

DeepGrasp: Modeling digital humans with task-driven learning

Bartłomiej Borzyszkowski

*School of Computer and Communication Sciences
EPFL, Swiss Federal Institute of Technology Lausanne*

Abstract—Task-driven learning with deep neural networks is a powerful, emerging approach to elucidate and model computations in the brain. This work focuses on human-object interaction and aims to better understand, predict, and model human behaviour using digital avatars. We train a Graph Convolutional Network (GCN) with spatial and temporal attention on multi-modal data, including body pose and touch signals. Our results show that the model learns a robust neural representation capable of solving multiple classification and regression tasks in 3D space, such as action recognition and object classification. The results demonstrate that task-driven multi-modal learning can capture meaningful structure in human grasping behavior, offering a step toward more accurate models of the sensorimotor system and advancing research in computational neuroscience.¹

I. INTRODUCTION

Body ownership is a fundamental aspect of our self-awareness and sense of self. It refers to the feeling of having a physical body that is under our control and that we identify with. Illusory body ownership refers to the phenomenon where individuals experience ownership over a body part that is not actually their own. This can occur in a variety of contexts, such as in virtual reality environments, where users may experience feeling as if they are inside a virtual body. Another well-known example is a rubber hand illusion wherein ownership over a hand model is experienced, and is generally believed to require synchronized stroking of real and fake hands [1].

Although much attention has been devoted to experimental work on body ownership, computational models that explain this phenomenon remain largely unexplored. In this work, we propose a novel approach using task-driven learning to model illusory body ownership. In task-driven learning, a model is trained to perform multiple related tasks simultaneously, building a coherent internal representation that supports efficient generalization across tasks. This approach is well-suited to modeling body ownership because it mirrors how the brain integrates multisensory signals. Compared to traditional methods, task-driven learning can capture more complex behavioral patterns while requiring less task-specific supervision.

We propose training the model on human-object interaction (HOI) data. This includes grasping, manipulating, and releasing objects, as well as exploring them through touch. Naturally, interaction with objects is a complex process that involves multiple sensory modalities. In this work, we focus on modeling touch and proprioception as we aim to gain new insights into their integration in the brain. When we interact with objects, we use touch to gather information about the object's shape, size, and texture. Proprioception is used to

control the movement and position of our hand and fingers in relation to the object.

As hands are the primary means by which humans manipulate objects in the real-world, analyzing human-object interaction holds great potential for understanding human behavior. However, existing datasets are limited in size and lack comprehensive awareness of the object's affordance and the hand's interactions with it. To address this gap, we introduce the HOPE Generator, a method for synthesizing large-scale interaction data. Our approach leverages GOAL [2] to generate naturalistic whole-body grasping motions, exploiting knowledge from GRAB [3], OakInk [4], and HOI4D [5]. The resulting dataset comprises diverse, dynamic interactions spanning over 2600 objects, providing a foundation for task-driven learning.

We hypothesize that models achieving high accuracy on tasks requiring multisensory integration will exhibit illusion-like behavior when presented with conflicting sensory inputs. To test this hypothesis, we evaluate our task-driven models on a rubber hand illusion (RHI) paradigm and compare their predictions to human responses. We establish an experimental pipeline for recording the RHI using Microsoft Kinect depth cameras, which capture detailed 3D hand position and movement. From these recordings, we reconstruct participants' bodies using the SMPL-X mesh [6] and provide them as input to the neural networks. This allows direct comparison between model predictions on the body localization task and the proprioceptive drift reported by participants.

The summary of our contributions is threefold:

- 1) We generate a large-scale dataset of naturalistic, dynamic whole-body human-object interactions spanning over 2600 objects, suitable for training deep neural networks on multimodal sensorimotor tasks;
- 2) We introduce a task-driven learning framework employing recurrent and graph neural networks, trained jointly on classification and regression tasks for human-object interaction. The models take tactile and proprioceptive inputs derived from the human body mesh.
- 3) We establish an experimental pipeline for evaluating model susceptibility to body ownership illusions by comparing predictions against human behavioral data from the RHI.

II. RELATED WORK

Illusory body ownership refers to the illusory perception of non-bodily objects (e.g., rubber hand) as being parts of one's own body [7, 8]. A number of experimental studies have been conducted to investigate the rubber hand illusion (RHI) and

¹Code and data are publicly available at:
<https://github.com/Borzyszkowski/DeepGrasp>

its underlying mechanisms [9–11]. Some studies investigated the role of inter-sensory integration in the illusion [12], while others examined the effect of different types of visual-tactile stimulation [13]. Additionally, research has been conducted to explore the potential applications of RHI in the treatment of phantom limb pain [14] and body image disorders [15]. The RHI provides important insights into our body awareness and how this process can be manipulated. Moreover, several computational models for body ownership have been proposed, but this field has yet to develop.

M. Samad et al. [16] showed that the process of the RHI can be modeled as a Bayesian Causal Inference (BCI). The BCI model explained that synchronous stroking of a dummy hand and a real hand produces the perception of a common cause for visual and tactile stimuli. Interestingly, their model predicted that the RHI can occur based on visual-proprioceptive integration absent tactile stimulation. These findings show that visual-proprioceptive integration is critical for the RHI, however they did not investigate somatic (non-visual) RHI setup.

Similar work has been done by D. Rezende et al. [17]. They applied a Bayesian Ideal Observer and proposed perception and response models to explain the major features of the RHI. Their model used visual and proprioceptive estimates of the hand position as well as the visuo-tactile vibrating stimulus. The results showed that the perceived hand position during the RHI is a combination of prior beliefs (top-down influence) and sensory input to three sensory modalities (vision, proprioception, touch).

Another work on modelling the RHI has been proposed by T. Rood et al. [18]. They apply a deep active inference model, where an artificial agent directly operates in a 3D virtual reality environment. The environment provides proprioceptive information on the shoulder and elbow. Interestingly, to learn the visual forward model they apply two neural networks: convolutional decoder and a variational autoencoder. Their results show that the inference model is able to produce similar perceptual and active patterns to those found in humans.

Early works assumed that proprioceptive drift can be used as a behavioral measure to assess the subjective feeling of body ownership [19], but recent studies call into question the relationship between the two [20]. K. Matsumiya proposed to investigate separate multisensory integration processes for ownership and localization of body parts [21]. He applied a statistically optimal cue combination paradigm and used maximum likelihood estimation to combine multiple sensory inputs. The results of his model show strong behavioral evidence that separate mechanisms of multisensory integration underlie ownership and localization of body parts.

Although preliminary work on computational models of body ownership has been proposed, it is fundamentally different from our approach. First of all, previous models are explicitly designed to account for RHI and are not capable of performing any other task. We argue that illusions are the effect of conflicting sensory information in the brain, therefore we aim to develop a model that is able to perform various hard inference tasks that require multisensory integration. We

hypothesise that a model that achieves reliable results in the task-driven procedure is susceptible to the illusions as a brain-like machine. This hypothesis can be verified by comparing models' predictions to neural and behavioural data. That way our work could provide deep insights into understanding processing of proprioception and touch in the brain.

III. LARGE-SCALE DATA SYNTHESIS

Our prior work on haptic perception [22] identified dataset scale as a key bottleneck for achieving high classification performance in task-driven learning. In this section, we address this limitation by developing a large-scale human-object interaction data synthesis pipeline, detailed in our work on the HOPE Generator [23].

A. Generation procedure

We propose to exploit knowledge from the real HOI datasets and extend them using our synthetic data. Firstly, we use GRAB [3], a dataset of whole-body humans that manipulate objects. As GRAB was recorded using MoCap, it is characterized by high precision at a cost of scale. It contains 51 objects and a total of 1334 dynamic interactions that are used for training and evaluation of our models. Secondly, we adapt OakInk [4], a large-scale dataset that contains 3D object meshes with virtual, single-hand, static grasps. We extend them by applying our generator to synthesize dynamic whole-body grasps for all 1800 objects from OakInk. Finally, we include HOI4D [5], a large-scale 4D egocentric dataset that contains another 800 different articulated and rigid-body object. Altogether, our method scales to dynamic interactions with over 2600 object instances from the three collected datasets.

B. Methodology

We propose to leverage GOAL [2], a novel approach to generating whole-body motion and object grasping (Fig. 1).

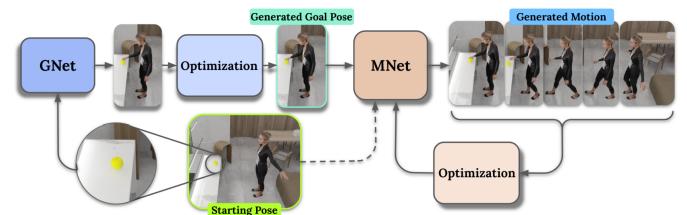


Fig. 1. Architecture overview: (1) GNet is an autoencoder that takes as input a 3D object, its position and orientation, and generates a static whole-body grasping pose; (2) MNet is an auto-regressor that takes as input a starting and final human poses, and generates the motion between them.

This method uses two neural networks: first, GNet generates a target grasp with a realistic body and hand-object contact; second, MNet generates the motion between the initial and target pose. We apply the method zero-shot and generalize it to the wide range of unseen data from OakInk. Specifically, interactions are generated without retraining the models and the original weights from GOAL are used.

C. Results

Although GOAL achieves remarkable results, the authors show its operation only on a small number of items. We adapt this method to over 1800 new objects and generate diverse and dynamic interaction for each of them (Fig. 2).

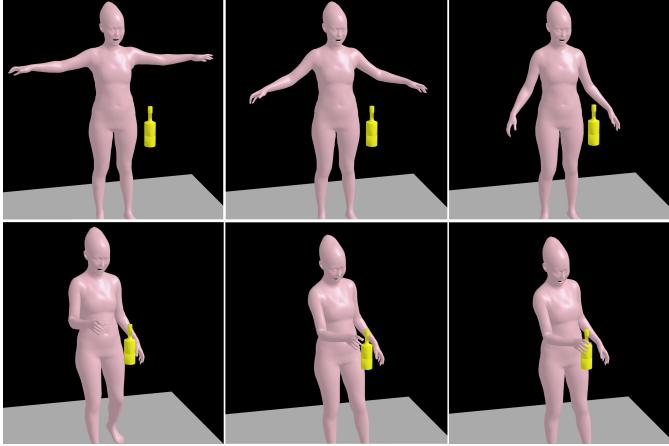


Fig. 2. Sample interaction generated for an unseen object from the OakInk dataset. We generate realistic final grasp (bottom right) and a sequence of motion that leads to it from the initial pose (top left). Our detailed results are available in Ref. [23].

D. Conclusion and future work

The qualitative and quantitative evaluation presented in our work on the HOPE Generator demonstrates its ability to synthesize physically plausible grasping motions at scale. We use these generated data in our task-driven learning procedure and release them to the research community. However, to make the data fully beneficial for studying stereognosis, several improvements can be proposed:

- conditioning on the object affordance;
- complex interactions (e.g., beyond the prehension phase);
- implementation of tactile and proprioceptive receptors, i.e., generating biologically plausible signals.

We plan to address these improvements in future work.

IV. TASK-DRIVEN LEARNING

The main contribution of this work is a novel task-driven learning approach to modeling illusory body ownership. The method builds on our prior work on haptic perception [22] and gradually increases the complexity of the models and the range of tasks they perform (Fig. 3). We evaluate our method on a variety of tasks and demonstrate that it learns a robust body representation, as evidenced by performance on the body self-localization task. We describe our task-driven learning procedure in the following sections.

A. Overview of the tasks

The tasks are learned in a supervised fashion and can be broken down into classification (1-2) and regression (3-4) sub-tasks as detailed below. They are performed jointly by a single model that learns a shared representation and extracts features that are useful across the sub-tasks.

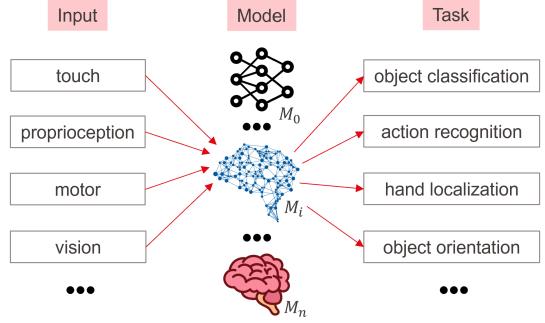


Fig. 3. Overview of the task-driven learning procedure. We develop a number of models of varying complexity, architecture, and parameters (middle). The models interpret multi-modal input (left), including touch and proprioception. We train the models on hard inference problems in a multi-task fashion (right) so that they learn internal representation of inputs. We hypothesise that a group of models achieving the highest performance on the relevant tasks is the most brain-like, i.e. the most susceptible to illusions.

1) *Object classification*: This task categorizes everyday objects of varying shape and size (e.g., water bottle, banana, wine glass). Initially, we consider 51 classes in the base GRAB dataset, which can be extended to 2600 classes (Sec. III).

2) *Action recognition*: This task categorizes each interaction as one of four intents: “use” and “pass” (to someone), as well as “lift” and “off-hand pass” (between the hands).

3) *Body localization*: This task learns to predict the 3D position of body joints at each timestep. During preprocessing, we subtract the neck position (in Cartesian coordinates) from all joint positions, yielding the relative distance from the neck to each joint. The model receives the axis-angle representation of joints and body shape parameters (bone lengths) as input, and predicts the Euclidean distance from the neck to each joint:

$$|JN| = \sqrt{(x_J - x_N)^2 + (y_J - y_N)^2 + (z_J - z_N)^2} \quad (1)$$

where $N = (x_N, y_N, z_N)$ denotes coordinates of a neck and $J = (x_J, y_J, z_J)$ denotes coordinates of a single joint.

4) *6D object pose estimation*: This task predicts the 6D object pose (3D position and 3D orientation) at each timestep. Given proprioceptive and tactile inputs, the model estimates the position and rotation of the object in 3D space. For evaluation, we use the average distance (ADD) metric, as proposed in [24]:

$$ADD = \frac{1}{m} \sum_{x \in M} \| (Rx + T) - (\tilde{R}x + \tilde{T}) \| \quad (2)$$

The ADD metric computes the mean pairwise distance between the 3D model points transformed by the ground truth pose (rotation R , translation T) and the estimated pose (rotation \tilde{R} , translation \tilde{T}), where M denotes the set of 3D object points and m is the number of points.

B. Methodology

Training models for multi-task human-object interaction presents several challenges, including handling variable sequence lengths, representing spatial body structure, and learning shared features across diverse tasks. In this section, we describe how our methodology addresses these challenges.

1) Data preparation: We observe that the class distribution of GRAB, the base training dataset, is imbalanced, which is particularly problematic in the multi-task setup where labels for each task should be evenly distributed across training, validation, and test sets. Sorting sequences by object and action type reveals a long-tail distribution (Fig. 4, left). For example, there are 28 sequences of a subject using a camera, but only one sequence for several other cases (e.g., passing a banana). To mitigate the impact of imbalanced data, we split sequences randomly while stratifying jointly on both labels. Incorporating our data augmentation procedure (Sec. III) further balances the samples, addressing this issue completely.

Varying sequence length is another challenge (Fig. 4, right). The longest interaction spans 4460 frames while the shortest contains only 491 frames, corresponding to 37s and 4s respectively at a recording rate of 120 FPS. This complicates training, as shorter sequences provide fewer features to the model. To address this, we extract the prehension phase and subsample all sequences to a fixed length of 80 frames. This step is also necessary for mini-batch training. An alternative approach based on characteristic poses has been proposed by Diller et al. [25]. Exploring this method may yield improved performance and is left for future work.

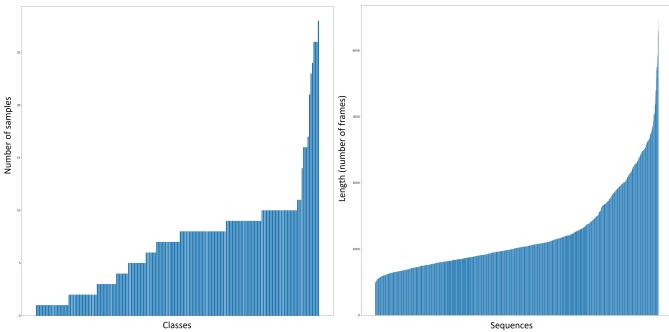


Fig. 4. Histograms showing data distribution for the number of classes (left) and sequence length (right). We notice that in both cases the dataset has long-tail distribution. The difference between two extreme classes is of order 28, and for two extreme sequence lengths of order 10.

2) Representing body as a graph: The baseline recurrent models such as Long Short-Term Memory (LSTM) excel at capturing temporal dynamics in sequential data. However, their architecture limits their ability to model spatial relationships required to build fundamental body awareness. It has been shown that Bodily Self-Consciousness (BSC) involves strong integration of proprioceptive, vestibular, and visual bodily inputs [13]. Consequently, we argue that modeling body ownership requires strong reliance on both spatial and temporal features.

We propose to represent the body as a graph, where joints correspond to nodes and bones to edges. We use a bi-directional, static graph with dynamic features (Fig. 5, left). The graph structure remains constant across subjects, while shape parameters and joint positions vary. In addition to proprioceptive input, we represent the hand surface as a separate graph with 778 nodes corresponding to the MANO hand model [26]. Binary contact values are assigned to each node to mimic the sense of touch and indicate hand-object contact (Fig. 5, right).

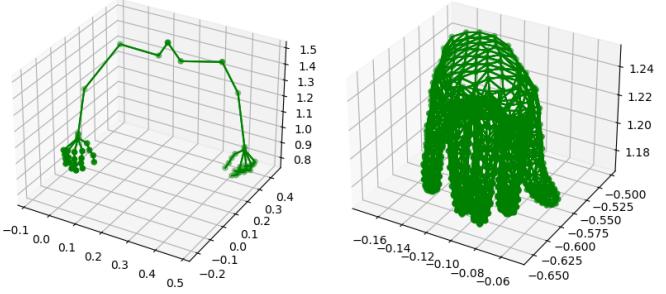


Fig. 5. Spatial representation of the body as a graph. We create bi-directional, static graphs with dynamic features for proprioception (left) and touch (right). The graphs are given as input to the Graph Convolutional Network models.

3) Architecture overview: We introduce the Topographic Deep Learning Baseline (TDLB) architecture, based on the Two-Stream Adaptive Graph Convolutional Network (2s-AGCN) [27]. This architecture has outperformed previous state-of-the-art methods by a significant margin and remains competitive on recent benchmarks. The model combines a Graph Convolutional Network (GCN) for spatial processing with an attention mechanism for temporal processing. Each AGCN block consists of a spatial GCN, a temporal GCN, and a batch normalization layer. The key advantage of this model is its adaptive graph convolutional layer, which processes two graph types: a global graph capturing patterns common across all data, and an individual graph capturing patterns unique to each sample. Additionally, the two-stream design processes skeleton data (first-order features) as well as bone lengths and directions (second-order features), fusing them to improve overall performance (Fig. 6). We leverage the implementation available in the PyTorch Geometric Temporal package [28] and modify it for our task setup and graph representation.

As input, the architecture takes the spatiotemporal graphs of the human skeleton. The graph topology is described by an adjacency matrix composed of three parts: A_k , B_k and C_k . Specifically, A_k represents the physical structure of the skeleton. B_k indicates strength of connections between two neighbouring joints, and C_k captures the similarity between two vertices. The graph convolution operation is defined as:

$$f_{out} = \sum_k^{K_v} W_k f_{in}(A_k + B_k + C_k) \quad (3)$$

where K_v denotes a kernel size; W_k is a weight vector and f corresponds to a feature map.

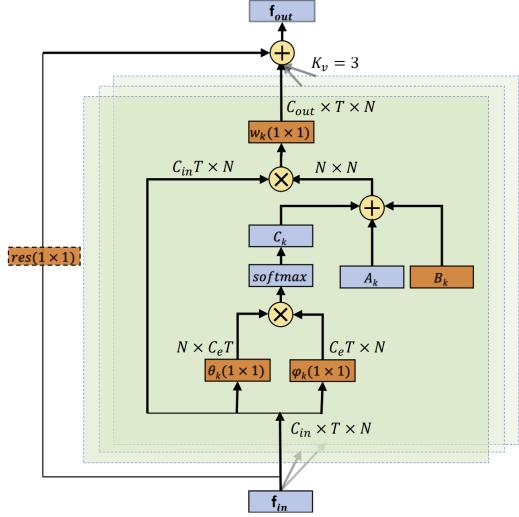


Fig. 6. Architecture of the adaptive graph convolution layer. Learnable parameters are marked in orange. Three distinct graphs, which form the adjacency matrix are symbolized as A_k , B_k , and C_k .

4) Training procedure and hyperparameter search: We develop a training procedure that supports neural architecture search and hyperparameter optimization at scale. To systematically explore the model configuration space, we employ Ray Tune [29], which provides a comprehensive suite of hyperparameter optimization algorithms including grid search, random search, and Bayesian optimization. For experiment tracking and comparison, we integrate Neptune [30], enabling real-time monitoring of training metrics and efficient management of multiple experimental runs. Our pipeline supports automated selection of model architecture, graph structure, input modalities, and task configuration, facilitating systematic evaluation across experimental conditions.

C. Results

We evaluate our task-driven learning approach using two architectures: the Recurrent Deep Learning Baseline (RDLB) and the Topographic Deep Learning Baseline (TDLB). The RDLB serves as a temporal baseline, while the TDLB extends this with spatial graph-based processing. Below, we present results for both architectures across all tasks.

1) Object classification (Tab. I): The TDLB architecture substantially outperforms the RDLB baseline on object classification. Using the full configuration (both hands, tactile and proprioceptive inputs), TDLB achieves a test accuracy of 40.65%, representing a 22 percentage point improvement over RDLB (18.71%). For reference, chance-level performance on this 51-class problem is approximately 2%. This improvement demonstrates the importance of spatial modeling through graph representations for distinguishing between object categories. We note that this task is particularly challenging due to the fine-grained similarity between many object classes. For instance, differentiating between cylindrical objects such as bottles and cans requires subtle tactile distinctions.

TABLE I
RESULTS OF THE OBJECT CLASSIFICATION TASK

Object Classification				
Model	Hands	Intent (classes)	Input	Test Accuracy [%]
TDLB	both	all (51)	T+P	40.65
RDLB	both	all (51)	T+P	18.71
TDLB	both	all (51)	T	35.48
RDLB	both	all (51)	T	14.19
TDLB	both	all (51)	P	35.98
RDLB	both	all (51)	P	23.87
TDLB	right	all (51)	T	36.25
TDLB	right	all (51)	P	34.25
RDLB	right	all (51)	T+P	17.42

2) Action recognition (Tab. II): Action recognition achieves high accuracy due to the clear distinctiveness of the four action classes. Unlike object classification, where categories share fine-grained similarities, actions involve fundamentally different movement patterns. For instance, lifting an object requires vertical motion while passing involves lateral extension toward another person. Consequently, this task can be solved with high accuracy using proprioceptive input alone, achieving 91.85% on the full four-class problem and 98.28% on the binary lift-versus-pass classification.

TABLE II
RESULTS OF THE ACTION RECOGNITION TASK

Action Recognition				
Model	Hands	Intent classes	Input	Test Accuracy [%]
TDLB	both	all (4)	P	91.85
TDLB	both	lift + pass (2)	P	98.28

3) Body localization (Tab. III): We report localization error as the Euclidean distance (in centimeters) between predicted and ground truth joint positions. Results indicate that centrally located joints (neck, collars, shoulders) are predicted with low error, while dynamically moving body parts (elbows, wrists) exhibit larger errors. TDLB consistently outperforms RDLB across all joints, with an average arm error of 7.39 cm compared to 14.91 cm. These results demonstrate that the models learn a coherent body representation encoding relative joint positions. However, prediction errors remain substantial for end effectors, motivating further work to meet the precision of proprioceptive position sense [31].

TABLE III
RESULTS OF THE BODY SELF-LOCALIZATION TASK

Body self-localization		
Body part	RDLB error [cm]	TDLB error [cm]
Arms average	14.91	7.39
Neck	9.51×10^{-4}	2.57×10^{-7}
Right collar	1.99	1.35
Left collar	2.41	1.32
Right shoulder	4.74	2.49
Left shoulder	6.26	3.16
Right elbow	11.49	6.68
Left elbow	9.46	5.81
Right wrist	14.79	7.56
Left wrist	14.29	6.98

V. BODY OWNERSHIP

Having demonstrated that our models learn robust body representations through task-driven learning, we now investigate whether these representations exhibit properties analogous to human body ownership. In this section, we describe our experimental setup for evaluating model predictions against human behavioral data from the rubber hand illusion.

A. Recording the Rubber Hand Illusion

We conducted recordings in the EPFL Smart Kitchen, a controlled environment equipped with nine side-view depth cameras (Microsoft Kinect) and a first-person view camera (HoloLens). The experimental setup captured the movements and interactions of participants' real hands alongside a rubber hand placed in their field of view. The Kinect depth sensors enabled precise 3D tracking of both hand positions throughout the experiment.

We recorded participants and measured illusion strength using questionnaire ratings on a 7-point Likert scale, following established protocols [32]. We systematically varied experimental conditions, including the distance between real and rubber hands, stimulation frequency, and stimulation location (Fig. 7). Participants reported the perceived location of their hand, providing a direct measure of proprioceptive drift for comparison with model predictions.



Fig. 7. Example frame from a recording of the RHI experiment in the EPFL Smart Kitchen. To induce the illusion, an experimenter strokes a left rubber hand with the blindfolded participant's right index finger while simultaneously stroking the corresponding part of the participants' left hand. The stimulation typically takes around 15s. During that time the participant experiences a proprioceptive drift towards the rubber hand.

B. Model evaluation on RHI data

To evaluate our models on recorded RHI data, we developed a processing pipeline that reconstructs participant body pose from RGB-D recordings. We employed Keypoint Communities [33] for 2D pose estimation and fit the SMPL-X body model to obtain full 3D body reconstructions. This enabled us to create a virtual representation of the RHI experiment (Fig. 8) that can be processed by our neural networks. We evaluated our models on the body self-localization task using the reconstructed body poses from the RHI recordings. Our preliminary results are promising: models that achieved high performance on the multisensory integration tasks exhibited behavior consistent with proprioceptive drift, predicting hand positions shifted toward the rubber hand location. This pattern mirrors the behavioral responses observed in human participants during the experiment. These findings provide initial evidence

that task-driven learning on problems requiring multimodal integration gives rise to body representations with human-like properties. Comprehensive results from this experimental study, including quantitative analysis across participants and conditions, will be presented in a forthcoming part of work.

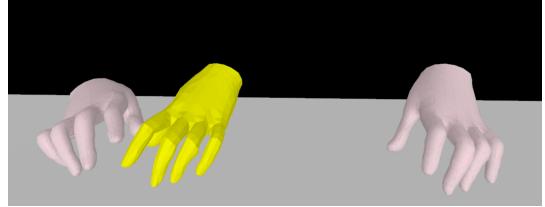


Fig. 8. Reconstructed RHI experiment using SMPL-X body model fit to RGB-D recordings. This virtual representation enables direct comparison between model predictions on the body localization task and participants' reported proprioceptive drift.

VI. DISCUSSION

This work presents a novel approach to modeling illusory body ownership using task-driven learning. We introduced a large-scale data synthesis procedure for human-object interaction that scales to over 2600 objects and generates diverse, dynamic, full-body grasping motions. This dataset provides a foundation for training models on multimodal sensorimotor tasks. We outlined potential extensions to further enhance its utility for studying stereognosis, including generating biologically plausible tactile signals.

We developed a task-driven learning framework employing recurrent networks and graph convolutional networks with spatial and temporal attention. We defined four tasks (object classification, action recognition, body localization, 6D object pose estimation) that require strong integration of tactile and proprioceptive features. Our proposed TDLB architecture, substantially outperformed the recurrent baseline (RDLB) across all tasks, demonstrating the importance of spatial modeling for learning coherent body representations.

Finally, we established an experimental pipeline for evaluating model predictions against human behavioral data from the rubber hand illusion. Our preliminary findings indicate that models achieving high performance on multisensory integration tasks exhibit behavior consistent with proprioceptive drift, mirroring human susceptibility to body ownership illusions. These results provide initial evidence supporting our central hypothesis: that task-driven learning on problems requiring multimodal integration gives rise to body representations with human-like properties.

This work contributes to the growing intersection of deep learning and computational neuroscience. By demonstrating that task-driven models can capture aspects of body ownership, we offer a new computational framework for investigating the mechanisms underlying bodily self-consciousness. We believe this approach has significant potential to advance our understanding of self-awareness and embodiment, with implications for fields ranging from cognitive neuroscience to prosthetics and virtual reality.

ACKNOWLEDGMENT

I would like to thank my advisors, Prof. Alexander Mathis and Prof. Olaf Blanke, for many valuable insights in formulating the research questions and methodology of this work. I would also like to express my gratitude to the lab members of AMG and LNCO for their advice and inspiring discussions on research directions in computational neuroscience and machine learning.

REFERENCES

- [1] Kiltени, Konstantina, et al. "Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception." *Frontiers in human neuroscience* 9 (2015): 141.
- [2] Taheri, Omid, et al. "Goal: Generating 4d whole-body motion for hand-object grasping." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- [3] Taheri, Omid, et al. "GRAB: A dataset of whole-body human grasping of objects." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV* 16. Springer International Publishing, 2020.
- [4] Yang, Lixin, et al. "OakInk: A Large-scale Knowledge Repository for Understanding Hand-Object Interaction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- [5] Liu, Yunze, et al. "HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- [6] Pavlakos, Georgios, et al. "Expressive body capture: 3d hands, face, and body from a single image." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
- [7] Tsakiris, M. My body in the brain: a neurocognitive model of body-ownership. *Neuropsychologia* 48, 703–712 (2010).
- [8] Kilteni, K., Maselli, A., Kording, K. P. and Slater, M. Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Front Hum Neurosci* 9, 141, <https://doi.org/10.3389/fnhum.2015.00141> (2015).
- [9] Ehrsson, H. Henrik, Nicholas P. Holmes, and Richard E. Passingham. "Touching a rubber hand: feeling of body ownership is associated with activity in multisensory brain areas." *Journal of neuroscience* 25.45 (2005): 10564-10573.
- [10] Brusa, Federico, Mustafa Suphi Erden, and Anna Sedda. "Influence of the Somatic Rubber Hand Illusion on Maximum Grip Aperture." *Journal of Motor Behavior* (2022): 1-19.
- [11] Lopez, Christophe, et al. "Tactile and vestibular mechanisms underlying ownership for body parts: a non-visual variant of the rubber hand illusion." *Neuroscience Letters* 511.2 (2012): 120-124.
- [12] Gallagher, Maria, Cristian Colzi, and Anna Sedda. "Dissociation of proprioceptive drift and feelings of ownership in the somatic rubber hand illusion." *Acta Psychologica* 212 (2021): 103192.
- [13] Blanke, Olaf, Mel Slater, and Andrea Serino. "Behavioral, neural, and computational principles of bodily self-consciousness." *Neuron* 88.1 (2015): 145-166.
- [14] Fang, Wen, et al. "Attenuation of pain perception induced by the rubber hand illusion." *Frontiers in Neuroscience* 13 (2019): 261.
- [15] Christ, Oliver, and Miriam Reiner. "Perspectives and possible applications of the rubber hand and virtual hand illusion in non-invasive rehabilitation: Technological improvements and their consequences." *Neuroscience and Biobehavioral Reviews* 44 (2014): 33-44.
- [16] Samad, Majed, Albert Jin Chung, and Ladan Shams. "Perception of body ownership is driven by Bayesian sensory inference." *PloS one* 10.2 (2015): e0117178.
- [17] Jimenez Rezende, Danilo. Behavioral and computational mechanisms of multi-sensory integration in humans. No. THESIS. EPFL, 2013.
- [18] Rood, Thomas, Marcel van Gerven, and Pablo Lanillos. "A deep active inference model of the rubber-hand illusion." *Active Inference: First International Workshop, IWAI 2020, Co-located with ECML/PKDD 2020, Ghent, Belgium, September 14, 2020, Proceedings 1*. Springer International Publishing, 2020.
- [19] Blanke, O. Multisensory brain mechanisms of bodily self-consciousness. *Nat Rev Neurosci* 13, 556–571 (2012)
- [20] Rohde, M., Di Luca, M., Ernst, M. O. Te Rubber Hand Illusion: feeling of ownership and proprioceptive drift do not go hand in hand. *PLoS One* 6, e21659, <https://doi.org/10.1371/journal.pone.0021659> (2011).
- [21] Matsumiya, Kazumichi. "Separate multisensory integration processes for ownership and localization of body parts." *Scientific reports* 9.1 (2019): 652.
- [22] Borzyszkowski B., Mathis A., Blanke O., "Creating datasets and machines for haptic perception", 1st semester PhD report in EDIC at EPFL.
- [23] Borzyszkowski B., Ahmadli M., Scurria A., Mathis A., "HOPE Generator: human-object interaction data synthesis", Project report in ML4Science, CS-433 at EPFL.
- [24] Hinterstoesser, Stefan, et al. "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes." *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I* 11. Springer Berlin Heidelberg, 2013.
- [25] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3D poses of human actions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- [26] Romero, Javier, Dimitrios Tzionas, and Michael J. Black. "Embodied hands: Modeling and capturing hands and bodies together." *arXiv:2201.02610* (2022).

- [27] Shi, Lei, et al. "Two-stream adaptive graph convolutional networks for skeleton-based action recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
- [28] Rozemberczki, Benedek, et al. "Pytorch geometric temporal: Spatiotemporal signal processing with neural machine learning models." Proceedings of the 30th ACM International Conference on Information Knowledge Management. 2021.
- [29] Liaw, Richard, et al. "Tune: A research platform for distributed model selection and training." arXiv preprint arXiv:1807.05118 (2018).
- [30] Neptune.AI: A metadata store for MLOps, built for teams that run a lot of experiments. <https://www.neptune.ai>
- [31] Van Beers, Robert J., Anne C. Sittig, and Jan J. Denier van der Gon. "The precision of proprioceptive position sense." Experimental brain research 122 (1998): 367-377.
- [32] Pozeg, Polona, et al. "Crossing the hands increases illusory self-touch." PLoS One 9.4 (2014): e94008.
- [33] Zauss, Duncan, Sven Kreiss, and Alexandre Alahi. "Key-point communities." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.