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 a different distance, may project to the same 2D 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 might be implemented in biologically plausible ways. State-of-the-art approaches to depth estimation in machine learning based on deep neural networks implicitly describe the brain as a (hierarchical) feature detector. Instead, we propose a biologically plausible 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 predicts the 2D projection of the eyes from a 3D belief over an object. The generative model updates its beliefs about depth by averaging the prediction errors coming from simultaneous 2D projections. Model inversion requires a series of biologically plausible, homogeneous transformations based on predictive coding principles. Under the plausible assumption of a nonuniform fovea resolution, depth estimation favours an active vision strategy that fixates the object with the eyes (by setting the object position in the camera plane as an attractor state), which renders the depth belief more precise. Interestingly, in our approach, active vision is not realized by first fixating on a target and then estimating the depth, but by combining the two processes together, through cycles of action and perception, with a similar mechanism of the saccades in higher-level processes of evidence accumulation for 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 this information travels 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 represen-

tations of objects are constructed. This view has inspired the development of Convolutional Neural Networks, which have led to remarkable results in object recognition tasks. 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 [6]. In recent years a different perspective has emerged, based on Predictive Coding theories, which 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 [7, 8]. Furthermore, vision is increasingly considered to be an active process that constantly tries to reduce the uncertainty of what will happen next. In other words, the biases we perceive are actually hints to better minimize the errors between our sensations and our predictions [9].

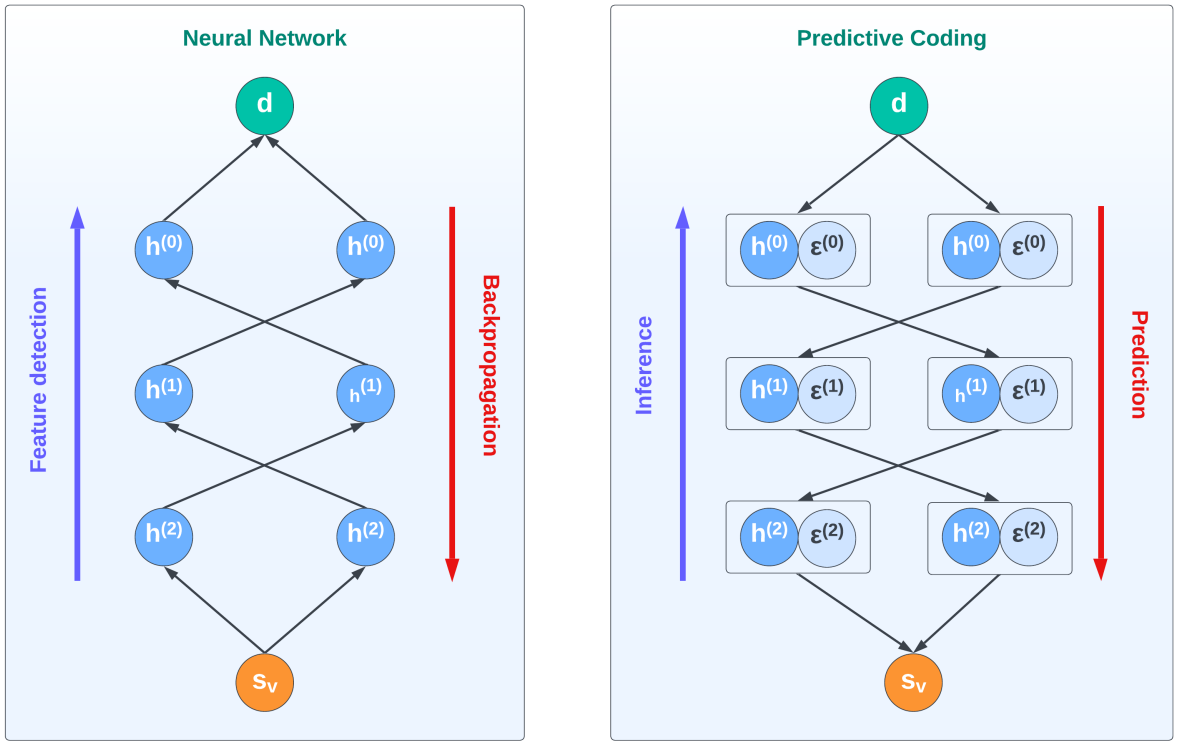

Figure 1: Representation of information processing in neural networks (left) and predictive coding (right). In a neural network, the visual observation $\mathbf{s}_v$ travels through the cortical hierarchy detecting increasingly complex features and eventually estimating, e.g., the depth of an object, and the descending signals are considered feedback signals that convey backpropagation errors. In contrast, in a predictive coding network the top-down signals generate, starting from a high-level belief, a visual prediction that is compared with the observation; the sensory prediction error then travels back the hierarchy eventually inferring the correct belief (for details, see below).

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 on the basis of two images of the object, through a process of prediction error minimization and active oculomotor behavior. In our model, depth estimation consists in the inference of top-down projective predictions from a high-level representation of the object, not as a bottom-up process that detects disparities in the images of the two eyes. In other words, as shown in Figure 1, the estimation of the depth of an object should naturally arise by inverting a visual generative model and the corresponding prediction errors that flow up the cortical hierarchy, in contrast with a direct process occurring with neural networks.

2 Methods

2.1 Oculomotor behavior and transformations between reference frames

Classical predictive coding models are passive, in the sense that the model cannot select its visual stimuli [7]. 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 [10, 11]. 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 [12, 13]. 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.

We previously showed that 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 [14]. The extrinsic configuration of the motor plant is computed hierarchically, as shown in Figure 2. For the relatively simple kinematic control tasks targeted in [14], these computations only required two simple transformations between reference frames: translations and rotations. However, that hierarchical kinematic model can be easily extended to afford more complex tasks, which require other transformations.

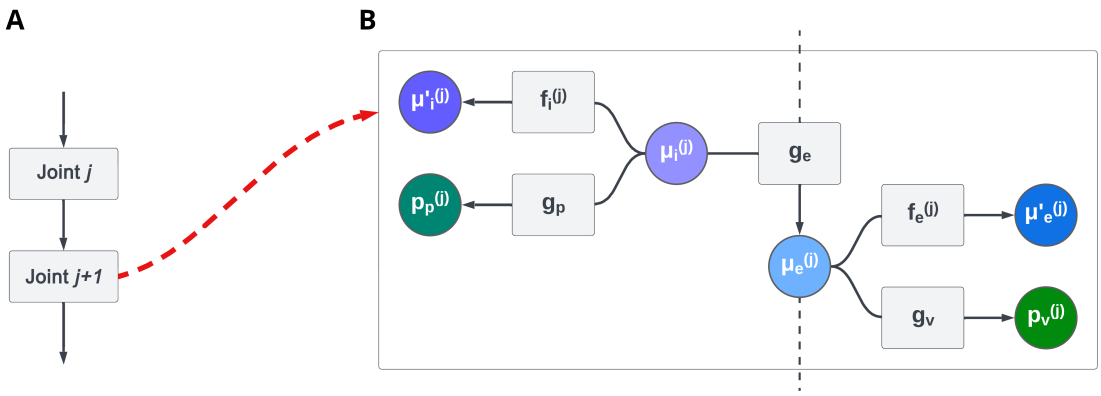

Figure 2: (**A**) An example of a portion of a kinematic plant. (**B**) Factor graph of a single level of the hierarchical kinematic model where each block has intrinsic $\boldsymbol{\mu}_i$ and extrinsic $\boldsymbol{\mu}_e$ beliefs. Note that the extrinsic belief acts as a prior for the layer below. See [14] 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 several different simpler steps where the homogeneous coordinates are multiplied once at a time through the chain rule, allowing more efficient computations. For example, affine transformations (which preserve parallel lines) have the following form:

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

where the last row ensures that every point always maps to the same plane. By the chain rule, a point in the plane can be rotated and translated by multiplication of the corresponding transformations:

$$\tilde{\boldsymbol{p}}^f = \begin{bmatrix} p_x^f \\ p_y^f \\ 1 \end{bmatrix} = \begin{bmatrix} c_\theta & -s_\theta & 0 \\ s_\theta & c_\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & l_x \\ 0 & 1 & l_y \\ 0 & 0 & 1 \end{bmatrix} \cdot \tilde{\boldsymbol{p}}^0 = \begin{bmatrix} c_\theta(p_x^0 + l_x) - s_\theta(p_y^0 + l_y) \\ s_\theta(p_x^0 + l_x) + c_\theta(p_y^0 + l_y) \\ 1 \end{bmatrix} \tag{2}$$

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:

$$\tilde{\boldsymbol{p}}^f = \begin{bmatrix} p_x^f \\ p_y^f \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ z_x & z_y & 1 \end{bmatrix} \tilde{\boldsymbol{p}}^0 = \begin{bmatrix} p_x^0 \\ p_y^0 \\ z_x p_x^0 + z_y p_y^0 + 1 \end{bmatrix} \tag{3}$$

so that the new point does not map on the same plane $z = 1$, and we thus have to divide it by the last element in order to obtain the Cartesian coordinates:

$$\begin{bmatrix} p_x^0 \\ p_y^0 \\ z_x p_x^0 + z_y p_y^0 + 1 \end{bmatrix} \equiv \begin{bmatrix} \frac{p_x^0}{z_x p_x^0 + z_y p_y^0 + 1} \\ \frac{p_y^0}{z_x p_x^0 + z_y p_y^0 + 1} \\ 1 \end{bmatrix} \tag{4}$$

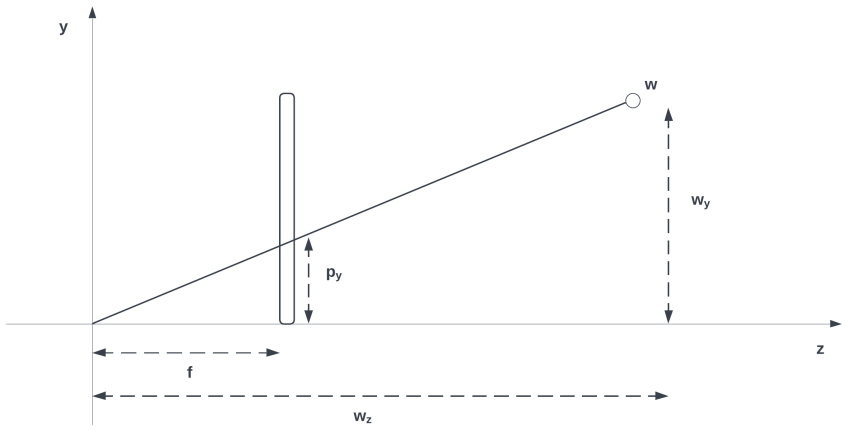

Figure 3: Projection of a 3D point in the camera plane (only two dimensions are shown).

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. In fact, if we have a 3D point $\tilde{\boldsymbol{w}}$ in homogeneous coordinates, we can obtain the corresponding 2D point $\boldsymbol{p}'$ projected in the camera plane by computing the product of two matrices:

$$\boldsymbol{p}' = \begin{bmatrix} p'_x \\ p'_y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & f & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}^T & -\boldsymbol{R}^T \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \tilde{\boldsymbol{w}} \tag{5}$$

The first one (called the *camera matrix*) scales and converts the point to 2D – where the image plane is defined by the focal length f – whereas the latter encodes where the camera is located and oriented (the *extrinsic parameters*). As before, the point $\boldsymbol{p}'$ is only up to a scale factor and to find the correct point that maps in the camera plane we just divide the camera coordinates by the depth coordinate p'_z (as shown in Figure 3):

$$p_x = \frac{p'_x}{p'_z} \qquad\qquad p_y = \frac{p'_y}{p'_z} \tag{6}$$

Keeping this in mind, we extend the deep kinematic model of [14] by considering that each hierarchical 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).

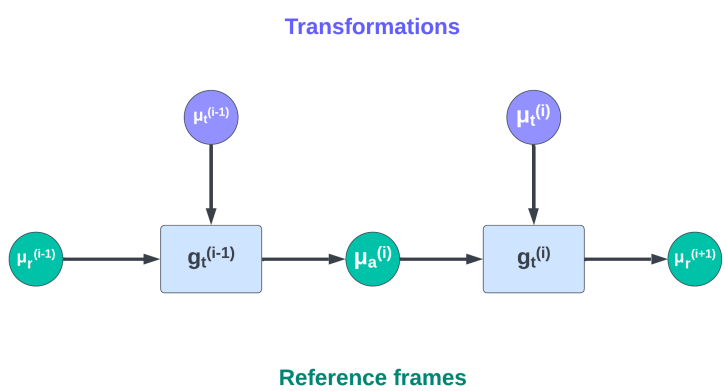

Figure 4: Representation of the hierarchical relationships of a generalized model with homogeneous transformations.

This belief then generates a homogeneous transformation relative to that Degree of Freedom (DoF), which can be implicit in the generative model. This matrix is then multiplied by a belief in a particular reference frame (as exemplified in Figure 4):

$$\boldsymbol{\mu}_r^{(i+1)} = \boldsymbol{g}_t^{(i)}(\boldsymbol{\mu}_t^{(i)}, \boldsymbol{\mu}_r^{(i)}) = \boldsymbol{T}^{(i)}(\boldsymbol{\mu}_t^{(i)}) \cdot \boldsymbol{\mu}_r^{(i)} \tag{7}$$

This equation leads to simple gradient computations through the prediction error $\varepsilon_r^{(i+1)}$:

$$\begin{aligned} \frac{\partial \boldsymbol{g}_t^{(i)}}{\partial \boldsymbol{\mu}_r^{(i)}}^T \boldsymbol{\varepsilon}_r^{(i+1)} &= \boldsymbol{T}^{(i)T} \cdot \boldsymbol{\varepsilon}_r^{(i+1)} \\ \frac{\partial \boldsymbol{g}_t^{(i)}}{\partial \boldsymbol{\mu}_t^{(i)}}^T \boldsymbol{\varepsilon}_r^{(i+1)} &= \frac{\partial \boldsymbol{T}^{(i)}}{\partial \boldsymbol{\mu}_t^{(i)}} \odot [\boldsymbol{\varepsilon}_r^{(i+1)} \cdot \boldsymbol{\mu}_r^{(i)T}] \end{aligned} \tag{8}$$

which are needed to iteratively update the two beliefs.

2.2 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. We construct the generative model hierarchically, starting from a belief $\boldsymbol{\mu}_a = (\mu_{a,x}, \mu_{a,y})$ on top containing information about the absolute 2D position of an object. This belief generates two specular predictions $\boldsymbol{p}_r^{(i)}$ in parallel pathways that take the absolute eye angle as additional input – which is computed by a combination of a vergence-accommodation belief $\boldsymbol{\mu}_\theta = (\theta_a, \theta_v)$ – and transform the absolute coordinates of the object into the two relative reference frames:

$$\boldsymbol{p}_r^{(i)} = \boldsymbol{g}_a^{(i)}(\boldsymbol{\mu}_a, \boldsymbol{\mu}_\theta) = \boldsymbol{T}^{(i)}(\boldsymbol{\mu}_\theta) \cdot \boldsymbol{\mu}_a \tag{9}$$

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

$$\boldsymbol{T}^{(i)}(\boldsymbol{\mu}_\theta) = \begin{bmatrix} c_\theta^{(i)} & s_\theta^{(i)} & -l^{(i)} s_\theta^{(i)} \\ -s_\theta^{(i)} & c_\theta^{(i)} & -l^{(i)} c_\theta^{(i)} \\ 0 & 0 & 1 \end{bmatrix} \tag{10}$$

and the absolute eye angles are:

$$\theta^{(0)} = \theta_a - \theta_v \qquad\qquad \theta^{(1)} = \theta_a + \theta_v \tag{11}$$

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

$$\boldsymbol{p}_c^{(i)} = \boldsymbol{g}_r(\boldsymbol{\mu}_r^{(i)}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & f & 0 \end{bmatrix} \cdot \frac{\boldsymbol{\mu}_r^{(i)}}{\mu_{r,x}^{(i)}} \tag{12}$$

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 predictions for the two eyes, proprioceptive predictions directly encode angles in the vergence-accommodation system, which is used for action execution [9]. The last level instead generates a visual image of the point – which in the simulations we approximated with the projected position directly.

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

$$\begin{aligned} \dot{\boldsymbol{\mu}}_a &= \sum_i \partial_a \boldsymbol{g}_a^T \boldsymbol{\varepsilon}_r^{(i)} + \partial \boldsymbol{f}_a^T \boldsymbol{\varepsilon}_{\mu,a} \\ \dot{\boldsymbol{\varepsilon}}_r^{(i)} &= \boldsymbol{\mu}_r^{(i)} - \boldsymbol{g}_a^{(i)}(\boldsymbol{\mu}_a, \boldsymbol{\mu}_\theta) - \boldsymbol{\varepsilon}_r^{(i)} / \pi_r^{(i)} \\ \dot{\boldsymbol{\varepsilon}}_{\mu,a} &= \boldsymbol{\mu}_a' - \boldsymbol{f}_a(\boldsymbol{\mu}_a) - \boldsymbol{\varepsilon}_{\mu,a} / \gamma_a \end{aligned} \tag{13}$$

where π_r and $\boldsymbol{\varepsilon}_r$ 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 $\varepsilon_r^{(i)}$ 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:

$$\begin{aligned} \dot{\boldsymbol{\mu}}_\theta &= \sum_i \partial_\theta \boldsymbol{g}_a^T \boldsymbol{\varepsilon}_r^{(i)} + \partial \boldsymbol{g}_p^T \boldsymbol{\varepsilon}_p + \partial \boldsymbol{f}_\theta^T \boldsymbol{\varepsilon}_{\mu,\theta} \\ \dot{\boldsymbol{\varepsilon}}_p &= \boldsymbol{s}_p - \boldsymbol{g}_p(\boldsymbol{\mu}_\theta) - \boldsymbol{\varepsilon}_p / \pi_p \\ \dot{\boldsymbol{\varepsilon}}_{\mu,\theta} &= \boldsymbol{\mu}_\theta' - \boldsymbol{f}_\theta(\boldsymbol{\mu}_\theta) - \boldsymbol{\varepsilon}_{\mu,\theta} / \gamma_\theta \end{aligned} \tag{14}$$

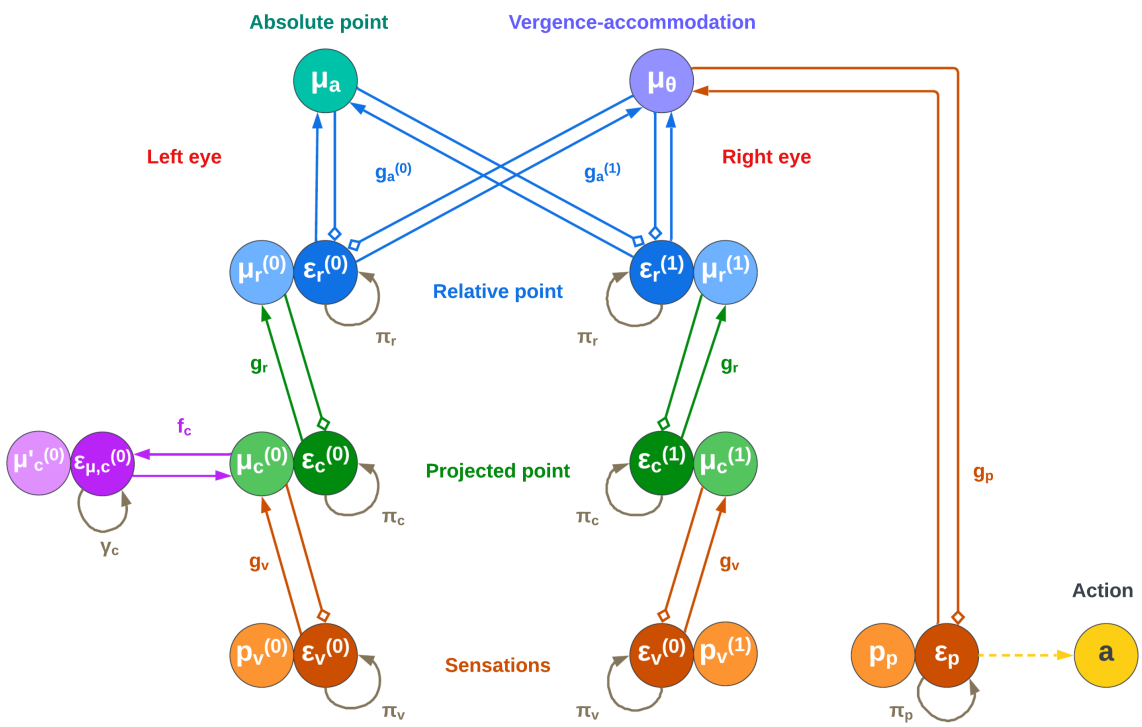

Figure 5: Neural-level implementation of a hierarchical generative model estimating the depth of a point in Active Inference. Small squares indicate inhibitory connections. For clarity, only the dynamics function of the belief over the projected point is shown. Differently from a neural network, depth is estimated by first generating a visual projection from a point in absolute coordinates. The predictions are then compared with the visual observations, and the prediction errors climb up the hierarchy, eventually driving the beliefs at the top toward the correct values. Note that action is realized by suppressing the proprioceptive prediction error, while goal-directed movement (e.g., target fixation) is achieved by defining an appropriate dynamics function at the last level.

where π_p, $\boldsymbol{\varepsilon}_p$, and $\boldsymbol{s}_p$ are the proprioceptive precision, prediction error, and observation.

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 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 modify 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:

$$
\begin{aligned}
\dot{\boldsymbol{\mu}}_c^{(i)} &= -\boldsymbol{\varepsilon}_c^{(i)} + \partial \boldsymbol{g}_v^T \boldsymbol{\varepsilon}_v^{(i)} + \partial \boldsymbol{f}_c^{(i)T} \boldsymbol{\varepsilon}_{\mu,c}^{(i)} \\
\dot{\boldsymbol{\varepsilon}}_c^{(i)} &= \boldsymbol{\mu}_c^{(i)} - \boldsymbol{g}_r(\boldsymbol{\mu}_r^{(i)}) - \boldsymbol{\varepsilon}_c^{(i)}/\pi_c^{(i)} \\
\dot{\boldsymbol{\varepsilon}}_v^{(i)} &= \boldsymbol{s}_v^{(i)} - \boldsymbol{g}_v(\boldsymbol{\mu}_c^{(i)}) - \boldsymbol{\varepsilon}_v^{(i)}/\pi_v^{(i)} \\
\dot{\boldsymbol{\varepsilon}}_{\mu,c}^{(i)} &= \boldsymbol{\mu}_c^{(i)\prime} - \boldsymbol{f}_c^{(i)}(\boldsymbol{\mu}_c^{(i)}) - \boldsymbol{\varepsilon}_{\mu,c}^{(i)}/\gamma_c^{(i)}
\end{aligned}
\tag{15}
$$

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 the *projective prediction error* $\boldsymbol{\varepsilon}_c^{(i)}$, and a visual likelihood from $\boldsymbol{\varepsilon}_v^{(i)}$.

2.3 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 [15], such that the projected position will be at the center of the camera planes:

$$\begin{aligned} \boldsymbol{e}_c^{(i)} &= \boldsymbol{\mu}_c^{(i)} - \mathbf{0} \\ \boldsymbol{\mu}_c^{(i)\prime} &= \boldsymbol{f}_c^{(i)}(\boldsymbol{\mu}_c^{(i)}) = \lambda \boldsymbol{e}_c^{(i)} \end{aligned} \tag{16}$$

where λ is an attractor gain.

This dynamics function makes the agent think that the projected point will be pulled toward the center, with a velocity proportional to the distance $\boldsymbol{e}_c^{(i)}$. 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.

This abnormal behavior can be avoided if we decompose the task into cyclical phases of action and perception [16]. 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 eye angles, resulting in the eyes moving toward the current depth belief. During a perception phase, action is blocked (e.g., either by setting to zero the attractor gain or the proprioceptive precision), 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.

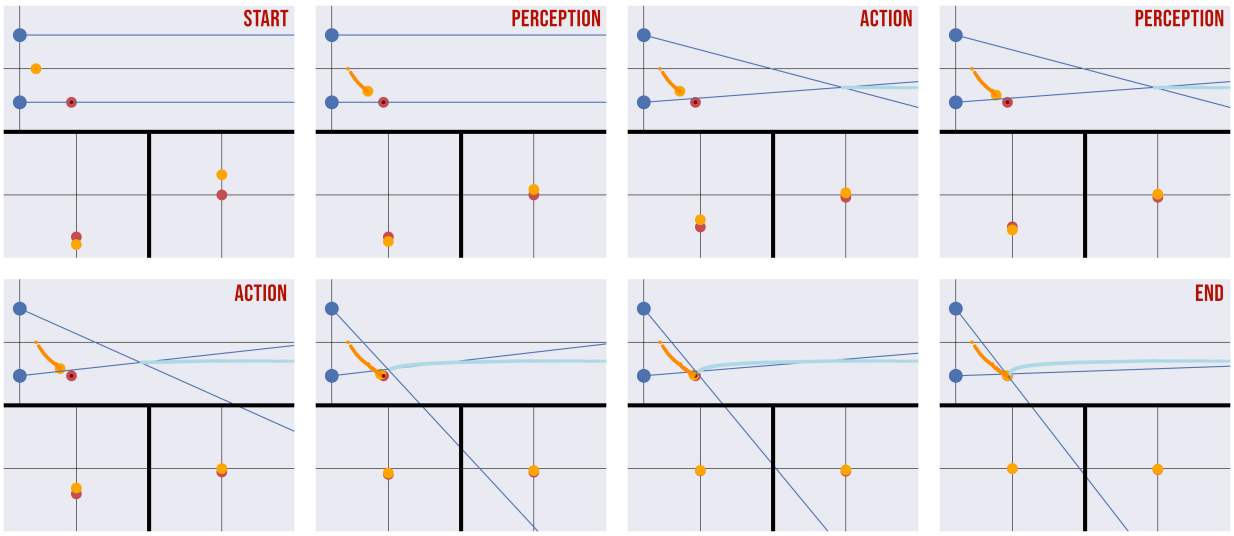

Figure 6: Sequence of time frames of a depth estimation task with simultaneous target fixation. The agent uses alternating action-perception phases to avoid being stuck during the minimization process. The eyes are represented in blue, and the real and estimated target position in red and orange. The fixation trajectory (when vergence occurs) is represented in cyan. Each frame is composed of three images: a view of the overall task (top), and the projection of the target to the two camera planes of the eyes (bottom left and bottom right).

3 Results

We test the model introduced in the Methods section in a depth estimation task that consists in inferring the 2D position of an object. Note that in this simulation, the fovea of the simulated "eye" has a nonuniform resolution and represents the object with greater accuracy when it is near the point of fixation. This reflects the fact that the fovea has far more receptors at the center of fixation than in the peripheral vision. Specifically, in our implementation, the variability of the Gaussian error in the visual observations exponentially increases with the distance between the point of fixation and the real target position.

We compare 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 shown in Figure 6).

Figure 7 shows the results of our simulations, which comprise an accuracy measure (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, center 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*) results in a very low accuracy, especially when the target position is far from the point of fixation. This is to be expected, since the fovea has low resolution at the periphery. Rather, if the angle of the eyes is set to fixate on the target, the accuracy is much higher and few fluctuations occur (*infer vergence*). Finally, the active vision model that simultaneously implements depth estimation and target fixation achieves a level of performance that is almost at pair with the model where the eyes are initialized at the correct position.

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 position at the end of every trial, and the time (right panel) measures the number of steps needed to correctly estimate the target.

This pattern of results shows two main things. First, the hierarchical active inference model endowed with a biologically realistic message passing scheme based on Predictive Coding, is able to solve 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 simulation

trial (shown graphically in Figure 6), the active vision model is able to obtain increasingly more precise estimates about 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 (see Figure 7). In this case, fixating the target does not help but rather hinder and slows down depth estimation, which is probably caused by the increased effort that the agent has to make 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.

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 2D projection of the eyes from a 3D 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 [7, 8, 9], which only require local (top-down and bottom-up) message passing. From a technical perspective, our model shows that 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 [14], 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 the combination of top-down and bottom-up message passing – our proposal departs from the trend of viewing cortical processing from a purely bottom-up perspective. This bottom-up perspective 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 [17], but recent studies have dramatically reduced the importance of this mechanism in depth estimation. Binocular fusion might not be strictly necessary for depth estimation [18] – 90% performance of depth estimation is attributed to diplopia [19] – and has never actually been tested as an absolute distance cue without eliminating all possible confounds [19]. Moreover, when fixating a target, there is always a vertical disparity of monocular images, with no line precisely intersecting to form a vergence angle [20] – and it has been demonstrated that vergence does not correspond to the exact distance of the object being gazed at [21]. 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 3D 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 3D 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 the 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 [18]. 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 [19], and that reflex-like vergence mechanisms serve only to eliminate small vergence errors and not to actively transfer the gaze to new depth planes [22].

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 inference 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 the latter 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 a prediction error that ultimately moves the eyes toward that object – hence emphasizing the importance of active sensing strategies to enhance inference [23, 24].

However, the fact that there are two open pathways through which the projective prediction errors can flow – the eye angles and the 3D beliefs – may be problematic in some cases, e.g., during simultaneous depth estimation and target fixation. While it seems natural to think that depth estimation follows target fixation, this would mean that when a 2D

image is presented to an eye, the latter might move into this projected space directly. However, our model assumes that the eye angles generate the projections by first performing a roto-translation in the 3D space using the estimated depth, so that target fixation is constantly biased by the high-level belief. 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 projective prediction errors 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 in decomposing the task into cyclical phases of action and perception [16]. During an action phase, the 3D belief is fixed and the agent can fixate the predicted projections. Instead, during a perception phase, the agent can infer the 3D position of an object but is not allowed to move its sight. This implies that the projective prediction errors will alternatively flow in different directions (3D 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 3D 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) [25, 26]. Here, we show that action-perception cycles are also useful in continuous time models, such as the one used in this article, to ensure an effective minimization of the free energy. Hence, both processes – recognizing a face or estimating the depth of an object – can be viewed as actively accumulating sensory evidence, although at different timescales. From a brain perspective, action-perception cycles have been often associated to hippocampal theta rhythms and cortical oscillations that might segment continuous experiences into discrete elements [27, 16]. From a more technical perspective, the cyclical scheme that we propose for action-perception cycles – which consists in keeping one aspect of the optimization objective fixed when updating the other – is commonly used in various optimization algorithms, such as expectation-maximization [28].

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 [29]. Under this assumption, the best performances are achieved when both eyes fixate on the object (as shown in Figure 7). As noted in [30], when stereo cameras have a nonsimmetrical 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. 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. In this scenario, the best estimation is achieved with fixed parallel eyes (see Figure 7), and active vision does not bring any advantage to the task.

The presented model 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 by 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. Another interesting direction for future research would be combining the architecture proposed here with a more sophisticated kinematic Active Inference model [14], for example embodying a humanoid robot with multiple DoF [31, 32, 33, 34, 35]. 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 [13], or with a latent space reconstructed from a Variational Autoencoder [36] – inference of the 3D position of an object through the projections of the eyes, which can then be used this belief 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.S., No. 945539 (Human Brain Project SGA3) to GP and No. 952215 (TAILOR) to GP; the European Research Council under the Grant Agreement No. 820213 (ThinkAhead) to GP, the Italian Ministry for Research MIUR under Grant Agreement PRIN 2017KZNZLN to I.S., and the PNRR MUR project PE0000013-FAIR to GP. 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] Richard Langton Gregory. Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):279–296, 1968.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] 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.

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

[20] 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.

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

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

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

[24] Thomas Parr and Karl J Friston. Attention or salience? *Current opinion in psychology*, 29:1–5, 2019.

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] Pablo Lanillos, Jordi Pages, and Gordon Cheng. Robot self/other distinction: active inference meets neural networks learning in a mirror. (Ecai), 2020.

[32] 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.

[33] 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. *arXiv preprint arXiv:2301.05832*, 2023.

[34] 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.

[35] 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.

[36] Thomas Rood, Marcel van Gerven, and Pablo Lanillos. A deep active inference model of the rubber-hand illusion. 2020.