and not just directly from perception. Besides depth priors, using an Active Inference model has also the advantage that, if one wants to fixate on a target, simple attractors can be defined either at the eye angles belief or at the last projection level, each in their own domain. For example, imposing that the agent should perceive a projection of an object at the center of the camera plane results in the generation of a prediction error that ultimately moves the eyes toward that object – hence emphasizing the importance of active sensing strategies to enhance inference [32, 33].

However, the fact that there are two open pathways through which the prediction errors of the projections can flow – the eye angles and the absolute beliefs – may be problematic in some cases, e.g., during simultaneous depth estimation and target fixation. It is natural to think that depth estimation follows target fixation. In fact, top-down processing to verge on a target is generally not necessary: when an image is presented to a camera, the latter might move into this projected space directly, resulting in a simpler control [34, 35]. Then, depth could be computed directly from vergence cues. However, our model assumes that the eye angles generate the projections by first performing a roto-translation in the 2D space using the estimated depth, allowing further mechanisms for more efficient inference. Under this assumption, target fixation in the projected space is possible through a top-down process that is constantly biased by the high-level belief. Nonetheless, a direct vergence control not considered here could be implemented by additional connections between the belief over the 2D or projected points, and the angles belief.

With these considerations, it would therefore appear plausible that the two processes of depth estimation and target fixation run in parallel. But when this is the case, the prediction errors of the projections drive the two high-level beliefs independently toward a direction that minimizes the free energy, leading the agent to be stuck in an intermediate configuration with incorrect object depth and eye angles. One way to solve this issue that we pursue here consists of decomposing the task into cyclical phases of action and perception [24, 23]. During an action phase, the 2D belief is fixed and the agent can fixate on the predicted projections. Instead, during a perception phase, the agent can infer the 2D position of an object but is not allowed to move its sight. This implies that the prediction errors of the projections will alternatively flow in different directions (2D position and eye angles) one step at a time, which results in (i) the object being pulled toward the center of the camera planes and (ii) the estimated 2D position converging toward the correct one, as shown in Figure 6. Action-perception cycles have been studied in discrete time models of Active Inference; for example, cycles of saccades and visual sampling allow an agent to reduce the uncertainty over the environment (e.g., by rapidly oscillating between different points for recognizing an object) [36, 37]. Here, we show that action-perception cycles are also useful in continuous time models, such as the one used here, to ensure an effective minimization of the free energy, but also when an agent is required to reach an object with the end-effector while inferring the lengths of its limbs [23]. In summary, the two processes of recognizing a face and estimating an object's depth can be viewed as actively accumulating sensory evidence, although at different timescales. From a brain perspective, action-perception cycles have been often associated with hippocampal theta rhythms and cortical oscillations that might segment continuous experiences into discrete elements [38, 24]. From a more technical perspective, the cyclical scheme that we propose for action-perception cycles – which consists of keeping one aspect of the optimization objective fixed when updating the other – is commonly used in various optimization algorithms such as expectation-maximization [39], and a similar approach is used for learning and inference in Predictive Coding Networks [40, 41].

Third, our results show that active vision improves depth estimation. But if vergence does not provide a useful cue for depth, how is this possible? The answer comes from the nonuniform resolution of the fovea, which has far more receptors at the center of fixation than in the peripheral vision. It is supposed that this nonlinear resolution allows gathering sensory processing resources around the most relevant sources of information [42]. Under this assumption, the best performances are achieved when both eyes fixate on the object (as shown in Figure 7). As noted in [43], when stereo cameras have a nonsymmetrical vergence angle, the error is minimum when the projections of a point fall at the

center of the camera planes. Hence, vergence can effectively play a key role in depth estimation, besides providing a unified representation of the environment. This can be appreciated by considered that the *infer vergence* and *active vision* models are more accurate than the *infer parallel* model, in the nonuniform resolution condition. Rather, with a uniform resolution, the error is larger when the eyes converge to the target, because the focal angle of the pixels in the center is larger than in the periphery [43]. In addition to the increased error, the estimation seems to be further slowed down by the inference of the different reference frames due to vergence. Taken together, the consequence is that in the uniform resolution scenario the best estimation is achieved with fixed parallel eyes (see Figure 7), and active vision does not bring any advantage to the task. Since maintaining parallel eyes in the uniform condition results in a slightly higher accuracy, such simulations may be useful to understand in which cases verging on a target would improve the model performance. This might be helpful for future studies in bio-inspired robotics, especially when extending the proposed model to implement high-level mechanisms, such as: higher-level priors that result from the integration of cues from different sensory modalities; or attentional mechanisms that unify the visual sensations into a single experience.

The model presented in this study has some limitations that can be addressed in future studies. For example, we used a fixed focal length f during all the simulations. However, in a more realistic setting, the focal length might be considered as another DoF of the agent – and might be thus changed through suppression of proprioceptive prediction errors – speeding up the inferential process for objects at different distances. Furthermore, although the presented simulations only estimate the depth of a 2D point, it could be potentially extended to deal with 3D objects and account for vertical binocular disparity [44]. This would involve augmenting all the latent states with the new dimension and performing a sequence of two rotations as intermediate levels, before the eye projections are predicted. Then, the vergence-accommodation belief would also be extended with a new DoF, letting the agent able to fixate on 3D objects. Future studies might also investigate how the scaling factor of Equation 29 of the nonuniform resolution – but also more realistic nonuniform transformations – affect performance and help to model the human data (e.g., [25]). Also, it could be useful to adopt an off-center fovea on one of the two retinas and analyze the agent's behavior to bring the two foveas on the target. Finally, another interesting direction for future research would be combining the architecture proposed here with a more sophisticated kinematic Active Inference model [21], for example embodying a humanoid robot with multiple DoF [45, 46, 47, 48, 49, 47]. This would allow – in contrast to state-of-the-art models that provide the agent either with the 3D environment directly as a visual observation [20], or with a latent space reconstructed from a Variational Autoencoder [50] – inference of the 3D position of an object through the projections of the eyes, which can then be used for subsequent tasks, such as reaching and grasping.

5 Acknowledgments

This research received funding from the European Union's Horizon H2020-EIC-FETPROACT-2019 Programme for Research and Innovation under Grant Agreement 951910 to I.P.S., No. 945539 (Human Brain Project SGA3) to G.P. and No. 952215 (TAILOR) to G.P.; the European Research Council under the Grant Agreement No. 820213 (ThinkAhead) to G.P., the Italian Ministry for Research MIUR under Grant Agreement PRIN 2017KZNZLN to I.P.S., and the PNRR MUR projects PE0000013-FAIR and IR0000011–EBRAINS-Italy to G.P. The GEFORCE Quadro RTX6000 and Titan GPU cards used for this research were donated by the NVIDIA Corporation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Ning Qian. Binocular disparity and the perception of depth. *Neuron*, 18(3):359–368, 1997.

[2] Andrew J. Parker. Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience*, 8(5):379–391, 2007.

[3] Jean Baptiste Durand, Koen Nelissen, Olivier Joly, Claire Wardak, James T. Todd, J. Farley Norman, Peter Janssen, Wim Vanduffel, and Guy A. Orban. Anterior Regions of Monkey Parietal Cortex Process Visual 3D Shape. *Neuron*, 55(3):493–505, 2007.

[4] 3D shape perception from combined depth cues in human visual cortex. *Nature Neuroscience*, 8(6):820–827, 2005.

[5] D. A. Wismeijer, R. Van Ee, and C. J. Erkelens. Depth cues, rather than perceived depth, govern vergence. *Experimental Brain Research*, 184(1):61–70, 2008.

[6] Takuya Isomura, Thomas Parr, and Karl Friston. Bayesian filtering with multiple internal models: Toward a theory of social intelligence. *Neural Computation*, 31(12):2390–2431, 2019.

[7] Richard Langton Gregory. Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):279–296, 1968.

[8] Rajesh P.N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[9] Karl J Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1456):815–36, apr 2005.

[10] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. Active inference: the free energy principle in mind, brain, and behavior. 2022.

[11] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

[12] Giovanni Pezzulo, Francesco Rigoli, and Karl J Friston. Hierarchical active inference: a theory of motivated control. *Trends in cognitive sciences*, 22(4):294–306, 2018.

[13] Giovanni Pezzulo, Francesco Rigoli, and Karl Friston. Active inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology*, 134:17–35, 2015.

[14] Karl Friston, Jérémie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny. Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234, 2007.

[15] Karl J. Friston, Thomas Parr, and Bert de Vries. The graphical brain: Belief propagation and active inference. 1(4):381–414, 2017.

[16] Rick A. Adams, Stewart Shipp, and Karl J. Friston. Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218(3):611–643, 2013.

[17] Thomas Parr and Karl J. Friston. Active inference and the anatomy of oculomotion. *Neuropsychologia*, 111(January):334–343, 2018.

[18] Rick A. Adams, Eduardo Aponte, Louise Marshall, and Karl J. Friston. Active inference and oculomotor pursuit: The dynamic causal modelling of eye movements. *Journal of Neuroscience Methods*, 242:1–14, 2015.

[19] Pablo Lanillos and Gordon Cheng. Adaptive Robot Body Learning and Estimation Through Predictive Coding. *IEEE International Conference on Intelligent Robots and Systems*, pages 4083–4090, 2018.

[20] Léo Pio-Lopez, Ange Nizard, Karl Friston, and Giovanni Pezzulo. Active inference and robot control: A case study. *Journal of the Royal Society Interface*, 13(122), 2016.

[21] Matteo Priorelli, Giovanni Pezzulo, and Ivilin Peev Stoianov. Deep kinematic inference affords efficient and scalable control of bodily movements. *bioRxiv*, pages 1–33, 2023.

[22] Matteo Priorelli and Ivilin Peev Stoianov. Flexible intentions: An active inference theory. *Frontiers in Computational Neuroscience*, 17:1128694:1–27, Mar. 2023.

[23] M. Priorelli and I. P. Stoianov. Efficient motor learning through action-perception cycles in deep kinematic inference. *4th International Workshop on Active Inference*, 2023.

[24] Rufin VanRullen. Perceptual cycles. *Trends in cognitive sciences*, 20(10):723–735, 2016.

[25] Ann E. Elsner, Toco Y.P. Chui, Lei Feng, Hong Xin Song, Joel A. Papay, and Stephen A. Burns. Distribution differences of macular cones measured by AOSLO: Variation in slope from fovea to periphery more pronounced than differences in total cones. *Vision Research*, 132:62–68, 2017.

[26] Qingpeng Zhu, Jochen Triesch, and Bertram E. Shi. Integration of Vergence, Cyclovergence, and Saccades through Active Efficient Coding. *ICDL-EpiRob 2020 - 10th IEEE International Conference on Development and Learning and Epigenetic Robotics*, 2020.

[27] Arthur J. Lugtigheid, Laurie M. Wilcox, Robert S. Allison, and Ian P. Howard. Vergence eye movements are not essential for stereoscopic depth. *Proceedings of the Royal Society B: Biological Sciences*, 281(1776):1–6, 2013.

[28] Paul Linton. Does vision extract absolute distance from vergence? *Attention, Perception, and Psychophysics*, 82(6):3176–3195, 2020.

[29] Alexander D. Logvinenko, Julie Epelboim, and Robert M. Steinman. The role of vergence in the perception of distance: A fair test of bishop Berkeley's claim. *Spatial Vision*, 15(1):77–97, 2002.

[30] Wolfgang Jaschinski. Fixation disparity and accommodation as a function of viewing distance and prism load. *Ophthalmic and Physiological Optics*, 17(4):324–339, 1997.

[31] G S Masson, C Busettini, and F A Miles. Vergence eye movements in response to binocular disparity without depth perception. *Nature*, pages 283–286, 1997.

[32] Francesco Mannella, Federico Maggiore, Manuel Baltieri, and Giovanni Pezzulo. Active inference through whiskers. *Neural Networks*, 144:428–437, 2021.

[33] Thomas Parr and Karl J Friston. Attention or salience? *Current Opinion in Psychology*, 29:1–5, 2019.

[34] R. Manzotti, A. Gasteratos, G. Metta, and G. Sandini. Disparity estimation on log-polar images and vergence control. *Computer Vision and Image Understanding*, 83(2):97–117, 2001.

[35] Agostino Gibaldi, Mauricio Vanegas, Andrea Canessa, and Silvio P. Sabatini. A portable bio-inspired architecture for efficient robotic vergence control. *International Journal of Computer Vision*, 121(2):281–302, Jan 2017.

[36] Karl Friston, Rick A. Adams, Laurent Perrinet, and Michael Breakspear. Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3(MAY):1–20, 2012.

[37] Ajith Anil Meera, Filip Novicky, Thomas Parr, Karl Friston, Pablo Lanillos, and Noor Sajid. Reclaiming saliency: Rhythmic precision-modulated action and perception. *Frontiers in Neurorobotics*, 16:1–23, 2022.

[38] John E Lisman and Ole Jensen. The theta-gamma neural code. *Neuron*, 77(6):1002–1016, 2013.

[39] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[40] James C.R. Whittington and Rafal Bogacz. Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, 23(3):235–250, 2019.

[41] Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Predictive Coding Approximates Backprop Along Arbitrary Computation Graphs. *Neural Computation*, 34(6):1329–1368, 2022.

[42] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[43] Hossein Sahabi and Anup Basu. Analysis of Error in Depth Perception with Vergence and Spatially Varying Sensing. *Computer Vision and Image Understanding*, 63(3):447–461, 1996.

[44] Jenny C A Read, Graeme P Phillipson, and Andrew Glennerster. Latitude and longitude vertical disparities. *J Vis*, 9(13):11.1–37, December 2009.

[45] Pablo Lanillos, Gordon Cheng, et al. Robot self/other distinction: active inference meets neural networks learning in a mirror. *arXiv*, 2020.

[46] Ahmadreza Ahmadi and Jun Tani. A novel predictive-coding-inspired variational rnn model for online prediction and recognition. *Neural computation*, 31(11):2025–2074, 2019.

[47] Tadahiro Taniguchi, Shingo Murata, Masahiro Suzuki, Dimitri Ognibene, Pablo Lanillos, Emre Ugur, Lorenzo Jamone, Tomoaki Nakamura, Alejandra Ciria, Bruno Lara, et al. World models and predictive coding for cognitive and developmental robotics: frontiers and challenges. *Advanced Robotics*, pages 1–27, 2023.

[48] Ozan Çatal, Tim Verbelen, Toon Van de Maele, Bart Dhoedt, and Adam Safron. Robot navigation as hierarchical active inference. *Neural Networks*, 142:192–204, 2021.

[49] Elias De Coninck, Tim Verbelen, Pieter Van Molle, Pieter Simoens, and Bart Dhoedt. Learning robots to grasp by demonstration. *Robotics and Autonomous Systems*, 127:103474, 2020.

[50] Thomas Rood, Marcel van Gerven, and Pablo Lanillos. A deep active inference model of the rubber-hand illusion. In *Active Inference*, pages 84–91, Cham, 2020. Springer International Publishing.