High-Resolution Pixelwise Contact Area and Normal Force Estimation for the GelSight Mini Visuotactile Sensor Using Neural Networks

Niklas Funk*[1], Paul-Otto Müller*[1], Boris Belousov[2], Anton Savchenko[3], Rolf Findeisen[3], Jan Peters[1,2,4,5]

***Abstract*— Visuotactile sensors are gaining momentum in robotics because they provide high-resolution contact measurements at a fraction of the price of conventional force/torque sensors. It is, however, not straightforward to extract useful signals from their raw camera stream, which captures the deformation of an elastic surface upon contact. To utilize visuotactile sensors more effectively, powerful approaches are required, capable of extracting meaningful contact-related representations. This paper proposes a neural network architecture called CANFnet that provides a high-resolution pixelwise estimation of the contact area and normal force given the raw sensor images. The CANFnet is trained on a labeled experimental dataset collected using a conventional force/torque sensor, thereby circumventing material identification and complex modeling for label generation. We test CANFnet using GelSight Mini sensors and showcase its performance on real-time force control and marble rolling tasks. We are also able to report generalization of the CANFnets across different sensors of the same type. Thus, the trained CANFnet provides a plug-and-play solution for pixelwise contact area and normal force estimation for visuotactile sensors. The models, dataset, and additional information are open-source at `https://sites.google.com/view/canfnet`.**

I. INTRODUCTION & RELATED WORKS

One of the biggest challenges in robotics is deploying autonomous systems in the real-world [1]. Particularly contact-rich tasks such as dexterous manipulation, reliable grasping, and precise assembly are still open research problems [2], [3], [4], [5]. Sensing object pose and other properties such as geometry, mass, etc., is especially difficult using external sensing only, e.g., due to gripper-object occlusions [6]. It is thus important to equip robots with sensors that can perceive contact properties directly. Tactile feedback has the potential to enhance robustness, precision, and reliability in tasks such as grasping, autonomous assembly, and stable object placing in complex and unstructured environments [7], [8].

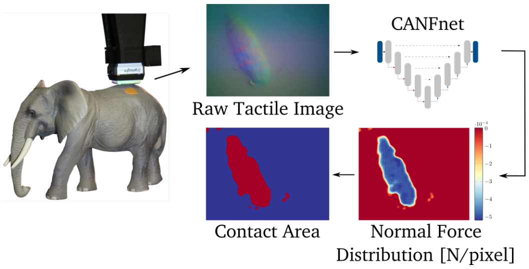


Fig. 1. Overview of the proposed method. A raw tactile image is passed through the CANFnet, which outputs a normal force distribution. Subsequently, the contact area is estimated via straightforward thresholding.

Visuotactile sensors measure contact properties through a camera capturing the deformation of an elastomer that interacts with the environment (cf. Fig. 1). They are promising due to potentially high spatial resolution, low cost, and ease of manufacturing. While many visuotactile sensors have been proposed already [9], [10], they are not readily available. Therefore, herein, we focus on the commercially available GelSight Mini [11], [12], thereby removing the most significant entry barrier into the field, i.e., manufacturing knowledge and facilities, and promoting reusability and reproducibility of the developed methods.

One of the significant challenges in visuotactile sensing is how to infer the contact information from the raw sensor images and which representation to choose. One possibility is to reconstruct the sensor's depthmap using photometric stereo [13], [14]. Other approaches use learning to directly regress from the image [14], [15], [16] or lower-dimensional representations, such as tracking markers inside the gel [17], [12], to force estimates. Alternatively, the images can be used directly in an end-to-end fashion to solve downstream tasks [18], [19], [20], [21]. However, this comes at the cost of losing interpretability and transferability to different tasks. Instead, learning policies or designing controllers on top of physically meaningful intermediate representations can effectively speed up the transfer to new tasks, mitigate overfitting and enhance generalization capabilities [19]. While there are various possibilities for an intermediate visuotactile representation, in this work, we learn a neural network that directly maps from input image to the normal force distribution that caused the deformation of the gel. This choice is because the force and force distribution directly relate to the statics and dynamics of a contact configuration [22]. Furthermore, force provides a natural interface in the context of control tasks.

Our proposed approach (cf. Fig. 1) focuses on maintaining the high spatial resolution benefit of visuotactile sensors. We train the network using solely experimental data collected using a conventional force/torque (F/T) sensor. We, therefore, neither require mathematical models and simulations nor other intermediate representations such as marker movements or depthmap estimation. Using real-world experimental data by nature amortizes potential inaccuracies and modeling errors. In contrast to other works, we estimate the normal force distribution pixelwise, i.e., at the same resolution as the

∗ Authors contributed equally.

This work is supported by the AICO grant by the Nexplore/Hochtief Collaboration with TU Darmstadt. Calculations for this research were conducted on the Lichtenberg high performance computer of the TU Darmstadt.

[1] Department of Computer Science, Technical University of Darmstadt, [2] German Research Center for AI (DFKI), Research Department: Systems AI for Robot Learning, [3] Department of Electrical Engineering and Information Technology, Technical University of Darmstadt, [4] Hessian.AI, [5] Centre for Cognitive Science. `name.surname@tu-darmstadt.de`

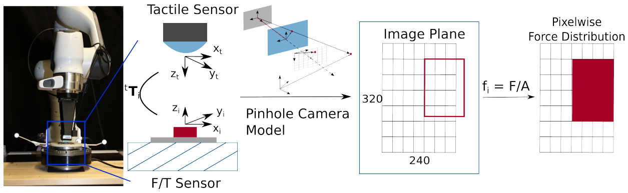


Fig. 2. Main steps for collecting the labeled dataset for training our proposed CANFnet. The left image depicts the real experimental setup in the lab. Next, we show a schematic close-up view of the most important components, i.e., visuotactile sensor, F/T, and the indenter (in red), mounted on top. By exploiting knowledge about the relative transformations between indenter and tactile sensor, we can project the indenter's contact area into the image of the visuotactile sensor using the Pinhole camera model. Finally, we obtain a labeled pixelwise force distribution by dividing the ground truth force measurement by the contact area and assigning this value to all pixels within the contact area.

original image. Furthermore, our model outputs a pixelwise estimate of the contact area between the external object and the sensor. Both normal force and contact area are essential for tasks that involve contact-rich object manipulation. To prevent loss of spatial information, we consider gels that do not have any impainted markers or dots.

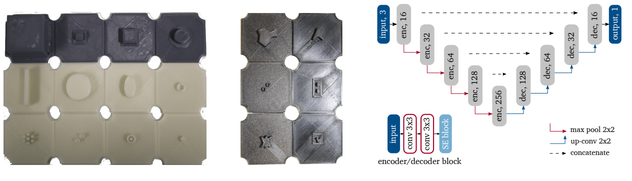
In summary, we make the following contributions. We propose CANFnet, an architecture and training procedure that provides a plug-and-play module for estimating normal force and contact area at a pixel level. We test CANFnet on GelSight Mini, and demonstrate that the trained network can generalize over different sensors, gels, and test objects. We showcase CANFnet's real-time capability by applying it to a dynamic marble rolling and force tracking task. We open-source our model and the training data to facilitate reproducibility and further advancements in this area.

II. METHOD

We next describe the experimental setup for collecting the labeled dataset containing pixelwise normal force values. Subsequently, we introduce a neural network for mapping from the raw sensor images to the high-resolution labels for **C**ontact **A**rea and **N**ormal **F**orce across the gel called CANFnet.

A. Experimental Setup for Creating High-Resolution Labels

One of our goals is to avoid complex and computationally expensive mathematical models for training data generation. Therefore, our method relies on labeled experimental data. F/T sensors only provide one single measurement, which contrasts with our idea of pixelwise normal force reconstruction. We propose the following procedure to collect pixelwise labeled experimental data. As shown in Fig. 2, we mount an object with known geometry, that we denote as *indenter*, on top of a F/T sensor and attach the visuotactile sensor at the robot's end-effector. The idea is to bring the sensor and indenter into a contact configuration where the normal force acts uniformly upon the gel. Exploiting a well-calibrated setup in which we know all the relevant transformations, we can project the indenter's geometry into the images of the tactile sensor. Subsequently, we obtain pixelwise labels by dividing the total measured normal force from the F/T sensor by the contact area. Note that we do not aim to reconstruct the normal forces acting inside the gel; instead, our labels relate to the external normal force exerted onto the gel. Our procedure is based on two main assumptions. First, the indenters, i.e., the objects pressing onto the visuotactile sensors, must have flat surfaces and known dimensions. Second, during contact, the indenter and visuotactile sensor surfaces are parallel. For network generalization, we collect data using indenters varying in shape and size (cf. Fig. 3(a)). We also press against the indenters using different positions and orientations. In total, we collected 277325 training samples with forces ranging from $0\,\mathrm{N}$ to $20\,\mathrm{N}$.



(a) Train Indenters. (b) Test Indenters. (c) U-Net architecture.

Fig. 3. These 3D-printed indenters have been used for training data collection (left) and testing (middle). The architecture of the proposed CANFnet, i.e., a U-Net architecture mapping from raw visuotactile images to a pixelwise normal force distribution (right).

B. CANFnet

We use a U-Net [23] to estimate the force distribution and contact area. The U-Net is a powerful convolutional neural network (CNN) architecture originally developed for segmentation in the biomedical field.

Architecture. Fig. 3(c) shows the proposed U-Net architecture, containing a contracting path (encoder) together with an expanding path (decoder), retrieving spatial information from the encoder's latent space. The U-Net thus has a 3-channel $320{\times}240$ input, and outputs the pixelwise normal force estimate $f_{x,y}^{\mathrm{UNET}}$ for each pixel location x, y. The contact area is obtained by thresholding the force estimates. For the details, we refer to our code which is publicly available.

Training. The U-Net has to resolve a pixelwise regression problem. The force distribution labels $f_{x,y}^{FT}$ have a unit of $\mathrm{N/pixel}$. A potential loss function for training the network's pixelwise force prediction $f_{x,y}^{\mathrm{UNET}}$ could thus be the mean squared error (MSE) per pixel, $\mathrm{MSE}(f_{x,y}^{\mathrm{UNET}}, f_{x,y}^{\mathrm{F/T}}) = 1/(\mathrm{WH}) \sum_{x=0}^{\mathrm{W}} \sum_{y=0}^{\mathrm{H}} (f_{x,y}^{\mathrm{UNET}} - f_{x,y}^{\mathrm{F/T}})^2$ with image width and height W, H, respectively. However, initial experiments revealed that this resulted in relatively big errors when comparing the integrated force values with the measurements from the F/T sensor $F_{F/T}$. Consequently, we added an integrated force loss (IFL) term comparing the integrated normal force prediction with the F/T measurement

$$\mathrm{IFL}(f_{x,y}^{\mathrm{UNET}}, F_{\mathrm{F/T}}) := \left(\sum_{x=0}^{\mathrm{W}} \sum_{y=0}^{\mathrm{H}} f_{x,y}^{\mathrm{UNET}} - F_{\mathrm{F/T}} \right)^2 . \quad (1)$$

The overall loss function is thus a weighted sum of IFL and MSE with weighting factor w_{ifl} controlling whether more focus should be put on local or global force reconstruction,

$$\mathcal{L} = \mathrm{MSE}(f_{x,y}^{\mathrm{UNET}}, f_{x,y}^{\mathrm{F/T}}) + w_{\mathrm{ifl}}\, \mathrm{IFL}(f_{x,y}^{\mathrm{UNET}}, F_{\mathrm{F/T}}). \quad (2)$$

TABLE I

EVALUATION OF CANFNET VS BASELINES FOR NORMAL FORCE AND CONTACT AREA ESTIMATION

Method	Sensor	IoU	MAE [F]
Image Diff	Train	0.301 ± 0.001	NA
(Pytouch [24])	Test	0.294 ± 0.001	NA
3D-Recon	Train	0.463 ± 0.002	1.472 ± 0.012
(GelSight [25])	Test	0.451 ± 0.002	1.805 ± 0.016
CANFnet (ours)	Train	$\mathbf{0.73 \pm 0.002}$	$\mathbf{0.82 \pm 0.009}$
(Neural Network)	Test	$\mathbf{0.718 \pm 0.002}$	$\mathbf{0.767 \pm 0.007}$

We reason that the additional global regularization term is needed since there might be small mismatches in the image data and the force labels, which might come from inaccuracies in the calibration. Small errors might accumulate to a large error in the overall force prediction due to the high pixel count. $w_{\text{ifl}}=0.05$ was found to be a good compromise.

To obtain the pixelwise contact area estimation, we use the normal force prediction. There is contact between the sensor and object $\text{ic}_{x,y}=1$ at pixel x, y, whenever the normal force exceeds the empirical threshold of $w_{\text{ic}}=10^{-4}$.

We also use data augmentation to reduce overfitting and enhance generalization abilities of the networks as well as their robustness against object and sensor variations.

III. EXPERIMENTAL RESULTS

After generating training data according to subsection II-A and training CANFnet, we validate the results. We want to point out that we did not only evaluate our models on the same sensor with which the training data was collected ('train sensor') but also on a different sensor ('test sensor') which may differ in lighting, gel properties, camera calibration, etc.

CANFnet Inference Speed. We test our network on a *NVIDIA GeForce RTX 3090* GPU and an *AMD Ryzen 9 5950X 16-Core* CPU. The mean inference time is 3.498 ms over 300 runs, which is sufficient for real-time operation, as the GelSight operates at 25 Hz.

Contact Area & Normal Force. We evaluate CANFnet and two baselines on contact area detection and static force reconstruction. We collected 10000 images for the train and test sensor, solely using new indenters ('unseen') that have not been used during training (cf. Fig 3(b)). We report the mean absolute error (MAE) for normal force estimation, and the intersection over union for contact area estimation. For the IoU, higher is better; for the force error, lower is better.

The baselines are: 1) Image Diff, which is based on PyTouch [24] and estimates the contact area using the differences between the current image and a reference image prior to contact, solely using classical image processing. 2) 3D-Recon reconstructs the depth map of the GelSight sensor [25]. For estimating contact area, a pixel is considered in contact if the depth value exceeds a threshold. For the normal force, we fit a linear model mapping from total gel deformation to normal force.

As shown in Table I, Image Diff results in lowest IoU values, and is outperformed by 3D-Recon. However, the IoU estimation of the 3D-Recon baseline suffers especially at higher forces and smaller objects. In these scenarios, the gel deforms in a much wider area than just at the contact area, resulting in lower IoU values. Our proposed CANFnets perform best and clearly outperforms the baselines (improving by a factor of more than two compared to Image Diff, by around 58% compared to 3D-Recon). Considering normal force reconstruction, the proposed CANFnet also performs best. Lastly, we also computed the average pixel error between the center of mass of the groundtruth contact area and the CANFnet's predictions. The error is 8.3 pixels which roughly corresponds to 0.45 mm, thereby underlining the high spatial resolution. Moreover, we observe that the model transfers seamlessly to the test sensor, although it was only trained on data from the train sensor.

TABLE II

EVALUATION ON THE TASK OF TRACKING DESIRED FORCE PROFILES FOR DIFFERENT FORCE RANGES AND INDENTER SIZES

Sensor	Indenter	Relative Size	Force Range [N]	MAE [N]
GelSight (Test)	seen	smaller	$0-10$	$\mathbf{0.491 \pm 0.015}$
		larger	$0-10$	0.863 ± 0.022
			$0-18$	2.073 ± 0.063
	unseen	smaller	$0-10$	0.919 ± 0.027
		larger	$0-10$	$\mathbf{0.602 \pm 0.02}$
			$0-18$	$\mathbf{0.961 \pm 0.029}$

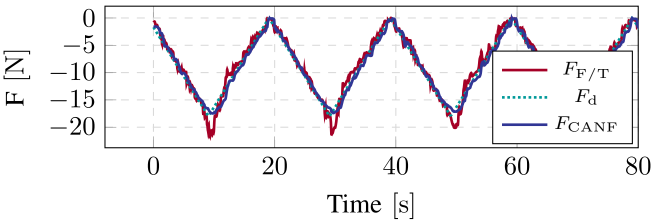


Fig. 4. Force tracking. The desired sawtooth profile F_d is tracked using our CANFnet F_{CANF}. The ground truth is given by the F/T sensor $F_{\text{F/T}}$.

Force Control. To get an impression of our model's capabilities on a more realistic task, i.e., force control, we mount the visuotactile sensor at the end-effector of a Franka Panda robot and track a sawtooth force profile.Thus, the visuotactile sensor presses against an indenter mounted on top of a F/T sensor, same as during data collection.An example trajectory can be seen in Fig. 4. The results in Table II are similar to the previous static experiments, demonstrating that CANFnet can also deal with more dynamic scenarios, i.e., changes in input images. The MAE is generally better at lower forces for smaller objects, covering only part of the sensing surface. While the force of a larger indenter is distributed over a bigger area and the gel deformation saturates at increasing pressure, the variation in the pixel values is only high at the edges of that object. Consequently, the network's potential to learn through image differences and the information contained in the images decreases.

Marble Roll. To showcase further advantages compared to a regular F/T sensor, we designed a marble roll task that requires joint force tracking and contact area estimation. A small marble is placed on a table and manipulated through the visuotactile sensor attached to a Franka Panda's tool flange (cf. Fig. 6(a)). The task is to move the marble to a desired position in the tactile image. The marble's centroid is estimated through CANFnet's contact area prediction. Fig. 5 shows the resulting trajectories. The marble is successfully moved to the desired set points. This illustrates that CANFnet can be used as a component in a larger task, jointly providing force and contact area information.

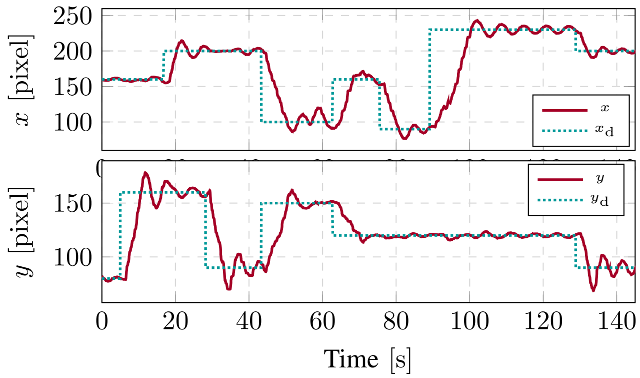
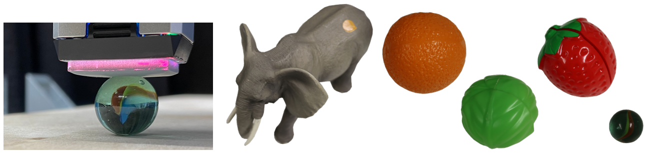


Fig. 5. Trajectories (x (top), y (bottom)) of the marble roll task. The desired position of the marble center (x_d, y_d, in pixel coordinates) is tracked by manipulating the marble through the visuotactile sensors mounted at the robot's end effector (cf. Fig. 6(a)).



(a) GelSight & marble. (b) Elephant, orange, salad, strawberry, marble.

Fig. 6. (a) An image of the marble roll task. (b) Five irregular toy objects used to evaluate the generalization of CANFnet to non-flat indenters.

Generalization to Irregular Objects. Given that CANFnet is trained only on flat indenters (cf. Fig. 3), how well can it generalize to irregularly shaped objects? We therefore repeated the force tracking task using five irregularily shaped objects (cf. Fig. 6(b)). We observer that the model also generalizes to the more irregular objects as the errors remain consistent with the previous experiments (see Table III).

IV. Conclusion

We presented CANFnet, a neural network for mapping from the raw images of visuotactile sensors to a pixelwise estimate of normal force and contact area, thereby not compromising any resolution. The method aims to maintain one of the main benefits of visuotactile sensors, their high spatial resolution, while at the same time counteracting one of their main drawbacks, namely, that the raw sensor signals are hard to interpret, making their integration cumbersome. To train CANFnet, we propose an effective experimental setup that allows for the creation of labeled high-resolution data using a standard F/T sensor and a robot manipulator. Our experimental evaluations underline the effectiveness of the trained representations in several experiments ranging from static force and contact area estimation to dynamic force control and a marble roll task. We open-source all models and data and plan to exploit CANFnet in more complicated robotic downstream tasks in the future.

TABLE III
Evaluation of CANFnet on Force Estimation using Irregular Toy Objects

Sensor	Object	Force Range [N]	MAE [N]
GelSight (Test)	elephant	0 − 10	**0.542 ± 0.008**
	marble	0 − 7	**0.936 ± 0.034**
	orange	0 − 9	**0.339 ± 0.006**
	salad	0 − 10	**1.138 ± 0.024**
	strawberry	0 − 8	**0.942 ± 0.023**

References

[1] F. Negrello, H. S. Stuart, and M. G. Catalano, "Hands in the real world," *Frontiers in Robotics and AI*, vol. 6, p. 147, 2020.

[2] N. Funk, C. Schaff, R. Madan, T. Yoneda, J. U. De Jesus, J. Watson, E. K. Gordon, F. Widmaier, S. Bauer, S. S. Srinivasa *et al.*, "Benchmarking structured policies and policy optimization for real-world dexterous object manipulation," *IEEE Robotics and Automation Letters*, 2021.

[3] N. Funk, G. Chalvatzaki, B. Belousov, and J. Peters, "Learn2assemble with structured representations and search for robotic architectural construction," in *Conference on Robot Learning*, 2022.

[4] N. Funk, S. Menzenbach, G. Chalvatzaki, and J. Peters, "Graph-based reinforcement learning meets mixed integer programs: An application to 3d robot assembly discovery," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[5] M. Schnaubelt, S. Kohlbrecher, and O. von Stryk, "Autonomous Assistance for Versatile Grasping with Rescue Robots," in *2019 IEEE International Symposium on Safety, Security, and Rescue Robotics*.

[6] B. Wen, C. Mitash, B. Ren, and K. E. Bekris, "se(3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains," in *IROS*. IEEE, 2020, pp. 10 367–10 373.

[7] C. Chi, X. Sun, N. Xue, T. Li, and C. Liu, "Recent Progress in Technologies for Tactile Sensors," *Sensors*, vol. 18, p. 948, 2018.

[8] L. Lach, N. Funk, R. Haschke, S. Lemaignan, H. J. Ritter, J. Peters, and G. Chalvatzaki, "Placing by touching: An empirical study on the importance of tactile sensing for precise object placing," *arXiv preprint*.

[9] N. F. Lepora, "Soft biomimetic optical tactile sensing with the tactip: A review," *IEEE Sensors Journal*, vol. 21, pp. 21 131–21 143, 2021.

[10] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using fingervision," in *Humanoids*. IEEE, 2017, pp. 241–248.

[11] "Gelsight — Products," https://www.gelsight.com/products/, accessed: 2023-02-28.

[12] W. Yuan, S. Dong, and E. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[13] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *CVPR*, 2009.

[14] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[15] C. Zhang, S. Cui, Y. Cai, J. Hu, R. Wang, and S. Wang, "Learning-based six-axis force/torque estimation using gelstereo fingertip visuotactile sensing," in *IROS*. IEEE, 2022, pp. 3651–3658.

[16] C. Sferrazza, A. Wahlsten, C. Trueeb, and R. D'Andrea, "Ground truth force distribution for learning-based tactile sensing: A finite element approach," *IEEE Access*, vol. 7, pp. 173 438–173 449, 2019.

[17] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *Humanoids*, 2016, pp. 1045–1051.

[18] B. Belousov, A. Sadybakasov, B. Wibranek, F. Veiga, O. Tessmann, and J. Peters, "Building a library of tactile skills based on fingervision," in *Humanoids*. IEEE, 2019, pp. 717–722.

[19] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-RL for Insertion: Generalization to Objects of Unknown Geometry," Apr. 2021.

[20] M. Lambeta, P.-W. Chou, S. Tian, B. Yang *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*, vol. 5, no. 3, pp. 3838–3845, 2020.

[21] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-RL: Learning Multimodal Manipulation Policies with Deep Reinforcement Learning," in *ICRA*, 2022, pp. 8298–8304.

[22] Y. Zhang, Z. Kan, Y. Yang, A. Y. Tse, and M. Y. Wang, "Effective Estimation of Contact Force and Torque for Vision-based Tactile Sensor with Helmholtz-Hodge Decomposition," Jun. 2019.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation."

[24] M. Lambeta, H. Xu, J. Xu, P.-W. Chou *et al.*, "PyTouch: A machine learning library for touch processing," *ICRA*, 2021.

[25] "Gelsightinc/gsrobotics: GelSight SDK for Robotic Sensors," https://github.com/gelsightinc/gsrobotics, accessed: 2023-02-28.