Self-Supervised Visuotactile Representation Learning for Dexterous Manipulation

Antonia Bronars (bronars), Neha Sunil (nsunil)

Abstract

The ability to integrate complementary information from vision and touch is a long-standing goal in robotic manipulation. While vision provides global information about an object's position and orientation, touch provides local signals of contact geometry and forces, which are important to supervise contact-rich interactions, especially in the presence of visual occlusions. In this work, we combine these modalities for two different robotics-minded goals: object classification and joint representation learning. To learn a joint representation, the visual and tactile images are embedded into a shared latent space using a cross-modal contrastive loss trained in a self-supervised manner. We implement and compare each network architecture with ResNets and Vision Transformers. Ultimately, we obtained the highest classification performance with ResNets and visuotactile data, achieving 95.3% *classification accuracy. We were also able to achieve semantic clustering in our learned joint representations. Code, datasets, and trained weights are available here.*

A. Introduction

Humans seamlessly integrate complementary information from vision and touch to enable more robust perception of the environment. While vision provides rich information about the appearance and spatial layout of objects in the scene, touch provides local information about contact geometry and forces during object interaction. Tactile information becomes especially useful in the presence of occlusions, whether from other objects in clutter or the gripper while grasping the object. We use Gelsight [15], a camera-based tactile sensor for high resolution tactile information. Figure 1 demonstrates that vision can help with initially grasping the USB connector and coarse alignment, but the high resolution tactile imprint is much more useful in a closed-loop controller to precisely plug it in. Consequently, there has been extensive research in robotic manipulation regarding effective ways to fuse visual and tactile modalities [9].

At the same time, Transformers have been demonstrably effective at fusing multi-modal data including text, images, and audio [1, 13]. The key advantage of the Transformer architecture is the attention mechanism, which allows the network to selectively focus on different portions of the input sequence when computing the output representation. This ability is especially helpful for global information processing when parsing through large quantities of data. Convolutional networks like ResNet [6] only have a field of view the size of the kernel, while vision transformers can have a much larger receptive field. In multi-modal representation learning, cross-modal attention, in addition to self-attention, provides a natural way to fuse modalities, which may contain very disparate information.

We propose to leverage Transformers for visuotactile representation learning by fine-tuning Vision Transformers [4] with vision data (overhead depth images of the object) and tactile data (Gelsight tactile sensing images at the contact interface). We compare the performance of Transformers with the same network architectures built with ResNet encoders [6]. Specifically, we combine separate encoders for object classification and learn a joint lower-dimensional

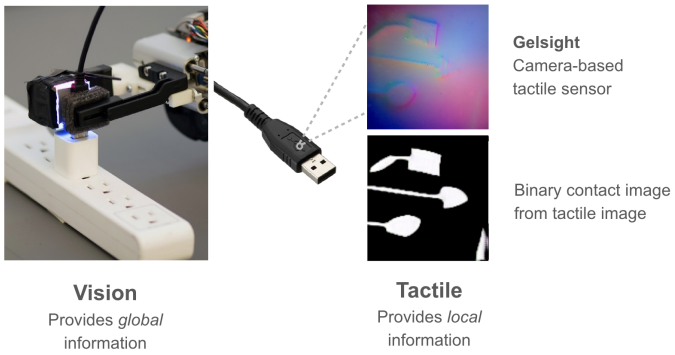

Figure 1. Visual and tactile modalities in the robotic task of plugging in a USB connector. While vision is useful when initially grasping the cable and coarsely aligning it with the socket, the gripper occludes the visual image, making precise object localization difficult. The tactile signal, on the other hand, allows for high resolution pose estimation that can be used in a closed-loop controller. We use Gelsight sensors for tactile information and threshold the depth reconstruction for a binary contact image.

representation using contrastive learning. Classification is useful in manipulation tasks in order to identify objects of interest in clutter and because different objects often require different manipulation skills or control parameters. A fused visuotactile latent space representation can be used to transfer state or goal specifications between modalities. Moreover, the lower-dimensionality of this representation compared to two input images simplifies policy learning for tasks that are dependent on both modalities.

B. Related Work

Previous approaches of fusing visual and tactile data combine overhead images with low-resolution, wrist-mounted force/torque (F/T) tactile sensing. [9] uses separate encoders for RGB images (a six-layer CNN), F/T sensor readings, and proprioceptive inputs, then concatenates the three feature vectors before passing it through a 2-layer MLP to produce the final fused representation. They train the representation in a self-supervised manner by predicting action-conditioned optical flow, contact state, and temporal alignment of the next state, given a current state representation/action pair. [3] uses similar self-supervised training signals as [9], but uses Vision Transformers [4] as the encoder architecture instead. This change allows them to achieve better performance in terms of sample efficiency and accuracy, when doing reinforcement learning on top of the representations to solve simulated manipulation tasks. Finally, [14] learns visuotactile representations of deformable objects, to predict object dynamics when subject to external forces.

In recent years, image-based tactile sensors [8, 10, 15] have become an increasingly popular tactile sensing paradigm, due to their high-resolution data stream. Image-based tactile sensors provide information about the contact geometry, normal, and shear forces, all of which are important signals to guide dexterous manipulation. [12] combines RGB and tactile image representations using maximum covariance analysis for cloth texture recognition. [11] learns a joint latent space by learning to synthesize RGB images from tactile images, and vice versa, using a conditional GAN. [7] learns a joint representation space for image-based tactile sensor images and collocated RGB images using contrastive learning, and demonstrates its efficacy in a range of tactile classification and control tasks for deformable objects.

While vision data is easy to simulate, tactile data is more challenging to generate. Therefore much of the previous work uses data collected from real robots, which is time-consuming and expensive. Since we can simulate tactile signals relevant to our sensor [2], we are able to generate our data entirely in simulation, which allows for a much larger dataset. Compared to [7], which demonstrates how a joint representation can be used in robotic tasks, our work is a systematic exploration of architectures that fuse visual and tactile modalities.

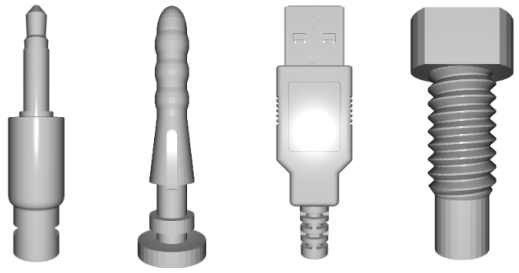

Figure 2. Four objects in dataset. The dataset consists of an AUX connector, pin, USB connector, and stud, all of similar size.

C. Methods

In our work, we train visual and tactile encoders for (1) object classification and (2) creating a joint latent space representation for downstream tasks. We implement both architectures with ResNets (ResNet-50) as well as Vision Transformers (vit-base-patch16-224) pretrained on ImageNet and compare performance.

Dataset. Our training dataset consists of 1,920 simulated depth images and simulated contact images (binary masks over the region of contact on the tactile sensor) for four different objects (Figure 2). The depth images are generated from a virtual depth camera that captures the object CAD model moving to discretized configurations on a grid with 5 degrees of rotational resolution, and 5mm of translational resolution. In order to obtain the tactile images, we threshold the signal of another virtual depth camera in the orientation of the gripper, as in [2]. The simulated depth images and contact images are aligned such that the center pixel of the depth image corresponds to the center of the gripper in a corresponding grasp (Figure 3).

To generate the heldout validation dataset, we apply a random pertubation (within ± 5 degrees and ± 5mm) to a subset (640) of the configurations in the training set, then render the vision and tactile images in the perturbed configuration. This method of generating the validation set ensures that the validation set is within the training distribution, but cannot have been observed during training.

We implemented several data augmentations that we selectively use when training different networks. Since we are training in simulation, we have implemented transforms relevant for sim2real transfer (added noise, random center crop, and occlusions for the visual image and slight rotations and morphological mask transformations for the tactile image). Variants of our networks use a subset of these transforms, as detailed below.

Classification network. For classification, we fine-tune

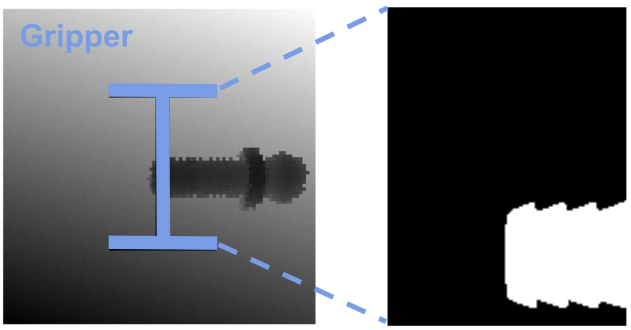

Figure 3. Simulated visuotactile data that we will use for training our models. The tactile data (right) consists of a white contact mask over the region of contact on the tactile sensor when the gripper grasps the object in the depth image (left). The blue "I" shape over the object in the depth image is a graphical representation of the gripper orientation with respect to the object.

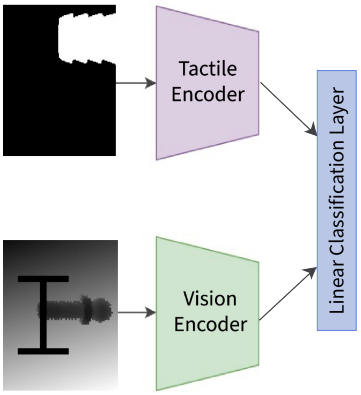

Figure 4. Visuotactile classification network architecture. Encoders are implemented in both ResNet and Vision Transformer versions.

tactile and vision encoders with an added linear classification layer. (Figure 4). We concatenate the outputs of the two encoders before passing it through the linear layer. We train with the categorical cross entropy loss, AdamW optimizer, batch size 64, and an exponential learning rate scheduler initialized at 1×10^{-4} for the ResNet encoders, and 1×10^{-5} for the ViT encoders. We train for 50 epochs, and evaluate the epoch with the lowest training loss against our validation set. Since the validation set contains purely simulated data, we omit data augmentations relevant to sim2real transfer when training the classification network.

We evaluate classification accuracy for three ablations of the task: (1) visuotactile classification, (2) vision-only classification, and (3) tactile-only classification. For the vision-only and tactile-only classification tasks, we pass only the outputs from the vision and tactile encoders, respectively, to the linear classification layer. We hypothesize that visuotactile classification will outperform tactile-only classification, but will provide only marginal benefit over vision-only classification in the absence of occlusion. This is because classification is more dependent on global object features, which vision readily provides. Additionally, since tactile information is especially useful in the case of occlusions, we train and compare a vision and visuotactile network on a dataset with the random occlusions included during data augmentation (Figure 5).

Figure 5. Data augmentations for introducing visual occlusions.

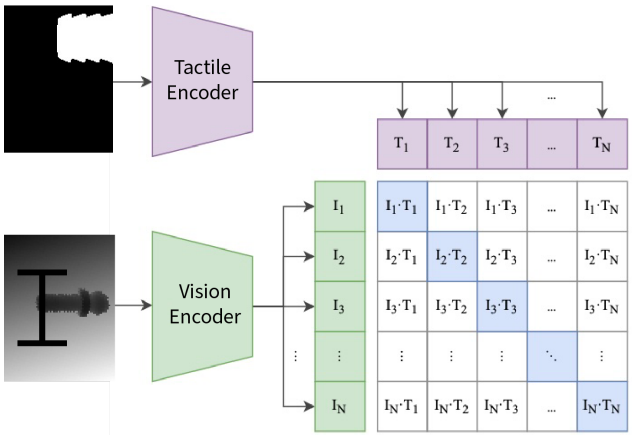

Figure 6. Visuotactile joint representation network architecture. Encoders are implemented in both ResNet and Vision Transformer versions.

Joint representation network. We train the learned joint representation using an InfoNCE contrastive loss in a self-supervised manner, since we have the ability to simulate corresponding pairs of tactile and overhead depth images (Figure 6). We use the same optimizer, learning rate scheduler, and hyperparameters as in the classification network (see above).

We compare the joint representations obtained from no data augmentation (same as classification network) and sim2real data augmentation (random center crop for the visual image and slight rotations and morphological mask transformations for the tactile image). We hypothesize that more data augmentation will lead to increased clustering of the representations by object class. We evaluate this hypothesis qualitatively by examining t_SNE plots of the resulting representations.

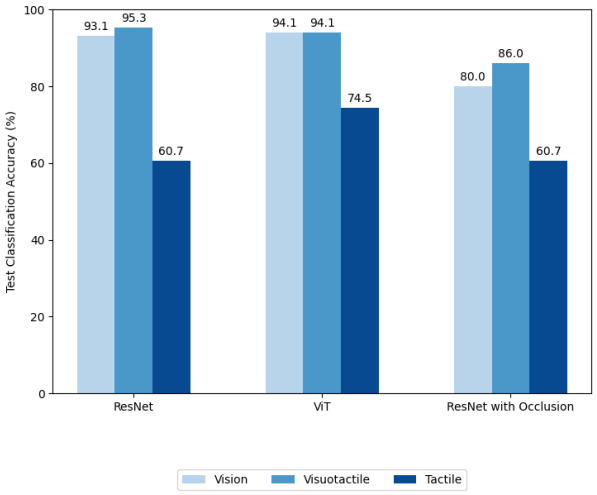

Figure 7. Test classification accuracies for vision-only, tactile-only, and visuotactile classification with both ResNet and Vision Transformer encoders. We also compare to a ResNet network trained with visual occlusions on a new test dataset with visual occlusions.

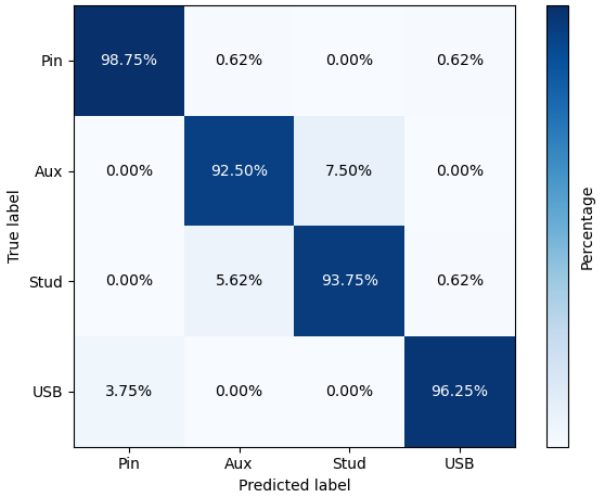

Figure 8. Confusion matrix for visuotactile object classification with ResNet Encoders.

D. Results and Discussion

We experiment with network architecture for two separate tasks, object classification and joint representation learning. For each task, we compare encoder architecture with ResNet and Vision Transformers and use different combinations of data augmentation to simulate different problems.

Classification network. We ablate our visuotactile classification networks (with both ResNet and Vision Transformer encoders) with vision-only and tactile-only classification (Figure 7). We found that the tactile-only classification performed significantly worse than the vision-only and visuotactile classification networks which had comparable performance. For the single modality classification networks, the Vision Transformer-based networks slightly outperform the ResNet-based networks. Both visuotactile networks had comparable performance, but the visuotactile network based on ResNet encoders had the best performance in our study with a test classification accuracy of 95.3%. AUX and stud were the two classes that created the most confusion because they are similar in both visual and tactile signals (Figure 8).

Tactile-only classification performing worse than the other modalities is expected because the local information this modality provides is less useful for object classification. Therefore, visuotactile does not significantly outperform just vision. ResNet-50 has over 23 million trainable parameters while ViT-Base has over 86 million parameters. Since both networks were pre-trained on the same ImageNet dataset, and transformers have a larger receptive field than convolutional neural networks as well as the self-attention mechanism, we hypothesized that Vision Transformers would outperform ResNets. However, this specific classification problem might not be complex enough to see a significant difference between the two architectures. Since tactile classification is a harder problem, we do see a more significant improvement in performance with Vision Transformers.

Joint representation network. For the representation learning task, we see that the learned representations for both encoders show some clustering based on the object class (Figure 9). The Vision Transformer joint representations appear to be more clustered by class. These were trained in a self-supervised manner, independent of class, so we see clustering based on local tactile features in the ResNet model like larger corner features and small periodic features (Figure 10). Convolutional networks are good at picking up local patterns while Vision Transformers with their larger receptive fields have more capacity for sorting these objects by class without explicit labels.

The Vision Transformer embeddings also appear more uniformly distributed on the unit sphere. The ResNet embeddings are of dimension 1K while the Vision Transformer embeddings are of dimension 151K. The added representational power of the Vision Transformer may allow for this more even distribution.

We do see more clustering in the representations learned with data augmentation, but do not see clear clustering based on the object class. Instead, visuotactile signals from multiple objects are collapsed to the same point in the representation space.

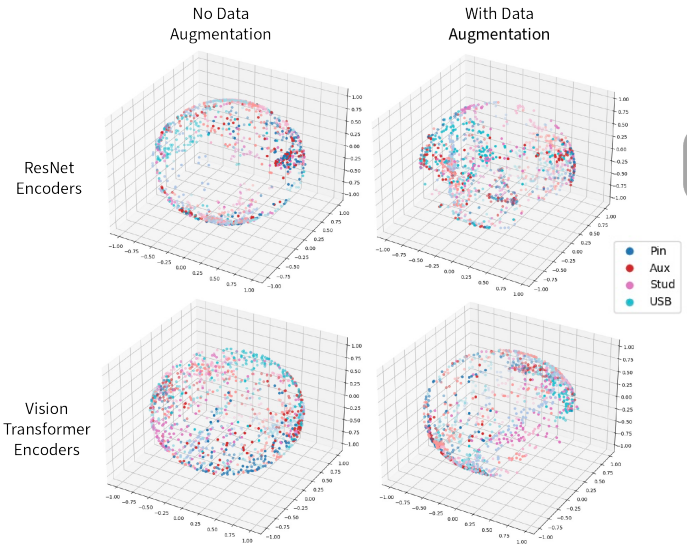

Figure 9. T-SNE plots of visuotactile joint representations. The colors correspond to the object type as labeled. The darker shades are for the tactile embeddings while lighter shades are the visual embeddings. The top row uses ResNet embeddings while the bottom row uses Vision Transformer embeddings. The left column uses no data augmentation (like our classification networks) while the right column has data augmentation.

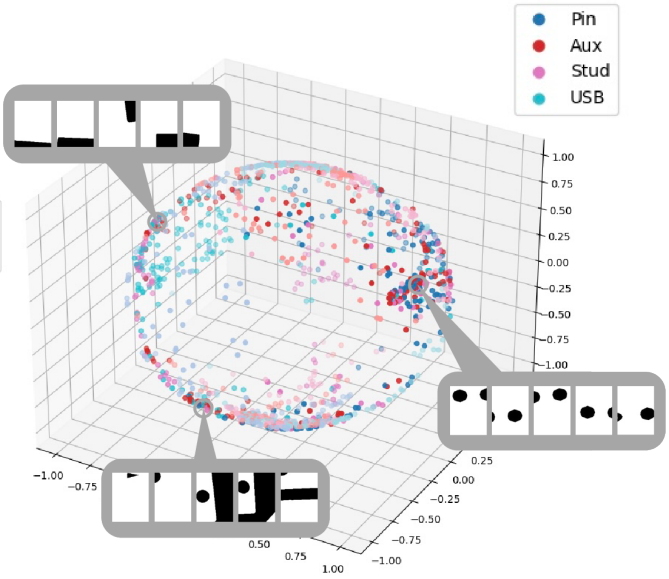

Figure 10. T-SNE plot of visuotactile joint representation for ResNet encoder and non-augmented images with tactile images sampled from major clusters. The colors correspond to the object type as labeled. The darker shades are for the tactile embeddings while lighter shades are the visual embeddings. The cluster on the top left shows large corner features, likely from the USB image. The cluster on the top right shows small periodic contacts. The sampled images are likely from the pin object. The bottom cluster seems to have less obvious patterns from our sample.

E. Conclusion

Visual and tactile modalities are complementary for several tasks in robotic manipulation and this work fuses these modalities for object classification and joint representation learning. We did not find significant differences in performance between ResNet and Vision Transformer encoders for object classification, except for the more difficult problem of tactile-only object classification. The vision and visuotactile networks performed similarly for the standard dataset, but once we introduced occlusions, the visuotactile classifier outperformed the vision-only classifier. Our highest performing network, visuotactile classification with ResNet, achieved a test classification accuracy of 95.3 %.

For the joint representation learning task, we found that the network based on Vision transformers was able to more evenly distribute the embeddings on the unit sphere because of the more expressive power of the network. Furthermore, the ResNet-based network seemed to cluster embeddings based on local features while the Vision Transformer-based network was able to cluster more based on class even without access to class labels. This result is possible because of the larger receptive field of Transformer networks compared to convolutional networks, thus the Transformer can connect features globally more effectively than convolutional networks.

Future Work One limitation of our existing approach to representation learning is that the constrastive loss only considers the relationship between samples in a given batch. This can lead the loss function to jump around, and lead to instability during training. In the future, we would like to implement contrastive learning as a dynamic dictionary with a queue and a moving averaged encoder, as in [5], to improve performance.

We attempted adding cross-modal attention between the encoders and final linear layer for our best classification network architecture, however ran into compute limitations when trying to implement this. Comparing the performance gains of self attention with cross-modal attention is interesting especially for our multi-modal classification task.

We would also like to see how the Sim2Real data augmentations we have implemented affect classification performance on the simulated test set as well as a test set generated on the real robot.

Eventually, we would like to implement our networks for real robot tasks. For example, the classification task can be used to identify objects in clutter if we add more objects in frame to the visual dataset. The joint representation is a lower-dimensional representation that simplifies policy learning for tasks like cable insertion, in-hand manipulation, or tool use.

F. Individual Contributions

Both group members worked together for assembling and debugging most components of this project. Specifically, Antonia contributed to dataset generation, the Transformers architecture, training at scale, overall code development, and visualization of results. Neha developed the ResNet and overall classification and contrastive loss network architectures, parameter tuning for training, and data augmentation.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. *CoRR*, abs/2104.11178, 2021. 1

[2] Maria Bauza, Antonia Bronars, and Alberto Rodriguez. Tac2pose: Tactile object pose estimation from the first touch, 2022. 2

[3] Yizhou Chen, Andrea Sipos, Mark Van der Merwe, and Nima Fazeli. Visuo-tactile transformers for manipulation, 2022. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 1, 2

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019. 5

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[7] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Learning self-supervised representations from vision and touch for active sliding perception of deformable surfaces. *arXiv preprint arXiv:2209.13042*, 2022. 2

[8] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. 2

[9] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. *CoRR*, abs/1810.10191, 2018. 1, 2

[10] Nathan F Lepora. Soft biomimetic optical tactile sensing with the tactip: A review. *IEEE Sensors Journal*, 21(19):21131–21143, 2021. 2

[11] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019. 2

[12] Shan Luo, Wenzhen Yuan, Edward Adelson, Anthony G Cohn, and Raul Fuentes. Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2722–2727. IEEE, 2018. 2

[13] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 1

[14] Youngsun Wi, Pete Florence, Andy Zeng, and Nima Fazeli. VIRDO: visio-tactile implicit representations of deformable objects. *CoRR*, abs/2202.00868, 2022. 2

[15] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 1, 2