Transferring Implicit Knowledge of Non-Visual Object Properties Across Heterogeneous Robot Morphologies

Gyan Tatiya$^1$ Jonathan Francis$^2$ Jivko Sinapov$^1$

Abstract—Humans leverage multiple sensor modalities when interacting with objects and discovering their intrinsic properties. Using the visual modality alone is insufficient for deriving intuition behind object properties (e.g., which of two boxes is heavier), making it essential to consider non-visual modalities as well, such as the tactile and auditory. Whereas robots may leverage various modalities to obtain object property understanding via learned exploratory interactions with objects (e.g., grasping, lifting, and shaking behaviors), challenges remain: the implicit knowledge acquired by one robot via object exploration cannot be directly leveraged by another robot with different morphology, because the sensor models, observed data distributions, and interaction capabilities are different across these different robot configurations. To avoid the costly process of learning interactive object perception tasks from scratch, we propose a multi-stage projection framework for each new robot for transferring explicit knowledge of object properties across heterogeneous robot morphologies. We evaluate our approach on the object-property recognition and object-identity recognition tasks, using a dataset containing two heterogeneous robots that perform 7,600 object interactions. Results indicate that knowledge can be transferred across robots, such that a newly-deployed robot can bootstrap its object-identity recognition tasks from observed data distributions, and resultant model that each robot learns. Where the focus of prior work has been limited to learning non-visual modalities such as audio and touch are essential outperforms EDN in both tasks indicating transferring knowledge from robots to a shared latent space boosts the performance of the target robot. Furthermore, we propose a data augmentation technique independent of the learning task and show that using our data augmentation technique improves the models’ generalization and prevents overfitting.

I. INTRODUCTION

Humans learn about object properties by physically interacting with objects and perceiving multiple sensory signals, including vision, audio, and touch [1]–[6]. Interactions based on non-visual modalities such as audio and touch are essential, because vision alone is insufficient for detecting intrinsic object properties [7]: e.g., detecting whether an opaque bottle is full of liquid or empty. Recent works show that learning implicit knowledge of non-visual object properties leads to robots’ improved downstream performance, in material classification [8], liquid property estimation [9], object categorization [10], and human-robot dialogue interaction [11].

A robot may learn about object properties by performing exploratory interactions on objects and analyzing the effects via a diverse set of sensors [12]–[14]. The immediate issue is that this process is time-consuming, as it must be repeated for each robot. A natural desire may be to transfer representation of the object properties to a new robot to enable it to learn faster and complete its downstream tasks more efficiently. However, if the new robot has different interaction capabilities (e.g., different sensor models, or a different physical embodiment or morphology), the implicit knowledge gained by the previous robot is not directly transferable to the new one. Indeed, a robot’s machine learning model for the interactive perception tasks cannot be naturally applied to another robot because these models are specific to each robot’s embodiment, sensors, and environment [15]. While there is a great need to transfer implicit knowledge of object properties across heterogeneous robot morphologies, obtaining a general-purpose representation to facilitate rapid learning has remained challenging.

To address this challenge, we propose a framework that leverages learned projection functions to transfer implicit knowledge of non-visual object properties from a more-experienced source robot to a newly-deployed target robot. Specifically, we consider the general encoder-decoder network (EDN) model class [16] and the kernel manifold alignment (KEMA) method [17]–[19] as projection functions for learning object property-based and object identity-based correspondences. To test our framework, we collected a dataset of two robots, Baxter and UR5, that performed eight behaviors on 95 objects. We evaluate our framework on two tasks: object-property and object-identity recognition tasks. The results of our experiments show that KEMA learned using object identity-based correspondence consistently outperforms EDN in both tasks indicating transferring knowledge from robots to a shared latent space boosts the performance of the target robot. Furthermore, we propose a data augmentation technique independent of the learning task and show that using our data augmentation technique improves the models’ generalization and prevents overfitting.

II. RELATED WORK

Interactive object perception: Studies in psychology and cognitive science show that humans manipulate objects in multiple stages to extract information about their properties, such as texture, stiffness, temperature, and weight [20]–[22]. In addition, the human brain leverages a multisensory representation when recognizing object properties, enabling flexible generalizability to unknown contexts [23], [24]. Recent advances in intelligent robotics consider integrating multisensory information acquired by object exploration [10], [25]–[30], where one challenge is that the implicit knowledge acquired by one robot through interactive perception cannot be directly transferred to another robot: the unique nature of the robot’s embodiment drastically affects the sensed data distribution and resultant model that each robot learns. Whereas the focus of prior work has been limited to learning
from scratch for each robot [31]–[33], this is prohibitively expensive at scale, e.g., for a fleet of heterogeneous robots.

We propose a framework for transferring implicit knowledge about object properties from a source robot to a target robot.

Transferring knowledge of object properties: Recent work demonstrates that implicit knowledge from the interactive object perception can be transferred across sensor models and robots [10], [17], [31]–[34]. In [31], a robot performed interactive object perception to improve object category recognition. As implicit knowledge transfer was not the focus of that work, experiments were conducted on only a single robot. Moreover, whereas object properties may sometimes be the same for objects in different categories (e.g., bottles and cups can have similar colors, contents, and weights), their method encouraged unconstrained feature similarity based on object category alone, compromising prospects for transferring the features across robots or tasks. Our cross-robot transfer approach jointly learns to distinguish between different categories while leveraging learned similarities across properties. In [33], authors consider object categorization under a transfer learning paradigm, wherein an encoder-decoder network was used to generate a “target” robot’s features from a “source” robot’s learned representation. The authors use only a single robot in their experiments; however, so inherent challenges introduced by different robot morphologies remain to be studied. The approach in [17] was used to project features from 3 robots with different embodiments to a shared latent space for object-identity recognition. However, their experiments consisted of only simulated robots that recorded only one sensor modality (effort) during interaction with objects that varied in only one dimension (weight). To address these shortcomings, we collect a multisensory dataset using two real robots with different morphologies, enabling embodied interaction with objects and transfer knowledge about the object-identity recognition. We consider learning two projection functions. First, the projection function $F_{Rs→Rt}$, that projects the observation features from the source robot’s feature space to the target robot’s feature space. More specifically, $F_{Rs→Rt}: x^c_{Rs} \rightarrow x^c_{Rt}$, where $x^c_{Rs}$ is the projected features in the target robot’s feature space. Second, the projection function $F_{Rs→Z}$, that projects the observation features from each robot’s feature space to a shared latent feature space. More specifically, $F_{Rs→Z}: x^c_{Rs} \rightarrow z^c_{Rs}$ and $F_{Rs→Z}: x^c_{Rs} \rightarrow z^c_{Rs}$, where $z^c_{Rs} \in \mathbb{R}^{D_Z}$ and represents the shared latent features of size $D_Z$. In the first mapping, we train the target robot in its own feature space; for the second mapping, we train the target robot in the shared latent space.

We also consider two ways to build correspondences between the source and the target robots, for learning the projection functions. First, object-identity correspondence, in which the source-target pair corresponds to the same object identity. It is applicable when both robots have access to the same objects. Second, object-property correspondence, in which the source-target pair corresponds to the same object property. It is applicable when both robots operate in different environments and do not have access to identical objects but have access to objects with the same properties (e.g., red and blue bowls containing rice).
B. Projection to Target Feature Space

We propose using an Encoder-Decoder Network (EDN) [33] to train the projection function $F_{R_s ightarrow R_t}$, mapping observation features from the source robot’s feature space to the target robot’s feature space (Fig. 1A). First, encoder $f_\theta$ transforms the observation feature of the source robot $x_i^{R_s}$ into a fixed-size lower-dimensional vector $z_i^{R_s} \in \mathbb{R}^{D_z}$ of size $D_z$. Then, decoder $g_\phi$ uses this code vector $z_i^{R_s}$ to generate the predicted observation feature of the target robot $\hat{x}_i^{R_t}$. We denote this overall non-linear mapping as $F_{R_s ightarrow R_t} : x_i^{R_s} \mapsto \hat{x}_i^{R_t} = g_\phi(f_\theta(x_i^{R_s}))$, where $\theta$ and $\phi$ are network parameter weights of encoder and decoder, respectively. For training the EDN, we use a dataset of source-target feature pairs $(x_i^{R_s}, x_i^{R_t})$ for $N$ training samples. We optimize EDN parameters by minimizing root mean-squared error (RMSE) between real features observed by target robot $x_i^{R_t}$ and “generated” target features $\hat{x}_i^{R_t}$ obtained by applying the projection to the corresponding source features: $\arg\min_{\theta, \phi} \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i^{R_t} - \hat{x}_i^{R_t})^2}$. Given a trained EDN, we generate the target robot’s feature to transfer knowledge about the source robot’s additional experience; then, using a standard multi-class classifier, we can train the target robot to recognize object properties with the “generated” features.

C. Projection to Shared Latent Feature Space

The projection $F_{R_s \rightarrow Z}$ can be achieved through distribution alignment—organizing observation features from each robot’s feature space within a shared representation (Fig. 1B). We illustrate this mapping via Kernel Manifold Alignment (KEMA) [17], which constructs a set of domain-specific projection functions for each robot $F_R = [F_{R_s}, F_{R_t}]^T$, such that the examples of the same object property would locate closer while examples of different object properties would locate distantly. To compute the data projection matrix $F_R$, we minimize the cost related to the projection functions being too dissimilar: $\{F_{R_s}, F_{R_t}\} = \arg\min_{F_{R_s}, F_{R_t}} (C(F_{R_s}, F_{R_t}))$. Here, $C(\cdot) = \frac{1}{\text{dim}} (\mu \ast \text{GEO} + (1 - \mu) \ast \text{SIM})$, where the geometry of a domain, class similarity, and class dissimilarity are represented as GEO, SIM, and DIS, respectively. GEO is minimized to preserve the local geometry of each domain by penalizing projections in the input domain that are far from each other. SIM is minimized to encourage examples with the same object property to be located close to each other in the latent space by penalizing projections of the same object property mapped far from each other. DIS is maximized to encourage examples with different object properties to be located far apart in the latent space by penalizing projections of the different object properties that are close to each other. The parameter $\mu \in [0, 1]$ regulates the contribution of the GEO and the SIM terms. For more details on KEMA, please see [36]. Data in the latent feature space are comparable and can be used to train a standard multi-class classifier for different robots. The target robot can use this classifier to recognize properties of objects it has never interacted with.

D. Model Implementation and Training

Specific EDN architectures (e.g., transformers, dense convolutions, etc.) may be chosen according to the form of the data observations; in our experiments, we used an architecture that consists of three fully-connected layers for both encoder and decoder, with 1000, 500, 250 units, activation via Exponential Linear Units (ELU), and a 125-dimensional latent code vector. We use Adam [37] with a learning rate of $10^{-4}$ to compute gradients according to RMSE, over 1000 epochs. We used Radial Basis Function (RBF) for KEMA’s kernel function, with $\mu = 0.5$. We train the target robot’s recognition model via a multi-class SVM with the RBF kernel. For the EDN approach, this recognition model is trained using the “generated” features from the source robot and the real features of the target robot used to train the EDN; for the KEMA approach, this recognition model is trained using the shared latent features corresponding to both robots’ datapoints used to learn the KEMA projection function.
IV. Evaluation

A. Experimental Platform and Feature Extraction

1) Robots and Sensors: We collected our dataset using two robots: Baxter [38] and UR5 [39] (Fig. 1A). Baxter has dual 7-degree-of-freedom (DOF) arms and a 2-finger gripper. We used the left Baxter arm for the data collection. UR5 has 6-DOF and 2-finger Robotiq 85 gripper. Baxter had a PrimeSense camera mounted on its head, which captures 640×480 images, and an Audio-Technica PRO 44 microphone placed on its workstation. Baxter hand camera captures 480×300 images. UR5 had an Orbbec Astra S 3D Camera mounted on its frame, which captures 640×480 images, and a Seeed Studio ReSpeaker microphone placed on its workstation. We recorded data from 14 and 11 sensor modalities for Baxter and UR5, respectively. For more dataset details, such as sampling rate, please see: [https://github.com/gtatiya/Implicit-Knowledge-Transfer](https://github.com/gtatiya/Implicit-Knowledge-Transfer).

2) Exploratory Behaviors and Objects: Both robots perform 8 behaviors: look, grasp, pick, hold, shake, lower, drop, and push (Fig. 2). We chose these diverse behaviors because they can capture various object properties. Look is a non-interactive behavior in which robots record visual modalities (RGB, Depth, and Point-Cloud) from their head camera. All other behaviors are interactive, encoded as robot joint-angle trajectories. For all behaviors, Point-Cloud was recorded for the first 5 frames. Both robots explore 95 objects (cylindrical containers) that vary in 5 colors (blue, green, red, white, and yellow), 6 contents (wooden buttons, plastic dices, glass marbles, nuts & bolts, pasta, and rice), and 4 weights (empty, 50g, 100g, and 150g) shown in Fig. 1D. There are 90 objects with contents (5 colors x 3 weights x 6 contents) and 5 objects without any content that only vary by 5 colors.

3) Data Collection: While recording sensor data, robots perform all 8 behaviors in a sequence on the 95 objects, in round-robin fashion, to minimize any transient noise effects after a single trial on an object. Both robots perform 5 such trials on each object, resulting in 7,600 interactions, overall.

4) Feature Extraction: We used all interactive behaviors in our experiments (i.e., all behaviors listed above except look). We used audio, effort at the robot’s joints, and force at the robot’s end-effector in our experiments, as they play crucial roles in the human somatosensory system for recognizing non-visual object properties. For audio, we used librosa [40] to generate mel-scaled spectrograms of the audio wave files recorded by robots with FFT window length of 1024, hop length of 512, and 60 mel-bands. Then, a spectro-temporal histogram was computed by discretizing both time and frequency into 10 equally-spaced bins, where each bin consists of the mean of values in that bin. Effort and force data were discretized into 10 equally-spaced temporal bins for joints and axes, respectively. Thus, audio and force data are represented as 100 and 30 dimensional feature vectors, respectively. For Baxter and UR5, effort data is represented as 70 and 60 dimensional feature vectors, respectively. Fig. 2 visualizes both robots’ audio, effort, and force features when they perform shake behavior on a blue-marbles-150g object.

5) Data Augmentation: To improve model generalization, we increase the number of object trials through data augmentation: we compute each bin’s mean and standard deviation in the discretized representation of all object trials and sample $k = 5$ additional trials of each object. The rationale behind augmenting data by constraining on trials is to generate realistic data that is less likely to be impossible to produce in the real-world. Furthermore, this data augmentation technique is independent of the downstream task and can be applied for both object-property and object-identity recognition.

B. Evaluation

We evaluated performance of the projection methods: 1) EDN projects source robot features to a target robot’s feature space, and 2) KEMA projects individual robot features to a shared feature space. To learn both projections, we evaluate two ways to build correspondence between source-target data pairs: 1) object identity-based pairs, wherein both source and target robots interact with the same object identity (e.g., baxter-buttons-50g and ur5-buttons-50g); and 2) object property-based pairs, wherein source and target robots interact with objects that share a property (e.g., baxter-buttons-50g and ur5-dices-50g, wherein weight is same and contents are different). In both correspondence types, we use the same behavior and modality for both source and target robots. We consider 2 tasks: object property-recognition and object identity-recognition. In property-recognition, the target robot must recognize content and weight of the object it interacts with; there are 7 content classes and 4 weight classes, including an empty class. In object identity recognition, the target robot must recognize the specific object identity.

1) Object Property Recognition Task: As a baseline condition, we train the target robot using data in its own feature space. For the transfer condition, we train the target robot using features obtained by applying the projections. In training each projection, we use all 95 objects for the source robot and increment the number of objects the target robot interacts with, from 4 (for weight-recognition) and 7 (content-recognition), to 76 objects (80% of objects). The remaining 19 objects (20% objects) are held-out for testing target robot performance. We randomly-sampled 76 objects for incremental training and used remaining 19 for testing; we repeated this process 10 times, in both conditions. For best target robot performance in the baseline condition, we train using all 95 objects and evaluate on test objects in each fold. In all cases, we used all 5 trials of each object.

2) Object Identity Recognition Task: The baseline and transfer conditions of the object identity recognition task are the same as in the property recognition task. We evaluated the target robot’s performance to recognize 12 randomly-sampled objects from the 95 objects. When training each projection method, we used all 5 trials of each object for the source robot and increment the number of trial per object from 1 to 4 (80% trials) for the target robot. The remaining 1 trial (20% trials) of each object is held-out for testing the target robot’s performance. For both conditions, we performed 5-fold cross-validation such that each trial of
all fewer due to using projected features (obtained by interacting with accuracy delta \( \Delta \)) in our experiments, defined as: \( m \Delta A \) accuracy delta of the least \( m \) repeated 10 times to compute performance statistics.

The mean accuracy computed over 10 folds using these real data collected is \( 22.31 \pm 8.05 \). This learning process is the same as in our baseline condition, where the robot learns using its own objects. Now, we additionally use the 5 trials with data augmentation and train \( UR5 \) to recognize the object’s weight using 40 examples: 4 weights \( \times 5 \) trials + 5 augmented. The mean accuracy computed over 10 folds using these real and augmented data is \( 28.21 \pm 6.09 \); the increased accuracy shows that using data augmentation improves recognition performance. Since, we consistently observed improvements from augmentation, we only report the performance of our baseline and transfer conditions using augmentation.

### Object Property Recognition Results

For the object property-recognition task, we evaluated both projection methods by building correspondences based on object-identity and object-property. For EDN, we built object-identity correspondences by mapping each source robot’s object trial to all the target robot’s trials of that object. We built object-property correspondences by mapping each source robot’s object with a property of the recognition task to all the target robots’ objects with that property. For example, for the weight recognition task, a 50g object interacted by the source robot will be mapped to all the 50g objects interacted by the target robot. For KEMA, we build the manifold alignment using all 95 objects of the source robot and incrementally vary the number of objects the target robot interacts with, for both object-identity and object-property correspondences.

Consider another case where \( UR5 \) interacts with each object of each weight 5 times and learns to recognize the object’s weight using 20 examples (4 weights \( \times 5 \) trials). The mean accuracy computed over 10 folds using these 20 examples is \( 22.31 \pm 8.05 \). This learning process is the same as in our baseline condition, where the robot learns using its own objects. Now, we additionally use the 5 trials with data augmentation and train \( UR5 \) to recognize the object’s weight using 40 examples: 4 weights \( \times 5 \) real trials + 5 augmented. The mean accuracy computed over 10 folds using these real and augmented data is \( 28.21 \pm 6.09 \); the increased accuracy shows that using data augmentation improves recognition performance. Since, we consistently observed improvements from augmentation, we only report the performance of our baseline and transfer conditions using augmentation.

### V. RESULTS

#### Illustrative Example

Consider the case where a source robot (Baxter) and a target robot (UR5) perform pick behavior and record force signal. Baxter interacts with all 95 objects, and UR5 interacts with only 20 objects; both robots perform 5 trials on each object. We use Principal Component Analysis (PCA) to visualize the robots’ feature spaces (Fig. 3A and 3B) and plot object weights with different colors. In Fig. 3A we only plot Baxter’s features of the common 20 objects, for comparison to original and projected features shown in Fig. 3C. We project UR5-pick-force to Baxter-pick-force, via EDN with object identity-based correspondences, and visualize with PCA in Fig. 3C. Compared to Baxter’s space (Fig. 3A), projected features are more tightly clustered for different weights. We also generate the shared latent features using KEMA with object identity-based correspondences. We plot first 2 dimensions of latent features in Fig. 3D: data collected by both robots of 4 different weights are clustered together, indicating both robots’ data distribution is aligned efficiently.

We consider another case where UR5 interacts with one object of each weight 5 times and learns to recognize the object’s weight using 20 examples (4 weights \( \times 5 \) trials). The mean accuracy computed over 10 folds using these 20 examples is \( 22.31 \pm 8.05 \). This learning process is the same as in our baseline condition, where the robot learns using its own objects. Now, we additionally use the 5 trials with data augmentation and train UR5 to recognize the object’s weight using 40 examples: 4 weights \( \times 5 \) real trials + 5 augmented. The mean accuracy computed over 10 folds using these real and augmented data is \( 28.21 \pm 6.09 \); the increased accuracy shows that using data augmentation improves recognition performance. Since, we consistently observed improvements from augmentation, we only report the performance of our baseline and transfer conditions using augmentation.

#### 3) Evaluation Metrics

We used two metrics to evaluate the target robot’s recognition performance. First, we consider accuracy \( A = \frac{\text{correct predictions}}{\text{total predictions}} \times 100 \); the second metric is the accuracy delta \( \Delta A \), which measures the drop in accuracy due to using projected features (obtained by interacting with fewer objects) versus using the target robot’s own features (obtained by interacting with all objects). We compute mean accuracy delta of the least \( m \) number of object interactions in our experiments, defined as: \( m \Delta A = \frac{1}{m} \sum_{j=1}^{m} (A_{\text{all}} - A_{\text{projected}})^\% \), where \( A_{\text{all}} \) is the accuracy obtained using 100\% of the target robot’s data, \( A_{\text{projected}} \) is the accuracy obtained using projected features, and \( m = 10 \) for object property recognition, and \( m = 4 \) for object identity recognition. For both metrics, we use recognition accuracy computed as a weighted combination of all the behaviors and modalities used, based on their performance on the training data.

**Fig. 2:** Examples of (A) audio, (B) effort and (C) force features when Baxter and UR5 perform shake on a blue-marbles-150g object.

**Fig. 3:** Original sensory features of (A) Baxter and (B) UR5 for pick-force performed on 20 objects in 2D space, and (C) the projected features from UR5-pick-force to Baxter-pick-force projection using EDN, and (D) first 2 dimensions of corresponding features in the shared latent feature space generated using KEMA.

---

**Fig. 4** shows results of EDN and KEMA on the weight- and content-recognition tasks, where Baxter is the source robot and UR5 is the target robot: all transfer conditions for both approaches perform better than the baseline condition when the target robot interacts with fewer objects. As the target robot interacts with more objects, KEMA still performs better than baseline condition, and EDN performs comparable to baseline condition. Overall, results indicate that the proposed knowledge transfer methods can boost target robot
Thus, we used 12 randomly-sampled objects with unique force signal; a force signal would not be helpful to distinguish those objects. The audio signal produced while performing behaviors, as the different contents, it would be very crucial to listen to the task. For example, if two objects have the same weight but different contents, it would be very crucial to listen to the task. We emphasize building correspondences based on object-identity. We recognize task, we evaluated EDN and KEMA approaches for object identity-based and property-based correspondences.

For a robot to learn about implicit object properties, it must perform object exploration while processing various non-visual modalities. This process is costly across multiple robots as object feature representations are unique to a robot’s morphology. We proposed a framework for transferring implicit object property knowledge across heterogeneous robots and evaluated two projection methods, on two interactive perception tasks; results showed that learning on a target robot is accelerated through transfer from source robot, even if it explores fewer objects. Although our framework expedites learning on the less experienced target robot, there are some limitations. We encoded different behaviors in robots for object exploration. In future work, we plan to enable robots to learn behaviors to extract different object properties, autonomously. Moreover, we assumed that both source and target robots explored objects using the same sensorimotor context; thus, we used this same context while learning the projections. We plan to select sensorimotor contexts for learning projections more efficiently. Furthermore, we envision a scenario where more than two robots explore objects with additional properties, e.g., shape, size, material, and stiffness.

![Fig. 4: Accuracy results of the baseline and transfer conditions, EDN (left) and KEMA (right), on the weight (top) and content (bottom) recognition tasks, for Baxter (source) and UR5 (target).](image1)

**TABLE I:** Mean accuracy delta (m\(\Delta A\)) results of EDN and KEMA for object identity-based and property-based correspondences.

| Method (correspondence) | Weight | Content | Weight | Content |
|-------------------------|--------|---------|--------|---------|
| EDN (object-identity)   | 57.72  | 71.26   | 31.91  | 32.88   |
| EDN (object-property)   | 58.57  | 72.54   | 30.49  | 32.78   |
| KEMA (object-identity)  | 44.88  | 42.17   | 28.84  | 19.25   |
| KEMA (object-property)  | 51.88  | 47.53   | 34.88  | 22.60   |

**TABLE II:** Mean accuracy delta (m\(\Delta A\)) results of EDN and KEMA on the object identity recognition tasks.

| Method (correspondence) | UR5     | Baxter |
|-------------------------|---------|--------|
| EDN (object-identity)   | -14.29  | -11.67 |
| KEMA (object-identity)  | -40.67  | -12.00 |

![Fig. 5: Accuracy results of the baseline and transfer conditions on the object identity recognition tasks.](image2)
