2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Tactile Embeddings for Multi-Task Learning

Yiyue Luo[1], Murphy Wonsick[2], Jessica Hodgins[2], Brian Okorn[2]

***Abstract*— Tactile sensing plays a pivotal role in human perception and manipulation tasks, allowing us to intuitively understand task dynamics and adapt our actions in real time. Transferring such tactile intelligence to robotic systems would help intelligent agents understand task constraints and accurately interpret the dynamics of both the objects they are interacting with and their own operations. While significant progress has been made in imbuing robots with this tactile intelligence, challenges persist in effectively utilizing tactile information due to the diversity of tactile sensor form factors, manipulation tasks, and learning objectives involved. To address this challenge, we present a unified tactile embedding space capable of predicting a variety of task-centric qualities over multiple manipulation tasks. We collect tactile data from human demonstrations across various tasks and leverage this data to construct a shared latent space for task stage classification, object dynamics estimation, and tactile dynamics prediction. Through experiments and ablation studies, we demonstrate the effectiveness of our shared tactile latent space for more accurate and adaptable tactile networks, showing an improvement of up to 84% over the single-task training.**

I. INTRODUCTION

Humans leverage tactile sensations to extract a variety of information while completing tasks [1]. We use this data to perceive the tightness or friction associated with surface interactions as well as to estimate an object's weight or center of mass. This information can only be directly perceived from tactile sensing and understanding it allows us to estimate the progress of the task, classify failure states, and predict how much pressure needs to be applied at any given moment [2]. Imparting this understanding of the physical world to robotic agents could improve their effectiveness on a variety of tasks.

With the advancements in AI and computer vision algorithms, robotic agents have been able to learn visual features from human demonstrations, allowing them to perform delicate tasks via visual observations [3], [4]. However, modern robots are mostly tactile-blind and rely on visual information for perceiving their environments. While many object qualities can be inferred from visual data, this perception suffers when there is visual occlusion, which commonly occurs during robotic manipulation tasks. Additionally, force information can only be partially observed from visual sensors, whereas tactile sensors can directly perceive forces and pressures. This real-time tactile information can enable intelligent systems to directly understand task constraints, such as contact, slippage, and collision, as well as allow them to estimate the dynamics properties of manipulated objects and the environment, such as friction, mass, and inertia [5].

Recent efforts in sensor design [6], [7], [8] and tactile learning [9], [10], [11], [12] have imbued robots with some amount of tactile intelligence. Researchers have used tactile information for object classification [13], [14], grasping patterns discovery [15], and hand-object interaction estimations [16]. Nevertheless, challenges persist in translating insights gained from human tactile sensations to robotics due to the wide and diverse array of tactile sensor form factors, manipulation tasks, and learning objectives involved. To begin to address these challenges, we aim to develop a unified tactile latent space capable of accommodating the diversity of manipulation tasks encountered in robotics.

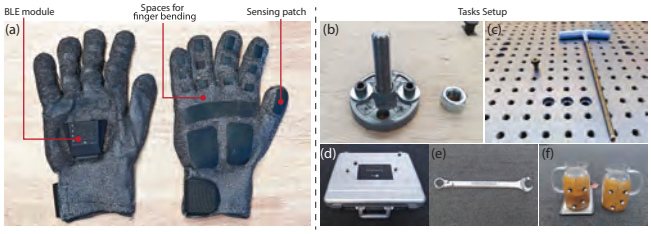

Fig. 1. **Tactile sensing gloves (a) and target tasks setup (b-f).** Tasks are hand tightening a bolt (b), using a T-handle screwdriver (c), opening a case (d), reorienting a wrench (e), and pouring a pitcher (f).

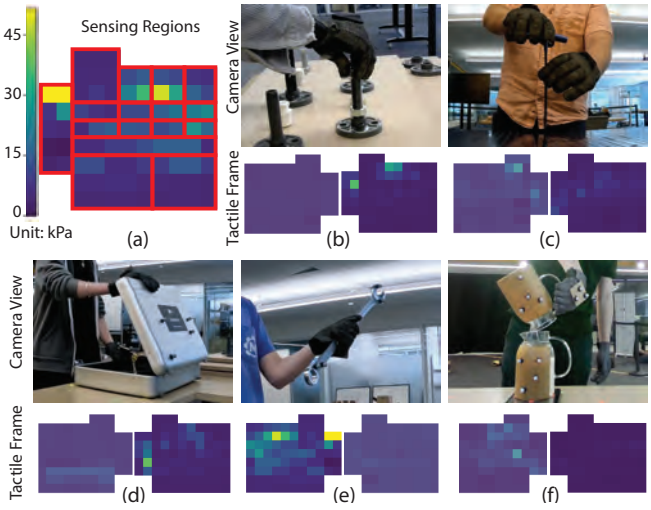

Fig. 2. **Mapping of tactile signal (a) and representative captured tactile frames during different manipulation tasks (b-f).** Each of the red boxes indicates a sensing region, such as the thumb, index finger tip, middle region of the index finger, lower region of the index finger, etc. Color correlates to pressure in the tactile frames, with yellow indicating high pressure and blue indicating low pressure.

[1]Authors are with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA. The work was done during an internship with the Boston Dynamics AI Institute. yiyueluo@mit.edu

[2]Authors are with the Boston Dynamics AI Institute, 145 Broadway Cambridge, MA 02139, USA. {mwonsick, jkh, bokorn}@theaiinstitute.com

In this work, we developed and evaluated the utility of shared latent spaces and cross-task prediction to tactile learning. To test this method, we used a commercially available sensor (Fig. 1) to collect diverse tactile data from ten people performing five different tasks: tightening a bolt, tightening a screw, opening a briefcase, reorienting a wrench, and pouring from a pitcher (Fig. 2 b-f). We trained a model to encode tactile sequences across all manipulation tasks into a shared embedding space. We leveraged the shared embedding space for three objectives: task stage classification, object dynamics estimation (estimating object orientations and flow rates), and tactile dynamics prediction (predicting future tactile frames). Through the use of unified tactile embeddings across manipulation tasks, we are able to identify task stages with an average accuracy of 88.2%, estimate object orientation with an average angular error of 6.04°, estimate the weight of poured rice with an average weight error of 44 g, and predict future tactile frames with an average pressure error of 0.262 kPa. We show that the use of this shared embedding space improves task classification and dynamics estimation performance through a series of experiments and ablation studies. Furthermore, we performed additional ablation studies to determine the importance of both temporal and spatial sensing resolution as well as the relative informativeness of individual sensing regions. We believe these shared tactile embeddings are the first step in transferring tactile learning to future robot embodiments.

II. RELATED WORK

A. Tactile Sensing

Advances in materials, electronics, and manufacturing techniques have enabled the development of a wide variety of tactile sensors using diverse sensing mechanisms. Commercial tactile sensors, such as Tekscan Grip system [17] and BioTac [18] have been widely used to facilitate robot manipulation with real-time tactile feedback. While these sensors are easily mounted to robotic grippers, they are not suitable for wearable sensors. Gelsight [19], and other vision-based tactile sensors [20], [21], [22], embed a camera and LEDs in transparent silicone with a reflective coating. The 3D shape and texture of the contacted surfaces are obtained through the reflection of internal light sources. Coupled with computational pipelines, vision-based tactile sensors have been used to predict geometry and slip [23], object properties [24], and liquid dynamics [25] during robot manipulation. These vision-based tactile sensors obtain high sensing resolution but suffer from the limitation of bulky designs which restricts the size, area, and complexity of surfaces to which they can be applied. This makes them unsuitable for capturing human demonstrations. Resistive/capacitive-based tactile sensors utilize the change of resistance/capacitance in materials when subjected to pressure, effectively transforming pressure stimuli into electrical signals through a coupled readout circuit. Such tactile sensors are made by aligning orthogonal electrode matrices over force-sensitive films. They can be easily scaled up to large areas and conform to complex surfaces. This has allowed resistive tactile sensing matrix to be fabricated into wearable sensors [26], [27], smart carpets [28], and robot manipulator coverings [26], [29] for use in object classification, human-object interactions characterization, and human-robot collaboration. In this work, we leverage commercially available capacitive-based tactile sensing gloves [30] to capture tactile frames from humans as they complete several different manipulation tasks.

B. Shared Embedding Space

Computer vision has long used shared embedding spaces and pretrained features to improve performance. Pretrained ImageNet [31] and COCO [32] features have been used to improve model robustness [33], and have formed the basis of saliency estimation [34], object pose [35] and image correspondence estimators [36], and robotic manipulation algorithms [37]. Cross-task learning has also been used to produce more robust features. XSkill [38] learned a cross-embodiment, cross-skill embedding space to transfer between human and robotic embodiment. More recently, image foundation models [39], [40] have used extremely large datasets and multiple training objectives to train massive models capable of producing state-of-the-art results in multiple domains. In this work, we evaluate the efficacy of similar cross-task training and shared embedding spaces for tactile learning, though we still lack the amount of tactile data required to train a tactile foundation model.

C. Human Demonstrations

Capturing large datasets of human demonstrations has emerged as a promising approach to impart complex skills and behaviors to robots through the observation of humans performing similar tasks. Some notable works have explored the use of vision as a means of transferring skills from human demonstration to robots through virtual teleoperation [41], [42] and motion-mirroring [3], though these demonstrations lacked any tactile or haptic information. More recently, researchers have started to capture tactile information from human demonstrations. Zhang [27] collected tactile information from simple dynamic human-object interactions, such as waving and balancing objects. ActionSense [43] captured a wide variety of human-centric information from people completing kitchen-based activities, including tactile data, body pose, and eye tracking. These datasets, however, use custom tactile sensors, making it difficult to add new activities or sensing modalities to them. In our work, we use commercial, off-the-shelf tactile gloves to evaluate what task information can be predicted from human tactile data and to study the utility of shared latent space to these predictions.

III. METHOD

To evaluate the utility of a shared tactile embedding space, we first require a cross-task dataset to train and analyze our methods. To this end, we designed a data collection pipeline to capture tactile frames in conjunction with synchronized visual imagery and object poses, both of which will be used to produce ground truth annotation. Using this dataset, we

trained and evaluated our shared tactile embedding framework on the objectives of task stage classification, object dynamics estimation, and tactile dynamics prediction. The data collection methodology and tactile learning module are described in more detail below.

A. Data Collection

Five manipulation tasks (Fig. 2) were selected to gather tactile data:

- Tightening a bolt using only hands: This task encompasses 4 stages - lifting the bolt, inserting it into the screw, threading, and finally releasing the bolt after tightening (Fig. 1b and Fig. 2b).
- Tightening a screw using a T-handle screwdriver: This task involves 5 stages - picking up the T-handle screwdriver, inserting it into the screw, tightening the screw, releasing the screwdriver from the secured screw, and placing the screwdriver down (Fig. 1c and Fig. 2c).
- Opening a case: The task is comprised of 4 stages - unlocking the right clip, unlocking the left clip, lifting the case lid, and releasing the lid once lifted (Fig. 1d and Fig. 2d).
- Reorienting a wrench: This task unfolds over 6 stages - picking up the wrench, rotating it to the left, straightening it from the left orientation, rotating it to the right, returning it to the starting right orientation, and setting the wrench down (Fig. 1e and Fig. 2e).
- Pouring from a pitcher: The task includes 4 stages - lifting the pitcher, tilting it, straightening the pitcher, and placing the pitcher in its original position (Fig. 1f and Fig. 2f).

These tasks were selected since they feature delicate task manipulation, exemplified by tasks of tightening bolts and screws, and incorporate intricate dynamics, as seen in tasks like rotation and pouring. These attributes stand out as potentially useful features in the realm of robot manipulation.

Tactile data was captured for all tasks via a commercially available capacitive-sensing tactile glove [30]. The tactile gloves contain 130 individual capacitive sensing elements spread across the palm and finger, shown in Fig. 1, which wirelessly transmit real-time pressure readings at 30 Hz (Fig. 2). The arrangement of sensing patches allows users to manipulate objects and perform tasks in a relatively natural manner. Object orientation for the pitcher, wrench, and case lid are captured via a fiducial-based motion capture system (OptiTrack). The pouring flow rates are captured using a digital scale (SparkFun OpenScale). These values are used as ground truth for object dynamics estimation. Synchronized videos were also captured using two cameras for task-stage annotation.

For each task, participants were initially asked to perform the task without any guidance, allowing them to become acquainted with the intricacies of the tasks. Then, participants were directed to execute the tasks following the predefined sequences, as defined above, using only their dominant hand, followed by a separate session using only their non-dominant hand for 3 - 5 rounds based on the time each participant spent on the specific task.

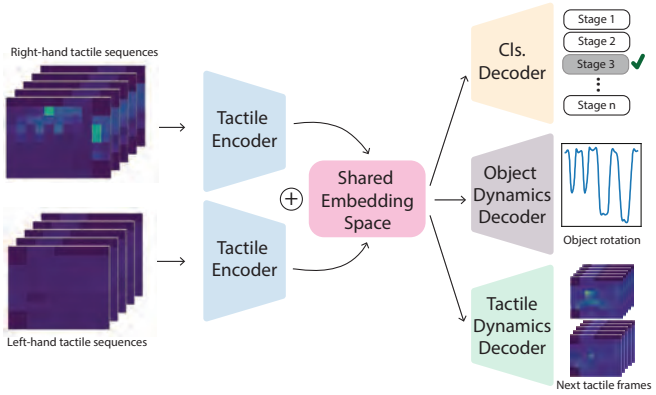

Fig. 3. **Model Diagram.** Our model takes in a sequence of tactile frames captured by the left and right sensing glove, and encodes them into the shared embedding space via the tactile encoders. Individual task classification decoders, object dynamics decoders, and tactile dynamics decoders output the corresponding prediction respectively from this shared embedding space.

Ten participants were recruited and more than 100,000 tactile frames were collected among all participants. Fig. 2 showcases representative tactile imprints featuring each of the target tasks. Signals from each of the sensing units were mapped to pixelized images, which were used as input to our model. Unused portions of these tactile images are set to zero. The recorded object orientation, pouring rate, and visual frames were synchronized with the tactile frame using a manual calibration procedure. We first align characteristic frames from each sensing modality (usually a clap in front of the camera), and then synchronize all the sequential frames by matching their logged timestamps with appropriate offsets. Ground truth for task stage classification was manually annotated by referencing the synchronized visual frames. The data was split into training, validation, and testing sets with the ratio of 80%, 10%, and 10%, respectively, with results reported on the test set. Each set was generated using completed task demonstrations, ensuring frames from individual demonstrations were not separated.

B. Model

As illustrated in Fig. 3, our model utilizes a conventional image encoder-decoder structure, where a sequence of tactile frames is taken as input and encoded into a shared latent space via convolutional layers. The model outputs three categories of parameters for each of the 5 tasks, including classification of the task stages, estimation of object dynamic parameters, such as object orientations and pouring flow rate (object dynamics), and predictions for future tactile frames (tactile dynamics). The outputs for each category of estimates and each task are separately computed using their own individual decoder but all share a single encoder.

Each tactile encoder consists of 5 layers of 2D convolutional layers with 3×3 kernels and 1×1 paddings. The outputs from each of the tactile encoders are flattened and concatenated to form our shared embedding space with a dimension of 2048. Each of the decoder branches consists

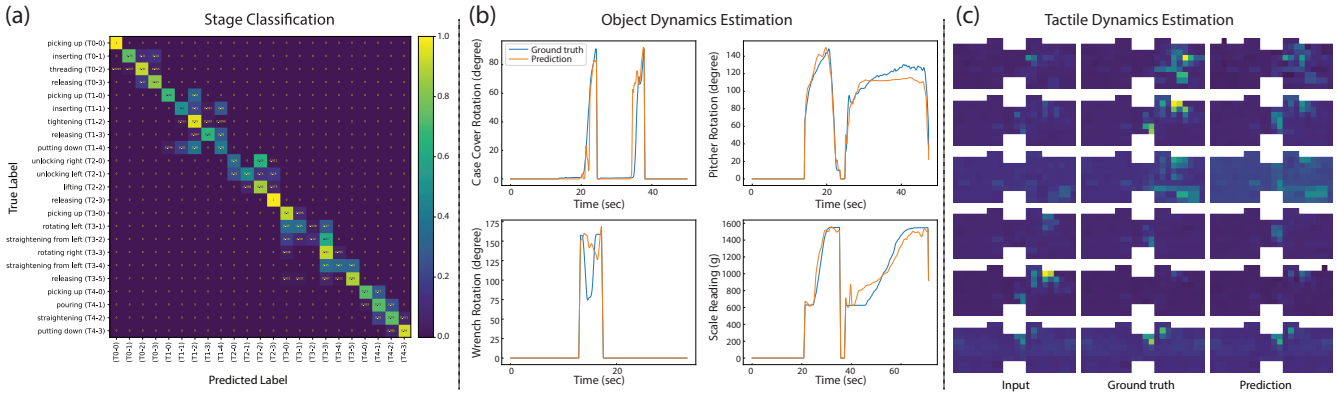

Fig. 4. **Qualitative results.** (a) the confusion matrix on 23 task stages across 5 manipulation tasks. Yellow represents high confusion and blue low confusion. (b) An example of ground truth (blue) and predicted (orange) object rotation and flow rate for object dynamics prediction. (c) The input, ground truth, and predicted tactile frames from tactile dynamics prediction.

of 5 fully connected layers. All convolutional and fully connected layers are followed by ReLU activation functions.

We used 40 contiguous tactile frames, $n-40$ to n, from each glove as input. This is equivalent to approximately 1.3 seconds of interaction time. For stage classification, we identify the current task stage at the n^{th} frame. For object dynamics regression, we predict the relevant object dynamic parameters for frames $n-10$ to n. We find that predicting a sequence of parameters produces significantly more accurate results than just predicting the parameters at the n^{th} frame. We believe this improvement is due to the continuity of the object motion, allowing the network to learn more complex, multi-frame trajectories that correspond to meaningful human actions. For the tactile dynamics prediction, we predict the tactile frames from $n+20$ to $n+40$. The model exhibits improved performance when it excludes the neighboring 20 future frames (from n to $n+20$). This exclusion strategy helps prevent the model from getting trapped in local minima, where it predicts identical to the last frame of the input tactile sequences.

For training, we use the weighted sum of three Mean Squared Error (MSE) losses between the predicted parameters from each decoder branch and the ground truth values:

$$\mathcal{L} = \sum_{K} w_K \mathcal{L}^K, \tag{1}$$

where K represents the three output categories, i.e. stage classification, object dynamics estimation, and tactile dynamics prediction, $\mathcal{L}_k$, represents each output category loss, and w_K defines their corresponding weighting. We use 1, 1, and 100 for the weights of stage classification, object dynamics estimation, and tactile dynamics prediction, respectively. The loss, $\mathcal{L}_k$ for each of these output categories is defined as:

$$\mathcal{L}^K = \frac{1}{N} \sum_{i=1}^{N} \| A_i^K - \hat{A}_i^K \|, \tag{2}$$

where N denotes the number of frames, and A_i^K and $\hat{A}_i^K$ represent the ground truth and the predicted value from each output category, respectively. While the use of MSE loss to train a classifier is nonstandard, we find that this loss achieves more accurate results when compared to the more traditional log likelihood loss for this setting. We optimize all parameters using the Adam optimizer [44] with a learning rate of $1e^{-3}$ and a batch size of 128.

IV. RESULT

When predicting the current stage for each task, we achieved an average accuracy of 88.2% $\pm$ 10% for stage classification. The full confusion matrix is demonstrated in Fig. 4 a. We find that our model exhibits a notable tendency to misclassify the various stages during the task of wrench reorientation. For instance, it mistakenly classifies the stage of straightening from the left as rotating to the right. This behavior is likely due to the similarity in tactile sequences between these specific stages.

For object dynamics prediction, we achieve a mean angular error of 6.04° $\pm$ 0.026° when predicting object orientation, and a weight error of 44 $\pm$ 1.2 g for pouring weight estimation. As a reference, the pitcher, when filled with the entire quantity of rice, weighs approximately 1.6 kg, (2.75% absolute difference with respect to the maximum weight). Qualitative results can be found in Fig. 4 b. Estimates of the wrench's rotation during reorientation yield a noticeably significant margin of absolute difference. This can likely be attributed to the fact that users tend to execute wrench rotations with greater vigor and at greater speeds as compared to the rotations of the pitcher and the briefcase (typically completing a 90° rotation in under half a second). This abrupt and substantial change in rotation angle introduces increased complexity for the predictive model.

Finally, when predicting future tactile frames, we obtain a mean pressure error of 0.262 ± 0.179 kPa as compared to the ground truth pressures, with the maximum recorded pressure of 45 kPa. Qualitative results can be found in Fig. 4 c. We believe that this tactile dynamics prediction can be used to predict how a robot should interact with an object, how much pressure it should exert, and where it should exert this pressure during manipulation tasks.

A. Utility of Shared Embedding Space

To evaluate the utility of our shared latent space, we perform an ablation study over the types of tasks and types of

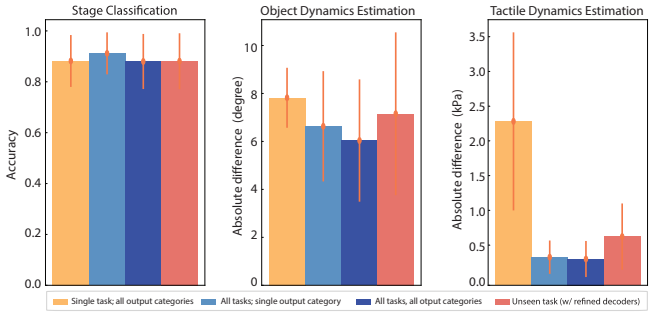

Fig. 5. **Utility of shared latent space.** Results on task stage classification (higher is better), object dynamics estimation (lower is better), tactile dynamics prediction (lower is better) from models with encoders trained for individual output categories (orange), models with encoders trained for individual output categories but shared over all tasks (light blue), and models with encoders trained with the full datasets (all tasks, and all output categories), dark blue. The generalization of shared latent space to unseen tasks is also investigated by pre-training the shared embedding space with 4 out of the 5 tasks and refining the specific decoder branches for the left-one-out task (red). Error bars indicate standard deviation.

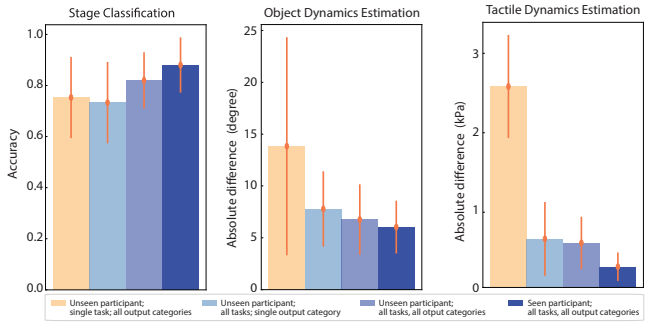

Fig. 6. **Generalization to unseen participants.** Results on unseen participants using models with encoders trained for individual output categories (orange), models with encoders trained for individual output categories but shared over all tasks (light blue), and models with encoders shared over all tasks and outputs (medium blue). Results from a model with encoders shared over all tasks and outputs trained with all participants is included as reference, (dark blue). Error bars indicate standard deviation.

learning objectives used to train the embedding space, shown in Fig. 5. We compare our model, trained with unique encoders for each individual manipulation task over all output categories (stage classification, object dynamics estimation, and tactile dynamics prediction), shown in orange, to a model with uses unique encoders for each output category, but shares these encoders over all tasks, (light blue bars), and a model whose encoder is shared over all task and all output categories (dark blue bars). The datasets for each setting are re-sampled to have approximately the same number of total training examples across all settings to mitigate the effects of purely having more data. While we find that the greatest improvement comes from cross-task learning (light blue bars), generating an improvement of 22.8% and 84.0% for object dynamics estimation and tactile dynamics prediction, respectively, as compared to the fully isolated case (orange), we do see additional improvement of 8.9% and 6.5% for object dynamics estimation and tactile dynamics prediction, respectively, when also sharing the embedding space across output categories.

We find that task stage classification demonstrates comparable performance across all models regardless of training strategy. However, for more demanding objectives, like object dynamics estimation and tactile prediction, the models trained with shared tactile embeddings exhibit significant reductions in estimation error.

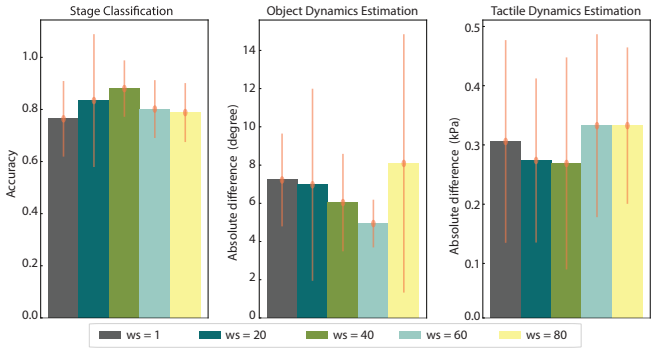

Fig. 7. **Ablation study on the temporal window size of tactile input frames.** Results include task stage classification accuracy (higher is better), the absolute difference between predicted object orientation and the ground truth captured by the motion capture system (lower is better), and the absolute difference between predicted future tactile frames and the ground truth (lower is better). The average results over all tasks are reported. Error bars indicate standard deviation.

Further, we tested the utility of this embedding space as pre-trained features for unseen tasks. We first pre-trained an encoder using data from four out of the five tasks. We then freeze the encoder weights and train a new decoder on the held-out task, represented by the red bars in Fig 5. The results achieved using the pre-trained embedding for unseen tasks, while slightly inferior to those of models trained over all tasks, outperform a model whose encoder was trained only on the held-out task by 8.5% and 70.3% for object dynamics estimation and tactile dynamics prediction, respectively.

Additionally, we explored the generalization capability of the shared tactile embedding space to unseen participants, shown in Fig 6. Compared with predictions from seen participants (dark blue), results on unseen participants (medium blue) show a lower performance by an average of 20.3% for all three output categories. However, when both are tested on unseen participants, the model with shared embeddings still outperforms the ones with task or output-specific encoders (orange and light blue). This improvement underscores the effectiveness of our proposed shared tactile embeddings in generalizing to novel tasks and participants.

B. Ablation Studies on Tactile Data

We investigate the relative importance of our design decisions using a series of ablation studies. Results over all tasks are averaged for each reported metric. To evaluate the sensitivity of our method to the duration of the tactile signal, we evaluate our method while varying the temporal window size of our input tactile sequences. As demonstrated in Fig. 7, the input window size of 40 frames (around 1.3 seconds) obtains the overall best performance. As a reminder, higher classification accuracy is better whereas lower absolute difference is better for dynamics estimation and prediction.

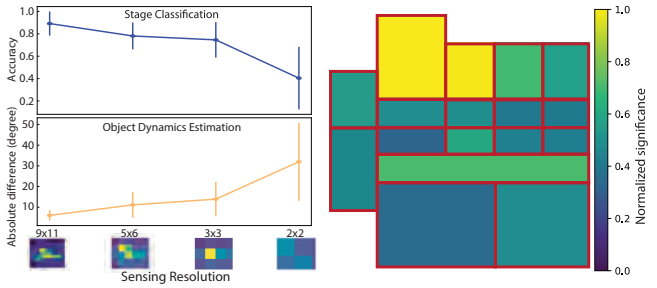

Fig. 8. **Ablation study on tactile sensing resolution and tactile sensing regions.** Left: Task stage classification accuracy (top) and object orientation estimation error (bottom) performance drops as we reduce the tactile sensing resolution. Right: Importance of tactile regions to task stage classification. In general, tactile sensing from fingertips and the top palm obtains higher importance.

To determine the importance of tactile resolution, we train and evaluate our model while progressively reducing the input resolution from 9×11 to 5×5, 3×3, and 2×2. As we decreased the effective sensing resolution, we observed a corresponding decrease in stage classification accuracy and an increase in absolute difference on object rotation. This suggests that higher tactile sensing resolution may be advantageous for improving performance, though the importance of this resolution is far more pronounced when the reduction is extreme.

To assess the importance of each individual region of the tactile array, we pre-trained a model with the full dataset (with full tactile images) and evaluated the stage classification accuracy using masked tactile images as input during testing. We zero mask portions the input tactile images, taking into account the hand's anatomical structures. More specifically, we individually mask out the sensing areas located on the distal, middle, and proximal phalanges of each finger, as well as those on the top, lower-left, and lower-right areas of the palms. We report normalized significance, as one minus prediction accuracy, normalized to between zero and one, with zero being the unmasked results. As shown in Fig. 8, the fingertip regions emerge as the most influential factor affecting performance. This aligns with our intuition given that the fingertips were predominantly employed for interacting with the object, i.e. bolts, during data capture. Furthermore, the top palm also demonstrates its importance compared with the lower palm, as it experiences notable pressure during the manipulation of larger objects, such as wrenches and pitchers.

C. Visualization of Shared Embedding Space

We visualize the shared embedding space via t-SNE [45]. As illustrated in Fig 9, it is evident that tactile frames obtained from each of the tasks exhibit discernible clustering patterns within the projected three-dimensional space. When examining the features associated with an individual task demonstration, the stages of each specific task seamlessly trace a continuous path within this projected space. Additionally, it is noted that the distinctiveness of the traces formed by tactile sequences from different individuals effectively encapsulates the unique behavioral characteristics of each participant. Learning more about the dynamics of these traces may better allow us to map tactile information between participants and eventually between sensor morphologies.

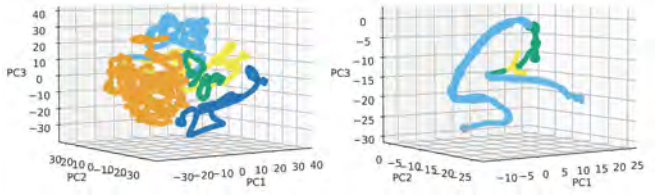

Fig. 9. **Visualization of the shared embedding space.** Features are projected into 3D space via t-SNE. Tactile features captured from different task manipulations form distinctive clusters. The orange, light blue, green, yellow, and dark blue dots represent data points from the task of tightening a bolt, tightening a screw, opening a briefcase, reorienting a wrench, and pouring from a pitcher respectively. Tactile features captured from different stages during the pouring task form unique traces in the projected space. The orange, light blue, green, and yellow dots represent data points from the stages of picking up the pitcher, pouring, straightening the pitcher, and putting down the pitcher.

V. DISCUSSION

We recognize the need for even larger datasets to further strengthen our models. Additionally, to capture fine-grained details in certain scenarios and further enhance our data capture capabilities, sensors with higher resolution may be required.

In future work, we plan to extend our shared embedding space to cover a variety of tactile sensing modalities and hand morphologies. In particular, we aim to map the shared tactile embedding space from tactile sensing gloves to various tactile sensors used by different robotic manipulators, including parallel grippers and robotic hands, to transfer tactile information gained from human demonstrations to robotic agents. Furthermore, we aspire to fuse tactile information with other sensory modalities, such as vision, audio, and language. This multi-modal approach has the potential to unlock new avenues for a tactile foundation model, enabling robust perception and interaction in robotics.

VI. CONCLUSION

In this paper, we evaluated the effectiveness of a unified tactile embedding space across diverse manipulation tasks and output categories. To quantify the utility of this approach, we collected a tactile dataset using 10 people performing five unique manipulation tasks. We showed that training tactile learning models over a variety of tasks and output categories far outperforms models trained on only a single task. Further, we evaluated the relative importance of spatial and temporal resolution through a series of ablation studies. We see this work as a stepping stone toward generalized tactile learning and transferring tactile knowledge from human demonstrations to robot actuation.

ACKNOWLEDGMENT

We thank Osman Dogan Yirmibesoglu for reviewing the manuscript.

REFERENCES

[1] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345–359, 2009.

[2] M.-M. Mesulam, "From sensation to cognition." *Brain: a journal of neurology*, vol. 121, no. 6, pp. 1013–1052, 1998.

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *RSS*, 2023.

[4] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[5] S. Sundaram, "Robots learn to identify objects by feeling," *Science Robotics*, vol. 5, no. 49, p. eabf1502, 2020.

[6] K. Senthil Kumar, P.-Y. Chen, and H. Ren, "A review of printable flexible and stretchable tactile sensors," *Research*, vol. 2019, 2019.

[7] H.-L. Cao and S.-Q. Cai, "Recent advances in electronic skins: material progress and applications," *Frontiers in Bioengineering and Biotechnology*, vol. 10, p. 1083579, 2022.

[8] Y. Luo, C. Liu, Y. J. Lee, J. DelPreto, K. Wu, M. Foshey, D. Rus, T. Palacios, Y. Li, A. Torralba, *et al.*, "Adaptive tactile interaction transfer via digitally embroidered smart gloves," *Nature Communications*, vol. 15, no. 1, p. 868, 2024.

[9] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[10] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5580–5588.

[11] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[12] J.-T. Lee, D. Bollegala, and S. Luo, ""touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4276–4282.

[13] T. Corradi, P. Hall, and P. Iravani, "Bayesian tactile object recognition: Learning and recognising objects using a new inexpensive tactile sensor," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 3909–3914.

[14] J. M. Gandarias, A. J. Garcia-Cerezo, and J. M. Gómez-de Gabriel, "Cnn-based methods for object recognition with high-resolution tactile sensors," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6872–6882, 2019.

[15] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, 2019.

[16] L. Yang, B. Huang, Q. Li, Y.-Y. Tsai, W. W. Lee, C. Song, and J. Pan, "Tacgnn: Learning tactile-based in-hand manipulation with a blind robot using hierarchical graph neural network," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3605–3612, 2023.

[17] "Measure grip forces," Mar 2018. [Online]. Available: https://www.tekscan.com/measure-grip-forces

[18] J. A. Fishel and G. E. Loeb, "Sensing tactile microvibrations with the biotac — comparison with human sensitivity," in *2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 1122–1127.

[19] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[20] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 1045–1051.

[21] B. Fang, F. Sun, C. Yang, H. Xue, W. Chen, C. Zhang, D. Guo, and H. Liu, "A dual-modal vision-based tactile sensor for robotic hand grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4740–4745.

[22] H. Sun, K. J. Kuchenbecker, and G. Martius, "A soft thumb-sized vision-based sensor with accurate all-round force perception," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 135–145, 2022.

[23] S. Dong, W. Yuan, and E. H. Adelson, "Improved gelsight tactile sensor for measuring geometry and slip," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 137–144.

[24] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a gelsight tactile sensor," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 951–958.

[25] H.-J. Huang, X. Guo, and W. Yuan, "Understanding dynamic tactile sensing for liquid property estimation," *arXiv preprint arXiv:2205.08771*, 2022.

[26] Y. Luo, Y. Li, P. Sharma, W. Shou, K. Wu, M. Foshey, B. Li, T. Palacios, A. Torralba, and W. Matusik, "Learning human–environment interactions using conformal tactile textiles," *Nature Electronics*, vol. 4, no. 3, pp. 193–201, 2021.

[27] Q. Zhang, Y. Li, Y. Luo, W. Shou, M. Foshey, J. Yan, J. B. Tenenbaum, W. Matusik, and A. Torralba, "Dynamic modeling of hand-object interactions via tactile sensing," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2874–2881.

[28] Y. Luo, Y. Li, M. Foshey, W. Shou, P. Sharma, T. Palacios, A. Torralba, and W. Matusik, "Intelligent carpet: Inferring 3d human pose from tactile signals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 255–11 265.

[29] L. Zlokapa, Y. Luo, J. Xu, M. Foshey, K. Wu, P. Agrawal, and W. Matusik, "An integrated design pipeline for tactile sensing robotic manipulators," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3136–3142.

[30] [Online]. Available: https://pressureprofile.com/body-pressure-mapping/tactile-glove

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[33] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *International conference on machine learning*. PMLR, 2019, pp. 2712–2721.

[34] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," in *International Conference on Learning Representations (ICLR 2015)*, 2014, pp. 1–12.

[35] A. Ravi, "Pre-trained convolutional neural network features for facial expression recognition," *arXiv preprint arXiv:1812.06387*, 2018.

[36] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," *Advances in neural information processing systems*, vol. 29, 2016.

[37] L. Yen-Chen, A. Zeng, S. Song, P. Isola, and T.-Y. Lin, "Learning to see before learning to act: Visual pre-training for manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7286–7293.

[38] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "Xskill: Cross embodiment skill discovery," in *Conference on robot learning*, 2023.

[39] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[41] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," in *Conference on robot learning*. PMLR, 2017, pp. 357–368.

[42] J. DelPreto, J. I. Lipton, L. Sanneman, A. J. Fay, C. Fourie, C. Choi, and D. Rus, "Helping robots learn: a human-robot master-apprentice model using demonstrations via virtual reality teleoperation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 226–10 233.

[43] J. DelPreto, C. Liu, Y. Luo, M. Foshey, Y. Li, A. Torralba, W. Matusik, and D. Rus, "Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 800–13 813, 2022.

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

[45] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.