Modulating Multi-Modal Integration in a Robot Forward Model for Sensory Enhancement and Self-Perception

Guido Schillaci, Alejandra Ciria, Egidio Falotico, Bruno Lara, Cecilia Laschi

Modulating Multi-Modal Integration in a Robot Forward Model for Sensory Enhancement and Self-Perception

Guido Schillaci, Alejandra Ciria, Egidio Falotico, Bruno Lara, Cecilia Laschi

***Abstract*—This work investigates how different modalities (i.e., visual, proprioceptive, motor) can be optimally integrated by a humanoid robot during a visual prediction task. A multi-modal forward model inspired on a work from different authors (Shim and colleagues, [31]) is adopted for generating visual predictions, given the motor activity and the context the robot is situated in.**

We extend the application of this tool by exploiting its optimal integration and predictive capabilities in sensory attenuation processes. According to the predictive brain hypothesis, our brains make sense of the world by anticipating sensory input and by enhancing or, on the contrary, filtering out information according to our expectations, motivations, desires and current contexts and tasks.

We develop a series of robotic studies in which we focus on the role of sensory attenuation processes in cognitive development. In particular, we show how attenuating predicted visual information may enhance the perceptual capabilities of a humanoid robot in an object detection task.

Moreover, we analyse the dynamics of the model prediction and its prediction error during the robot movements. In line with similar studies, our experiments indicate the mismatch between visual predictions and observations as a computational candidate for the study of self-perception and self-other distinction in artificial systems.

Finally, the capability of the model to re-modulate its multi-modal integration weights under dynamical environmental conditions is tested. This work analyses the dynamic modulation of multi-modal integration, proposing this to be also an essential prerequisite for the development of subjective experience in artificial systems.

***Index Terms*—Internal models, multi-modal integration, perceptual optimisation, sensory enhancement, self-perception**

I. Introduction

Prediction has been considered as having a fundamental role in human cognition and in perception, at least since Helmholzt proposal of unconscious inferences [16]. In cognitive robotics, prediction and prediction error are seen traditionally as the key tool for learning and model updating. However, the effects of prediction error on perception has not been widely modeled [20]. Few examples link prediction error directly with perception [5, 19, 30, 9].

GS, EF and CL are affiliated with the BioRobotics Institute and the Dept. of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, Italy.

AC is affiliated with Facultad de Psicologia, Universidad Nacional Autónoma de México, México City, Mexico.

BL is affiliated with Centro de Investigación en Ciencias, Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico

CL is affiliated with Dept. of Mechanical Engineering, National University of Singapore, Singapore

This work presents a multi-modal forward model for a simulated iCub humanoid robot.This model allows predicting and attenuating visually perceived movements, either generated by the robot or produced by other objects in the environment. Moreover, sensory attenuation processes are implemented to enable the robot to maintain the sight of task-relevant objects that could otherwise be occluded by other moving entities. In fact, predictive processes can enable a low-cost enhancement of the perceptual capabilities of the agent. Previous studies [5, 19] are extended by including external moving objects during training and testing, and by analysing the impact of these events on sensory attenuation. In addition, the proposed model can modulate the integration of different input modalities (i.e., visual, proprioceptive, and motor command). Inspired by the work of Shim et al. [31] and Patel et al. [23], the proposed model modulates sensory integration by a set of values that dynamically weight the importance of the different modalities for their contribution to visual prediction. The impact of different modulations in the predictive process during an object detection task is analysed using a mechanism for extraction and manual modulation of multi-modal integration weights.

The paper is organized as follows. Sections I-A and I-B frame out this work within the research on sense of agency in human and robots, and on perceptual enhancement in robots. Section II introduces the proposed computational model and the experiments. Results are discussed in section III. Finally, section IV draws the conclusions and outlines venues for further research.

A. Sense of agency and multi-modal integration

During their ontogenetic development, starting in the uterus, foetuses acquire their own body schema, fundamental for their efficient action planning and interaction with the environment [35, 25]. These schemas have been proposed to be composed of associations of different sensory and motor regions in the brain. In the scientific literature of the last decades, inverse and forward models have emerged as the standard computational candidate to encode them. In the literature on sense of agency, the forward model is also known as the *comparator* model [15].Theories claim that, when a motor command is executed, an efference copy is generated to predict the consequences of such action on what is sensed by the organism. The reafference feedback allows the comparison between the predicted sensory consequences and the actual sensory consequences of an action. This comparison provides

the capability to distinguish sensory changes consequence of voluntary actions from those generated by external causes. The comparator model theory claims that this comparison is the necessary and sufficient condition for the sense of agency. Under this theory, the resulting magnitude of the prediction error is directly linked to the sense of agency. Prediction error is also thought to be behind sensory attenuation and enhancement depending on the context and the task at hand [17, 8, 12]. Furthermore, sensory attenuation has been argued by scholars to be one of the markers of pre-reflective subjective experience [4, 29, 30, 24, 13, 14] and as one of the pre-requisites for a sense of object permanence [7, 5, 19].

It has been suggested that *reafference* can help to understand the origin of a 'self'. Reafference is a concept introduced by von Holst and Mittelstaedt [32] to make a distinction between the sensed consequences of actions and externally caused sensations, or exafference. Reafference is a characteristic of living organisms, offering a mechanism to predict sensory changes consequence of their actions, such as optical flow during motion. Proprioception is a basic form of reafference, which, together with efferent motor signals and visual input, contributes to the visual self-recognition and, thus, to the sense of agency [3]. Daprati et al. [11] suggest that the contribution of proprioceptive feedback for the sense of agency is time-constrained: while movement is ongoing, a constant monitoring of movement, as well as an actualization of the sense of agency, are needed. It has been proposed that the central nervous system adjusts the weights given to the reafference inflow (e.g. visual and proprioceptive information) and efferent outflow (i.e. motor commands) to estimate the position and velocity of the hand [33]. Thus, optimal sensorimotor integration for state estimation is based on previous experiences and on the noise or reliability of the signals. This is in line with the idea of a higher weighting of efferent information to rapidly differentiate ourselves from objects and others when visual information is not reliable [26]. Here, it is suggested that the sense of agency can be partially explained by the capability to accurately predict self-produced sensory changes. A further clue to the sense of agency can be given by the capability to monitor prediction errors to dynamically assign weights to motor, proprioceptive, and visual signals for optimal state estimation during movement in a task within a context.

In artificial agents, the sense of agency has been modeled mostly using the magnitude of the prediction error as an indicator. In Schillaci et al. [28], an artificial agent observes trajectories performed by itself or by another agent. Given that the agent is better at predicting the consequences of its own movements, the magnitude of the prediction error on the velocity profile indicates the authorship of the trajectories. In line with this idea, in Schillaci et al. [30], Bechtle et al. [5], Lang et al. [19], prediction error has been used as the main marker for distinguishing between sensory data generated by a humanoid robot and by another entity. A similar paradigm is adopted here, although tested in a more complex experimental setting, where multiple moving objects are present in the scene. In addition, we analyse the dynamic modulation of multi-modal integration weights, and propose this to be also an essential prerequisite for the development of subjective experience in artificial system.

B. Enhancing perceptual capabilities in artificial systems

This work focuses on visual perception in robotics and on the particular issue of visual occlusions. Despite decades of development, occlusion is still a challenge, especially for accurate motion estimation [21]. In computer vision, motion is estimated through optical flow calculations. Optical flow can be defined as the flow of visual changes resulting from the relative motion between an observer and the scene, or objects in the scene. Detecting and estimating movements is of crucial importance for any living agent, as well as for robots. Movements can be produced by different sources, including the agent itself, and anticipating them is fundamental for interaction.

Machine learning and computer vision literature on optical flow estimation is substantial. Robotics studies can be found addressing different applications, including navigation [34] and object manipulation [2]. Visual occlusions have been the subject of interest also for non-engineering sciences. Several related studies can be found in the developmental psychology literature, in particular on how infants deal with visual occlusions and how object permanence is perceived. In fact, the sense of object permanence is considered to be a milestone in infant development [7]. Likely occurring already during the firsts months of life, it is through the acquisition of the knowledge that objects persist in the environment, even if they are occluded, that infants start to form internal representations of objects and to handle them. Very early stages of development are already characterised by predictive capabilities, which infants typically exhibit through anticipatory gaze behaviours [1] Studies show that infants' knowledge of the permanence or non-permanence of objects is embodied in their predictive tracking [6].

Few robotics studies have linked the investigation on movement estimations with the cognitive development of a sense of object permanence. Here and in previous work [19, 5], it is argued that such a capability may be in part the result of predictive and sensory attenuation processes. Attenuating expected information – such as visually detected self-generated movements – could help maintaining the focus of attention on an object of interest. In this work, a new end-to-end multisensory model for enhancing robot perpeption and that deals with self-generated body movements occlusions or body self-occlusions, as well as with movements generated by other objects in the environment, is presented.

II. Methodology

We propose a multi-modal forward model for predicting changes in the visual field. The model is learnt with sensory data coming from a humanoid robot and is used for sensory enhancement and self-perception. The proposed framework has two distinctive features: first, a mechanism for attenuating predicted visual changes that are either generated by the robot itself or by other objects in the scene; second, the capability to learn the influence that each sensory and motor input has in

the prediction and how this is modulated to perform selective sampling of sensory information.

A similar mechanism – i.e., *precision weighting* – has been proposed in the predictive processing literature (see [10] for a review). In precision weighting, information incoming from different modalities is weighted according to the expected confidence given a certain task in a specific context [22]. The architecture proposed here is capable of learning such weights and of modulating them. This is performed through a mechanism inspired on features extraction processes and typical of autoencoder neural networks.

The model architecture is inspired by the works of Shim et al. [31] and Patel et al. [23], which combine convolutional networks with sensor fusion techniques and fusion weights regularization through auxiliary uni-modal branches for classification purposes. In their NetGated architecture [23], authors employ a gating and sensor fusion approach to integrate different modalities.The authors aim at preventing sensor failures during a classification task, as interference of a noisy channel can be gated off when the corresponding fusion weight is suppressed. A late fusion approach is adopted, where low-dimensional features extracted through uni-modal convolutional branches are integrated. Fusing the different modalities consists of multiplying the uni-modal features with the corresponding fusion weights. Such fusion weights are learned through a further extraction process, which is applied on each uni-modal low-dimensional feature.

However, as noted by Shim et al. [31], similar gating architectures can lead to inconsistency in the sensor fusion weights. NetGated may develop unstable fusion weights, which do not well reflect the relevance of the corresponding modalities in the specific task. To overcome this issue, Shim et al. [31] propose a series of variations to the original NetGated architecture.Among the proposed extensions, there is the introduction of auxiliary uni-sensory processing paths during training, and of fusion weight regularisation using the losses generated by such auxiliary branches.

We adopt a similar approach and extend it. First, we develop a forward model which integrates different sensory inputs and motor commands using a gating system as proposed by [31], and that produces predictions about changes in the visual input. Second, we introduce the possibility to modulate the fusion weights, in fact regulating the integration of the different inputs during a prediction step. This is achieved by overriding the values of the multi-modal integration weights resulting from the model training with those of additional inputs passed to the model, when needed. Moreover, model predictions are combined with upcoming visual inputs, and used in sensory attenuation processes. The forward model is characterised by the following inputs and outputs.

Inputs:

- visual input V recorded at time (t), consisting of a 32×32 pixels grayscale image captured from the left camera of a simulated iCub robot;
- proprioceptive input P recorded at time (t), consisting of the absolute positions of the 16 joints of the left arm and hand (including fingers) of the iCub robot;

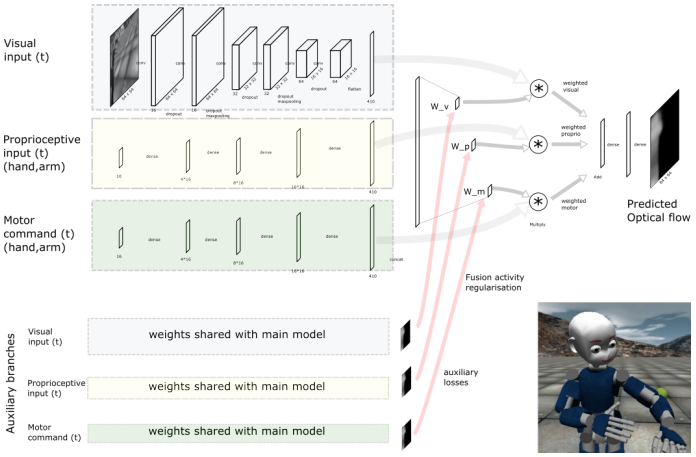

Fig. 1. An illustration of the implemented forward model and a screenshot of the humanoid robot iCub.

- motor command M applied at time (t), consisting of a 16-D vector representing the target positions for the same left arm and hand joints.

Outputs:

- the magnitude of the optical flow (OF) computed between the visual inputs at time (t) and at time (t+1). The resulting OF image consists of a 32×32 pixels matrix.

In the model structure, depicted in Figure 1, each input signal flows through a sequence of layers in the respective uni-modal branch. The three uni-modal branches differ from the type of processed input. The visual branch consists of a sequence of convolutional, dropout (rate: 40%) and max-pooling layers terminating on a feature layer of 256 neurons. The proprioceptive and motor branches consist of sequences of dense and dropout (rate: 40%) layers, terminating in two feature layers of the same dimension as the visual one.

Input signals flow through different paths in the feature layers. The path extracting the multi-modal integration weights concatenates, first, the three uni-modal 256-D feature layers. The concatenated tensor is reduced through a series of dense layers to a 3-dimensional tensor, which is then processed by a SoftMax layer. This ensures that the output activation of the three neurons sum up to one. The resulting layer is split into three single neurons, each representing the weight of the corresponding modality for the multi-modal integration. Finally, regularisation of the weight activity is performed. Multi-modal integration weights are multiplied to the uni-modal 256-D features extracted earlier, modulating, in fact, the output of each uni-modal branch. At this point, weighted multi-modal integration is performed by summing up the incoming signals. The result of the multi-modal integration is thus passed through a sequence of dense, reshape and upsampling layers, until the desired shape $(32, 32)$ of the model output is achieved.

Importantly, the three uni-modal branches (from the input layers to the 256-D feature layers) follow also a parallel, auxiliary path. In particular, each uni-modal input is mapped to an exact copy of the model output layer (layers and connections between them are shared with the main model). This allows

to calculate auxiliary uni-modal losses and thus to estimate how efficient is every single modality in generating the desired output. Hence, the full architecture is characterised by four outputs: the main model output and three uni-modal branch outputs. During training, the same optical flow information is passed to the four outputs of the model.

Model optimisation is performed through an Adam optimiser [18]. A loss function that combines the main model loss with a weighted average of the auxiliary losses is adopted. In particular, the following loss function has been implemented:

$$Loss = Loss_{main} + \alpha \cdot \sum_{n=1}^{N} w^n \cdot Loss_{aux}^n \qquad (1)$$

where $Loss_{main}$ is the loss of the main model, α is a constant (set to 0.5 in the experiments presented here), N is the number of modalities (i.e., visual, proprioception and motor), w is the weight used for multi-modal integration and $Loss_{aux}^n$ is the auxiliary loss of the modality n. Each loss is computed as the mean squared error between the targets and the predicted outputs. In addition, activities of the three multi-modal integration weights are regularised as follows:

$$w^n = w^n - \beta \cdot (w^n - e^{-(Loss_{aux}^n)^2}) \qquad (2)$$

where β is a constant set to 0.05 in the experiments presented here, and the exponential is passed through a sigmoid function. These equations are adapted from [31].

A. Sensory attenuation

We propose a mechanism that combines forward model predictions and upcoming visual inputs with the aim of attenuating visually perceived movements.Given the visual and proprioceptive inputs recorded at time (t) and a motor command, the forward model generates an optical flow (OF) prediction, i.e., an prediction of the movements to be perceived in the upcoming visual input. The predicted and the observed magnitudes of the OF images are then binarised applying a threshold to these values. The binarised OF prediction is used as a mask for the upcoming visual input, $V(t+1)$. In particular, pixels in $V(t+1)$ whose coordinates correspond to those OF pixels that have maximum magnitude values are *attenuated*. Here, attenuation is performed simply by substituting such pixels with those of a background image that has been stored at the beginning of the experiment.

In the experiments presented below, we measure sensory attenuation in an object detection task. In the experimental setup, a set of nine objects are placed in front of the robot. These objects are tagged with fiducial markers to ease their detection, as illustrated in Figure 2. We measure the quality of sensory attenuation by comparing the number of objects detected in the observed image – where occlusions due to the movements of the robot or of other objects may occur – with those detected in the image where predicted changes in the image are attenuated. Performances under different training and test configurations are examined, as well as under different modulations of the multi-modal integration weights.

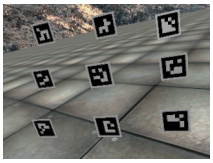

Fig. 2. The set of objects positioned in front of the robot.

B. Design of the experiments

A series of experiments has been carried out in which we analyse different aspects of the proposed model. First, we trained and tested the model on different datasets, analysing its learning performance under the different configurations. Second, we tested the capability of the proposed system to enhance robot perception in an object detection task (as described in Section I-B), as well as its capability to characterise self-generated movements by means of prediction errors. Finally, we measured sensory attenuation and object detection performance under different modulations of the multi-modal integration weights.

Three datasets have been collected for training and testing the model. Each sample of the datasets is characterised by visual, proprioceptive and motor information grabbed at a time (t), and the optical flow estimated between the visual inputs at time (t) and ($t + 1$). Datasets have been collected under three different experimental conditions simulated in the iCub humanoid robot simulator:

- DS1: the iCub robot, alone in the scene, randomly explores the movements of its left arm, hand and fingers.
- DS2: the robot is moving as in DS1, while coloured balls fall from the sky from random positions. A new ball is created every 20 seconds.
- DS3: same conditions as in DS2, but with more simulated objects (a new ball is created every second).

A set of nine static objects was also present in the scene, as depicted in Figure 2, during all the data collection processes. Objects were tagged with fiducial markers to ease their detection. Each dataset consisted of 25000 samples, gathered with a sampling rate of 5 samples per second.

We designed a set of experiments combining the use of the described sample collections either as training or test datasets, as depicted in Table I[1].

Exp. ID	Training dataset	Test dataset
1	DS1	DS1
2	DS1	DS2
3	DS1	DS3
4	DS1 (Phase 0), DS3 (Phase 1), DS1 (Phase 2), DS1 (Phase 3), DS3 (Phase 4), DS3 (Phase 5)	DS1
5	DS1 (Phase 0), DS3 (Phase 1), DS1 (Phase 2), DS1 (Phase 3), DS3 (Phase 4), DS3 (Phase 5)	DS3
6	Combined DS1 and DS3	DS1
7	Combined DS1 and DS3	DS3

TABLE I: Design of experiments. Legend: DS1: iCub alone in the scene; DS2: iCub and balls falling in the scene; DS3: iCub and many balls falling in the scene.

[1]When adopted as training dataset, 95% of the total 25000 samples in the corresponding collection were used, whereas only 5% of the total 25000 samples was used if adopted as testing dataset. Datasets were shuffled

Experiments 1, 2 and 3 have been carried out to test whether prediction errors can be used as a means to detect changes in sensory information that are not produced by the robot itself. In particular, three models are trained with the same dataset (DS1: iCub alone in the scene), however each one of them is tested with a different test dataset (DS1 in experiment 1; DS2 in experiment 2; DS3 in experiment 3). We expect prediction errors to be, in average, higher in exp. 2 and 3 – where moving objects are present in the scene – than in exp. 1. This study extends the experiments presented by Bechtle et al. [5] and Lang et al. [19] by adopting a more complex computational model and experimental setup. In addition, a sensory attenuation and object detection experiment has been carried out for each of the models trained as in Table I. We test the contribution of sensory attenuation processes in enhancing robot perception: attenuating self-generated movements, or those generated by other entities, allows detecting objects even if they are occluded in the raw visual input. In experiments 4 and 5 (see Table I), the proposed forward model has been exposed, sequentially, to sensorimotor data belonging to different training datasets. Each of these experiments consisted of six training phases. The training dataset was changed for each phase. The aim of these experiments is to analyse the re-modulation of the multi-modal integration weights of the model during the learning process. We expect the model to adapt to such dynamic circumstances, and to re-modulate its MIW when visual information becomes too unreliable to be used for predicting upcoming sensory input, as occurring in DS3. Finally, experiments 6 and 7 have been carried out using both DS1 and DS3 as training datasets.

Each experiment has been run 10 times. A run consisted of a 10-epochs offline training session, where batches of 32 samples were processed in each training iteration. Moreover, in each experiment we quantify the impact of multi-modal integration weights modulation onto the attenuation process. In particular, we analysed the object detection performance of each model on images of the corresponding test dataset, under the following configurations: (i) no sensory attenuation is carried out; (ii) sensory attenuation is performed using the model's prediction – i.e, using the multi-modal integration weights resulting from the model training (MIW from now); (iii) sensory attenuation is performed using the model's prediction resulting from manually modulating the multi-modal integration weights. In this last configuration, six different sets of weights have been tested, as described in Table II. Object detection performance has been measured as the average number of objects detected in the given image dataset. The results of the experiments are described in the following section. Datasets, scripts for setting up the experimental environment and source code are openly available[2].

[2]Source code is available here: https://github.com/guidoschillaci/icub_perceptual_optimisation.git. Tensorflow 2 has been used. Datasets are available in the Zenodo platform [27].Experiments have been tested on different platforms, including an NVidia Jetson Nano and cloud computing resources – gently offered by the OCRE (EU-H2020 Open Clouds for Research Environment) project. Scripts for setting up Docker environments for the different platforms are also available at the provided links.

	visual	proprio	motor
WS0	0.6	0.3	0.1
WS1	0.5	0.3	0.2
WS2	0.3	0.4	0.3
WS3	0.45	0.1	0.45
WS4	0.2	0.3	0.5
WS5	0.1	0.3	0.6

TABLE II: Custom multi-modal integration weights sets (*WS0-6* from now) tested in the object detection experiments.

III. Results

A. Experiments 1, 2 and 3

Experiments 1, 2 and 3 have been carried out to test whether sensory attenuation is better when objects are occluded by self-generated changes in the visual input than when they are due to other external moving objects (falling balls). Three models are trained with the same dataset (DS1: iCub alone in the scene), and tested with DS1, DS2 and DS3, respectively. The mean validation loss for each of the 10 epochs of the 10 runs has been compared over the experiments (see Fig. 3).

We used a one-way Welch's Heteroscedastic F Test with Trimmed Means and Winsorized Variances for independent samples to test for equal means (Welch's Test in what follows). The Welch's Test showed a significant difference in the validation loss across the experiments 1, 2 and 3, $F(2, 145.438) = 3846.481, p < 0.0001$. The post-hoc paired comparisons using a Holm correction showed a significant difference in validation loss between exp. 1, $M = 4.69e^{-05}, SD = 1.15e^{-05}$, and exp. 2 (DS2, $M = 8.31e^{-05}, SD = 1.47e^{-05}$, $p < 0.0001$), between exp. 1 (DS1) and exp. 3 (DS3, $M = 0.0003769, SD = 2.87e^{-05}, p < 0.0001$), and between exp. 2 (DS2) and exp. 3 (DS3, $p < 0.0001$). In line with the theories presented in section I and the results presented by Bechtle et al. [5] and Lang et al. [19], prediction errors were significantly lower in exp. 1 where the iCub was alone in the scene than exp. 2, where there were also other moving objects, and even to a greater extent when the rate of the presence of other moving objects increased, as occurred in exp. 3. In other words, average prediction error was significantly lower when the iCub was observing *only* its own movements.This supports the idea that prediction errors of a multi-modal forward model can be used as a cue for detecting activities performed by other individuals and, ultimately, for distinguishing self-generated sensory consequences from those generated by others.

Figure 4, shows the results of the object detection tests for exp. 1, 2 and 3. Each sub-figure shows a series of plots illustrating the mean number of objects detected in the original images and in those processed after sensory attenuation for each WS. In general, as can be seen in Figure 4, a perceptual enhancement of the objects occurred using the MIW resulting from the model training, as well as with the six custom multi-modal integration weights sets (WS0-5), in comparison to the number of objects detected where no sensory attenuation occurred (i.e, baseline condition).

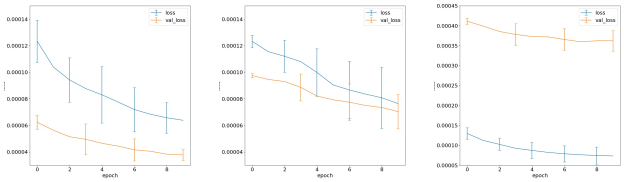

Fig. 3. Loss (blue plots) and validation loss (orange plots) of exp. 1, 2 and 3.

In addition, an object detection statistical analysis has been carried out, where optical flow predictions were generated using the MIW, as well as using the WS0-5 (Table I). We tested the contribution of the sensory attenuation processes in enhancing the robot object perception. Specifically, we tested if attenuating visually perceived movements, either self-generated or generated by other entities, allows detecting objects even if they are occluded in the raw visual input. The quality of sensory attenuation and the object detection performance under different configurations and multi-modal integration weight modulations was measured and analyzed. The mean of the total 'object detected' without any sensory attenuation was used as a baseline, and the differences on the 'objects detected' was calculated for each experiment. Thus, the data used for the statistical analysis was the result of subtracting the observed mean of each of the ten epochs of the ten total runs minus the corresponding baseline mean of each experiment. A total of 100 mean differences were calculated for each experimental condition, the MIW, and each one of the six WS0-5 (Table II), in each one of the three experiments.

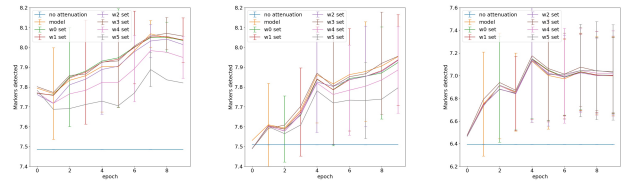

Fig. 4. Mean and standard deviation of objects detected in experiments 1, 2 and 3 over training epochs and applying sensory attenuation using the MIW and W0-5 (Table II). The blue line over training epochs, represents the baseline of objects detected, which is the mean of objects detected in the test image dataset for each experiment. The baseline is constant, as no sensory attenuation occurred and occlusions, due to robot movements or objects motion, may have occurred.

In exp. 1, the Welch's test showed a significant difference in the objects detected between the conditions $F(6, 245.343) = 9.652$, $p < 0.0001$. The post-hoc paired comparisons using a Holm correction showed a significant difference in the objects detected between the MIW ($M = 0.434$, $SD = 0.222$) and the *WS5* weights set ($M = 0.276, SD = 0.169$). This suggests that the dynamic MIW outperformed the *WS5* weights set (vision: 0.1, proprioception: 0.3, motor: 0.6). There were no significant differences in the objects detected between the MIW and the other remaining five different custom multi-modal integration weights sets.[3] Contrary to exp. 1, the Welch's test in exp. 2 ($F(6, 245.350) = 1.524, p = 0.17$), as well as in exp. 3 ($F(6, 245.677) = 0.114, p = 0.99$), showed no significant differences in the objects detected between the conditions.

From the trends depicted in Figure 4, it seems that the more the visual modality becomes noisy the more the model tends to rely to a greater extent on the motor modality. This trend could explain why, in exp. 1, a statistically significant difference was observed in the comparison between the object detection performance of the MIW and that of the model using *WS5*. This suggests that, when the visual modality is reliable, up-weighting the motor modality to predict the optical flow produces a worse sensory attenuation – and thus worse object detection – performance. Although the rest of the comparisons were not statistically significant, promising trends can be observed in Figure 4, where in exp. 2 and 3, the performances of *WS5* becomes closer to those of the main model, suggesting that the MIW increased to some extent the weights assigned to the motor modality.

Finally, two qualitative analyses on how the system modulates, during testing, the dynamic multi-modal integration weights were performed. The first analysis used a sample sequence of the iCub moving its arm occluding the objects (Figure 5), and the second, using a sequence of the iCub hand leaving the field of view of the robot camera (Figure 6). The aim of these qualitative analyses was to elucidate how the visual, proprioceptive, and motor weights were dynamically adjusted as a result of an incremental occlusion of the objects caused by a self-generated movement, as well as the opposite situation. First, as the robot arm incrementally occluded the objects throughout the movement, the proprioceptive and motor modalities weights were dynamically adjusted to be more influential for predicting the optical flow (see third row of Figure 5). It is plausible to think that, given the ample movement of the robot arm, proprioceptive and motor information needed to be highly weighted to correctly predict the OF product of self-generated movements.

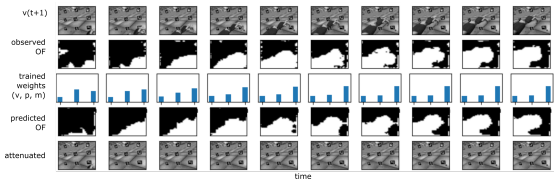

Fig. 5. A sample sequence showing the dynamics of the MIW while the robot hand is entering the field of view of the robot camera. The first row shows the observed images $v(t+1)$; the second row shows the observed OF; the third row shows the MIW; the fourth row shows the OF predicted by the model; finally, the fifth row shows the image resulting from the sensory attenuation.

In the second test, the iCub hand is leaving the field of view of the camera. At the beginning of the sample sequence (see Figure 6), the visual weights increased because the hand is

[3]The post-hoc paired comparisons of the MIW performance on the mean 'objects detected' between the six different custom multi-modal integration weights sets were of special interest for the present work. Therefore, only the post-hoc paired comparisons between the main model and the six different custom weights sets are reported.

almost out of the field of view. Then, the hand appeared again in the field of view and as a consequence the multi-modal weights were almost equally weighted. However, while the hand disappeared from the visual field throughout the sample sequence, the visual weights remained high.

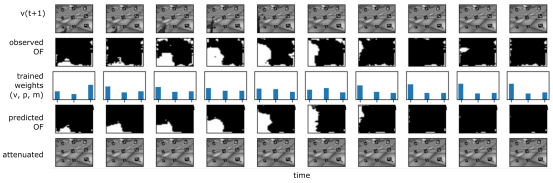

Fig. 6. A sample sequence showing the dynamics of the MIW while the robot hand is leaving the field of view of the robot camera.

B. Experiments 4 and 5

The aim of experiments 4 and 5 was to analyse the potential re-modulation of the MIW during the learning process, changing training dataset over six phases during the experiment.

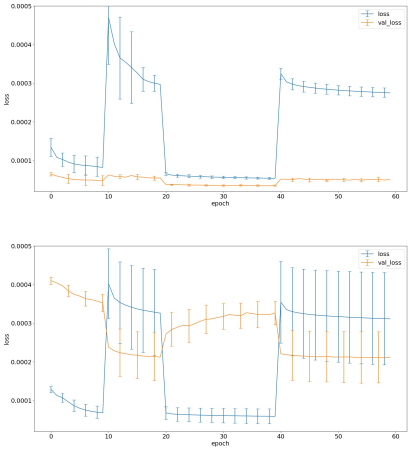

Fig. 7. Loss and validation loss of experiments 4 and 5.

Figure 7 shows the trends of the loss of the model using MIW in exp. 4 and 5, in each one of the six training phases with their respective training dataset (DS1, DS3, DS1, DS1, DS3, and DS3). Additionally, in Figure 7 the validation loss is shown for the dataset tests DS1 and DS3 used in exp. 4 and in exp. 5, respectively. Interestingly, even when both experiments were trained using the same sequences of datasets within phases, in exp. 4 the validation loss outstandingly outperforms the validation loss obtained in experiment 5. These findings can be interpreted as the model being able to learn and dissociate the visual consequences of self-generated movements and thus accurately predict the OF of these changes in all the test phases. In exp. 5, the validation loss curve shows that the model was not as good in predicting the optical flow caused by other sources than the robot. Although in phases 1, 4, and 5, the validation loss is lower than the loss during training, the prediction error is considerably higher in these phases than in exp. 4. Moreover, in exp. 5, the validation loss curve bumps within phases and only showed a tendency to decrease in the first two phases. This further supports the idea that prediction errors of a multi-modal forward model can be used as a cue for distinguishing self-generated sensory consequences from those generated by others, also in dynamic contexts. As expected, the model was able to adapt to dynamic circumstances, for instance, to re-modulate its weights when visual information became too unreliable to be used for predicting upcoming sensory input, as when being trained with DS3.

Figure 8 shows the average object detection rates in exp. 4 and 5. In comparison with the baseline, an enhancement of the robot visual perceptual capabilities when applying predictive and sensory attenuation processes can be seen in all the experimental conditions. Additionally, a tendency for switching the MIW can be seen depending on the dataset used. In both tests using the DS1 (exp. 4) and DS3 (exp. 5), the WS with high vision weights, such as *WS0* and *WS1*, tended to decrease their performance in phases 1, 4, and 5, where DS3 (whose samples have noisy visual inputs) was used during training. On the contrary, *WS4* and *WS5*, i.e. the sets with high proprioceptive and motor weights, tended to perform better in those phases.

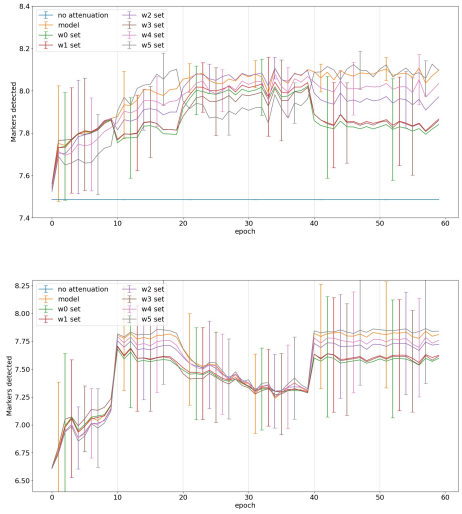

Fig. 8. Means of object detection in experiment 4 and 5 using DS1 and DS3 respectively.

As in previous experiments, a statistical analysis on the mean object detection has been carried out. For each phase, a total of 100 mean differences were calculated for each experimental condition, the MIW and each WS. In exp. 4, in its phase 0, the Welch's Test showed a significant difference in the objects detected between experimental conditions $F(6,\ 244.968) = 3.039$, $p = 0.009$. Although there were significant differences between some of the post-hoc paired comparisons, none of them were between the dynamic MIW and the six WS shown in Table III. In phase 1, there were significant differences $F(6,\ 242.956) = 32.390$, $p < 0.0001$. The post-hoc paired comparisons showed that the dynamic MIW significantly outperformed all the six WS. In phase 2, there were also significant differences $F(6,\ 243.565) = 31.337$, $p < 0.0001$. However, contrary to phase 0 and 1, in the post-

hoc paired comparisons, the *WS2* weights set ($M = 0.574$, $SD = 0.073$) performed equally well as the dynamic MIW ($M = 0.571$, $SD = 0.072$; $p = 1.0$). The same pattern of results were observed in phase 3 ($F(6, 243.508) = 26.241$, $p < 0.0001$), the *WS2* weights set ($M = 0.584$, $SD = 0.090$) performed equally well as the dynamic MIW ($M = 0.579$, $SD = 0.090$; $p = 1.0$), but here, also the *WS4* weights set ($M = 0.547$, $SD = 0.109$; $p = 0.14$). Finally, in phase 4 and 5 significantly differences were found $F(6, 239.687) = 40.078$, and $F(6, 238.417) = 36.647$, respectively. In both phases, the post-hoc paired comparisons showed that the *WS5* weights set (phase 4: $M = 0.606$, $SD = 0.086$; phase 5: $M = 0.606$, $SD = 0.092$) performed equally well as the dynamic MIW (phase 4: $M = 0.571$, $SD = 0.072$; phase 5: $M = 0.594$, $SD = 0.071$; $p = 1.0$).

Table III relates to exp. 4 and shows that the model re-adapted its MIW when switching training dataset. In phase 2 and 3, the model adjusted its MIW to perform as in those weights sets where proprioceptive and motor modalities have more influence than the visual one. Yet, the visual modality still has an influence in predicting on DS1. When the visual modality becomes too noisy during training (DS3), as in phases 4 and 5, the model seems to rely more on the motor modality. This trend can explain why, in phase 4 and 5, the *WS5* weights set performed equally well as the dynamic MIW to predict on DS1.

Exp 4 Train DS	Ph. 0 DS1	Ph. 1 DS3	Ph. 2 DS1	Ph. 3 DS1	Ph. 4 DS3	Ph. 5 DS3
Model vs WS0	n.s	***	***	**	***	***
Model vs WS1	n.s	***	**	**	***	***
Model vs WS2	n.s	***	n.s.	n.s.	***	***
Model vs WS3	n.s	***	***	***	***	***
Model vs WS4	n.s	**	**	n.s.	***	***
Model vs WS5	n.s	**	***	***	n.s.	n.s.

TABLE III: Comparisons of means of object detection with DS1 as test set in experiment 4. (Ph.: phase)

In exp. 5, the Welch's test showed that the difference in the objects detected between experimental conditions in phase 0 was not statistically significant ($F(6, 245.702) = 0.946$, $p = 0.46$). Therefore, post-hoc paired comparisons were not performed, nor reported in Table IV. In phase 1, there were statistically significant differences ($F(6, 238.699) = 20.872$, $p < 0.0001$). The post-hoc paired comparisons showed that the dynamic MIW ($M = 1.399$, $SD = 0.473$) significantly outperformed all the WS, except for the *WS5* weights set ($M = 1.43$, $SD = 0.482$; $p = 0.17$). In phase 2, there were also statistically significant differences ($F(6, 245.274) = 2.804$, $p < 0.0001$). Although there were significant differences between some of the post-hoc paired comparisons, none of them were between the dynamic MIW and the custom WS as shown in Table IV. In phase 3, the difference was not statistically significant ($F(6, 245.422) = 0.343$, $p = 0.91$), as in phase 0, hence post-hoc paired comparisons were not performed. Finally, in phase 4 and 5, statistically significantly differences where found $F(6, 237.872) = 18.343$, $p < 0.0001$, and $F(6, 235.715) = 25.749$, $p < 0.0001$, respectively. As in phase 1, in phase 4, the dynamic MIW ($M = 1.417$, $SD = 0.477$) significantly outperformed all the WS, except for *WS5* weights set ($M = 1.445$, $SD = 0.488$; $p = 1.0$). However, in phase 5, although the dynamic MIW ($M = 1.418$, $SD = 0.478$) significantly outperformed all the WS, the *WS5* weights set was outperformed the dynamic MIW ($M = 1.453$, $SD = 0.489$; $p = 0.01$).

Exp 5 Train DS	Ph. 1 DS3	Ph. 2 DS1	Ph. 4 DS3	Ph. 5 DS3
Model vs WS0	***	n.s.	***	***
Model vs WS1	***	n.s.	***	***
Model vs WS2	***	n.s.	***	***
Model vs WS3	***	n.s.	***	***
Model vs WS4	**	n.s.	**	***
Model vs WS5	n.s.	n.s.	n.s.	**

TABLE IV: Comparisons of means of object detection with DS4 as test set in experiment 5. (Ph.: phase)

C. Experiments 6 and 7

The aim of experiments 6 and 7 was to test if the dissociation between visual changes produced by self-generated movements and changes that are externally produced in changing contexts can occur when different datasets are combined in the same training phase. This experimental manipulation can be understood as being exposed to a noisy context, where abrupt visual external changes can suddenly occur, and the occurrence of these changes is not predictable. For the simulation of this type of context, DS1 and DS3 were combined and randomly presented during training. DS1 was adopted as test dataset for exp. 6 and DS3 as test dataset for exp. 7. Figure 9 shows the trends of the loss and the validation loss of the model under the different experimental conditions, as depicted in Table I. The Welch's test showed that the difference in the validation loss between exp. 6 and 7 was statistically significant ($F(1, 79) = 209.645$, $p < 0.0001$). As expected, the prediction error was significantly lower when the iCub was observing its own movements (DS1) than when there were other objects moving in the environment (DS4), despite during training the model was exposed to both datasets equally likely. This finding can be interpreted as the model being able to dissociate the self-generated visual changes from the externally visual changes and to accurately predict the resulting optical flow of its movements.

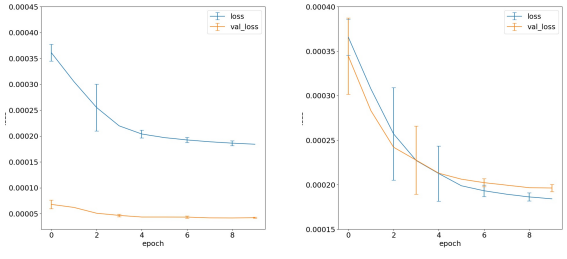

Fig. 9. Loss and validation loss of experiments 6 and 7.

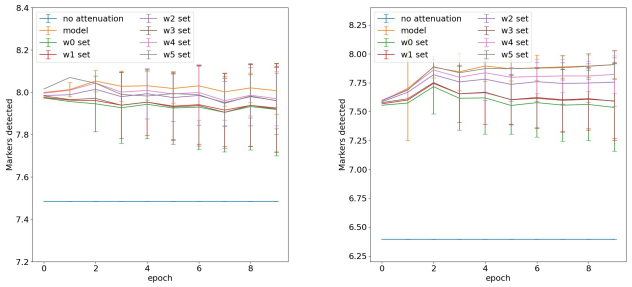

Fig. 10. Comparison of the object detection tests between experiments 6 and 7.

Figure 10 shows the results of the mean object detection tests for exp. 6 and 7, depicted in Table I. As in the previous experiments, an enhancement of the robot visual perceptual capabilities for object detection can be seen in all the experimental conditions of both experiments in comparison with the baseline. Results of the statistical analyses of the comparisons of object detection under the different configurations are reported in Table V.

	Exp. 6	Exp. 7
Model vs WS0	*	***
Model vs WS1	*	***
Model vs WS2	n.s	***
Model vs WS3	*	***
Model vs WS4	n.s	n.s
Model vs WS5	n.s	n.s

TABLE V: Comparisons exp. 6 and 7

In exp. 6, the Welch's test showed a statistical significant difference in the objects detected between experimental conditions ($F(6, 241.557) = 4.372$, $p = 0.0003$). The post-hoc paired comparisons indicate that the *WS2* weights set ($M = 0.495$, $SD = 0.115$; $p = 0.51$), the *WS4* weights set ($M = 0.511$, $SD = 0.097$; $p = 0.91$), and the *WS5* weights set ($M = 0.511$, $SD = 0.101$; $p = 0.31$), performed equally well as the dynamic MIW ($M = 0.534$, $SD = 0.071$). In exp. 7, a similar pattern of results was observed[4].

Therefore, in exp. 6, the dynamic MIW tended to rely mainly on proprioceptive and motor modalities, but also recruited the visual modality, as for the *WS2* weights set. Interestingly, in exp. 7, where DS3 was used as test dataset, the model tended to rely on proprioceptive and motor modalities to a greater extent, likely due to the fact that the visual modality was not as reliable as in exp. 6. Therefore, when the model was trained under a noisy context, where abrupt visual external changes occurred suddenly, and the occurrence of these changes were not predictable, the multi-modal trained weights tended to be higher on the proprioceptive and motor modalities. However, it is important to highlight that during training there were also data examples that were not noisy and only self-generated visual changes occurred, and even in some examples non visual changes occurred (DS1). In this examples, the model should up-weight the visual modality as can be seen in Figure 6, which can explain why the model, when being tested with DS1 performed equally well as the *WS2* weights set.

[4]A statistical significant difference occurred among experimental conditions ($F(6, 243.472) = 19.166$, $p < 0.0001$), but here, only the *WS4* weights set ($M = 1.388$, $SD = 0.285$; $p = 0.095$), and the *WS5* weights set ($M = 1.436$, $SD = 0.292$; $p = 1.0$), performed equally well as the dynamic MIW ($M = 1.441$, $SD = 0.282$).

IV. Conclusions and Future Works

This work has presented a computational model for multi-modal integration in a simulated humanoid robot. The model allows anticipating and attenuating movements from the visual input that are either self-generated by the robot or generated by other objects. Thanks to these processes, the visual perceptual capabilities of the robot can be enhanced. Here, sensory enhancement has been tested on a experiment on object detection and object permanence: through anticipation and attenuation processes, objects that are originally occluded by self-generated movements or those generated by other entities can be maintained in the field of view of the robot. The results presented in this study are in line with the idea that a sense of object permanence, an important capability acquired by infants during early developmental stages, may rely in part on predictive processes. Moreover, the prediction errors generated by the computational model under study can be used to characterise parts of the visual inputs as produced by the robot itself or by other objects as demonstrated by exp. 1, 2, and 3. This is a fundamental capability for self-perception and for self-other distinction.

An important feature of the computational model proposed here is the capability to extract multi-modal integration weights. This is the capability to learn how influential each input modality is in the process of predicting an upcoming visual input. We extended a model proposed by Shim et al. [31] and Patel et al. [23], providing the possibility to manually modulate the multi-modal integration weights, and analysing the model's capability to learn these features and to adapt them to dynamic environmental circumstances.

The observed trends in the modulation of multi-modal integration suggest different interesting venues for further research. The current process for attenuating expected movements from the visual input is based on the binarisation of the magnitude of the optical flow. This introduces an additional error that may bias the attenuation and object detection performances – which depend on the chosen threshold. Further work should test a more robust and comprehensive algorithm for combining predicted and observed optical flows and for allowing also partial attenuation. The direction of the movements could also be taken into account in the predictive model. Moreover, a better mechanism for encoding the background images should be considered. Convolutional networks could be used to encode the current background and context into a short-term and working memory representation. This would allow also extending the algorithm to support head movements, which are likely to produce dense optical flow information over the whole visual input. Moreover, measures to allow an incremental learning of the model and to prevent catastrophic forgetting issues in the dynamic context tests should be adopted. More complex object detection scenarios could be

also developed. Nonetheless, multi-modal integration should be also validated on an experiment that is agnostic to the object detection task. For instance, the comparisons between the multi-modal integration using weights resulting from the model training and the custom multi-modal integration weights could be also tested on the raw prediction error of the model. Moreover, custom multi-modal integration weights, not fixed for the entire duration of the test sequence, should be tested.

ACKNOWLEDGEMENTS

GS has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 838861 (Predictive Robots). Predictive Robots is an associated project of the Priority Programme "The Active Self" funded by the DFG (German Research Foundation).

REFERENCES

[1] J. B. Applin and M. M. Kibbe. Six-month-old infants predict agents' goal-directed actions on occluded objects. *Infancy*, 24(3):392–410, 2019.

[2] V. E. Arriola-Rios, P. Guler, F. Ficuciello, D. Kragic, B. Siciliano, and J. L. Wyatt. Modeling of deformable objects for robotic manipulation: A tutorial and review. *Frontiers in Robotics and AI*, 7, 2020.

[3] D. Balslev, J. Cole, and R. C. Miall. Proprioception contributes to the sense of agency during visual observation of hand movements: Evidence from temporal judgments of action. *Journal of Cognitive Neuroscience*, 19(9): 1535–1541, sep 2007.

[4] P. M. Bays, J. R. Flanagan, and D. M. Wolpert. Attenuation of self-generated tactile sensations is predictive, not postdictive. *PLoS Biol*, 4(2):e28, 2006.

[5] S. Bechtle, G. Schillaci, and V. V. Hafner. On the sense of agency and of object permanence in robots. In *IEEE ICDL-EpiRob*, pages 166–171. IEEE, 2016.

[6] B. I. Bertenthal, M. R. Longo, and S. Kenny. Phenomenal permanence and the development of predictive tracking in infancy. *Child Development*, 78(1):350–363, 2007.

[7] J. G. Bremner, A. M. Slater, and S. P. Johnson. Perception of object persistence: The origins of object permanence in infancy. *Child Development Perspectives*, 9(1):7–13, 2015.

[8] P. Cardoso-Leite, P. Mamassian, S. Schütz-Bosbach, and F. Waszak. A new look at sensory attenuation: Action-effect anticipation affects sensitivity, not response bias. *Psychological science*, 21(12):1740–1745, 2010.

[9] N. Cauli, E. Falotico, A. Bernardino, J. Santos-Victor, and C. Laschi. Correcting for changes: Expected perception-based control for reaching a moving target. *IEEE Robotics & Automation Mag.*, 23(1):63–70, 2016.

[10] A. Ciria, G. Schillaci, G. Pezzulo, V. V. Hafner, and B. Lara. Predictive processing in cognitive robotics: a review. *arXiv preprint arXiv:2101.06611*, 2021.

[11] E. Daprati, A. Sirigu, and D. Nico. Remembering actions without proprioception. *Cortex*, 113:29–36, apr 2019.

[12] F. P. De Lange, M. Heilbron, and P. Kok. How do expectations shape perception? *Trends in cognitive sciences*, 22(9):764–779, 2018.

[13] Y. K. Georgie, G. Schillaci, and V. V. Hafner. An interdisciplinary overview of developmental indices and behavioral measures of the minimal self. In *IEEE ICDL-EpiRob*, pages 129–136. IEEE, 2019.

[14] V. V. Hafner, P. Loviken, A. P. Villalpando, and G. Schillaci. Prerequisites for an artificial self. *Frontiers in neurorobotics*, 14, 2020.

[15] P. Haggard. Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18(4):196, 2017.

[16] H. v. Helmholtz. Concerning the perceptions in general. *Treatise on physiological optics,*, 1866.

[17] G. Juravle, F. McGlone, and C. Spence. Context-dependent changes in tactile perception during movement execution. *Frontiers in psychology*, 4:913, 2013.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] C. Lang, G. Schillaci, and V. V. Hafner. A deep convolutional neural network model for sense of agency and object permanence in robots. In *IEEE ICDL-EpiRob*, pages 257–262. IEEE, 2018.

[20] B. Lara, D. Astorga, E. Mendoza-Bock, M. Pardo, E. Escobar, and A. Ciria. Embodied cognitive robotics and the learning of sensorimotor schemes. *Adaptive Behavior*, 26 (5):225–238, jun 2018.

[21] P. Liu, M. Lyu, I. King, and J. Xu. Selflow: Self-supervised learning of optical flow. In *IEEE/CVF CVPR*, pages 4571–4580, 2019.

[22] T. Parr and K. J. Friston. Working memory, attention, and salience in active inference. *Scientific reports*, 7(1): 1–21, 2017.

[23] N. Patel, A. Choromanska, P. Krishnamurthy, and F. Khorrami. Sensor modality fusion with cnns for ugv autonomous driving in indoor environments. In *IEEE/RSJ IROS*, pages 1531–1536. IEEE, 2017.

[24] M. Pyasik, A. Salatino, D. Burin, A. Berti, R. Ricci, and L. Pia. Shared neurocognitive mechanisms of attenuating self-touch and illusory self-touch. *Social cognitive and affective neuroscience*, 14(2):119–127, 2019.

[25] N. Reissland, B. Francis, E. Aydin, J. Mason, and B. Schaal. The development of anticipation in the fetus: A longitudinal account of human fetal mouth movements in reaction to and anticipation of touch. *Developmental psychobiology*, 56(5):955–963, 2014.

[26] R. Salomon, M. Lim, O. Kannape, J. Llobera, and O. Blanke. "self pop-out": agency enhances self-recognition in visual search. *Experimental Brain Research*, 228(2):173–181, may 2013.

[27] G. Schillaci. Dataset from simulated iCub humanoid robot for sensory enhancement and perceptual optimisation, 2021. URL https://doi.org/10.5281/zenodo.4596464.

[28] G. Schillaci, V. V. Hafner, B. Lara, and M. Grosjean. Is that me? sensorimotor learning and self-other distinction in robotics. In *ACM/IEEE HRI*, pages 223–224, 2013.

[29] G. Schillaci, V. V. Hafner, and B. Lara. Exploration be-

haviors, body representations, and simulation processes for the development of cognition in artificial agents. *Frontiers in Robotics and AI*, 3:39, 2016.

[30] G. Schillaci, C.-N. Ritter, V. V. Hafner, and B. Lara. Body representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents. In *ALIFE*. MIT Press, 2016.

[31] M. S. Shim, C. Zhao, Y. Li, X. Zhang, W. Zhang, and P. Li. Robust deep multi-modal sensor fusion using fusion weight regularization and target learning. *arXiv preprint arXiv:1901.10610*, 2019.

[32] E. von Holst and H. Mittelstaedt. Das reafferenzprinzip. *Naturwissenschaften*, 37(20):464–476, jan 1950.

[33] D. Wolpert, Z. Ghahramani, and M. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232): 1880–1882, sep 1995.

[34] T. Zhang, H. Zhang, Y. Li, Y. Nakamura, and L. Zhang. Flowfusion: Dynamic dense rgb-d slam based on optical flow. In *IEEE ICRA*, pages 7322–7328, 2020.

[35] S. Zoia, L. Blason, G. D'Ottavio, M. Bulgheroni, E. Pezzetta, A. Scabar, and U. Castiello. Evidence of early development of action planning in the human foetus: a kinematic study. *Experimental Brain Research*, 176(2):217–226, 2007.