

Forces at your fingertips: Estimating Hand-Object Contact Forces using Newtonian Physics

Piyush Bagad

VGG, University of Oxford
piyush@robots.ox.ac.uk

Taein Kwon

Meta Reality Labs, Zurich

Abstract: In this work, our objective is to estimate the distribution of contact forces on the hand when holding an object. This is a fundamental problem at the core of multiple downstream tasks such as physically-plausible object manipulation, affordance, and precise hand-object reconstruction. Using simple Newtonian physics, we formulate this as an optimization recipe to estimate forces on hand vertices given the object is static (*i.e.*, zero net force and torque). The physical properties (*e.g.*, mass and friction coefficient) are estimated by leaning on the world knowledge of VLMs. The method is also extended to video data where we jointly optimize a temporal sequence of forces with smoothness regularizers. Since there is no single dataset of hand-held objects with both 3D annotations and force heatmaps, we build a robust pipeline for 3D hand and object reconstruction on EgoTactile which has force ground truth. Forces simulated by our method outperform existing baselines on EgoTactile. Furthermore, our method generalizes to in-the-wild data, *e.g.*, on samples from EPIC-Contact.

Keywords: Contact forces, Newton’s laws, 3D hand-object reconstruction

1 Introduction

The year is 1666. Newton is sitting under *that* apple tree. While pondering about why the apple fell straight into the ground, he picks it up, holds it steady and observes it carefully. While inspecting it, he loses his grip and the apple falls. So, he murmurs to himself, why does the apple remain static when held in a particular way and falls off as the grip loses? Thanks to Newton himself, today we can use high-school physics to reason about this problem. As shown in Fig. 1, as the apple is static, the net external forces on the apple are zero, *i.e.*, the gravitational pull is balanced by the contact-normal forces and the frictional forces. Likewise, since there is no rotation, the net torque is also zero. Together, they enable us to recover a plausible force profile. This is the crux of our paper.

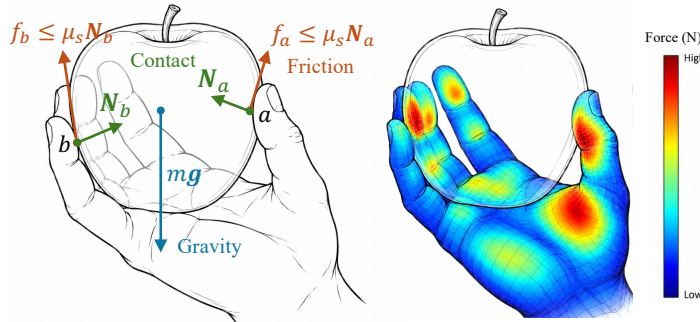


Figure 1: **Free-body diagram based on Newton’s laws.** On the left, an apple held by a hand is static: net external forces and torques add up to zero, naturally giving us an optimization objective to estimate the force profile illustrated on the right.

Given the image of hand-held object and the 3D mesh representation of the posed hand-object system, our goal is to recover the force profile as illustrated with a toy example in Fig. 1. There are several challenges to implement this idea based on Newton’s laws in practice. (i) *Tricky optimization*: this is an under-specified problem since there can be multiple ways to hold an object static. Even for a given pose, there can be multiple solutions to the optimization problem. (ii) *Physical*

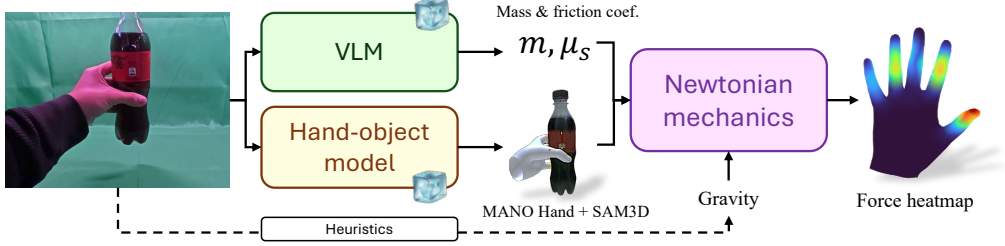


Figure 2: **Newtonian force estimation.** Given an image of hand-held object, we leverage a powerful VLM to estimate its mass m and (static) friction coefficient μ_s . Optionally, when required, gravity is estimated based on geometric heuristics. Then, m, μ_s , 3D reconstruction of hand-object system are fed to our optimization module which uses Newton’s laws to estimate the force field on the hand.

properties: We need to estimate physical properties (object mass m and friction coefficient μ) and direction of gravity in a robust manner from a single image. (iii) *Generalizability:* While this idea has been explored in prior work, it is usually evaluated in a very limited setting [1, 2, 3] or with specialized hardware [4]. (iv) *Evaluation data:* Finally, there is no single dataset with 3D hand and object as well as force annotations. Datasets with 3D hand-object poses do not have force measurements [5, 6, 7] and vice-versa [8, 9]. There are some datasets with ground truth forces but are collected with gloved hands which makes it difficult for 3D reconstruction [10, 11].

(i) We solve the optimization as a second-order cone program (SOCP) inspired by Pham et al. [1]. While the force/torque equations are linear in force F_x at point x , the friction equations are second-order cone constraints [12]. To restrict the solution space, we only consider $K_a=32$ anchor points (Yang et al. [13]) given the fixed topology of the MANO hand model [14]. Furthermore, we pick the minimum Euclidean norm solution intuitively picking a force profile that picks the object with the least contact effort. (ii) To obtain physical properties m, μ_s , we leverage the world knowledge of frontier VLMs [15, 16]. In cases where the camera axis is not perpendicular to gravity, we use simple geometric heuristics. (iii) The combination of these ingredients make our method generalizable: it works on in-the-wild samples from EPIC-Contact by Bansal et al. [17]. A brief overview of the complete pipeline is shown in Fig. 2. (iv) To address the data problem, we take an existing dataset, EgoTactile (Zeng et al. [8]), which has pressure grid annotations. We transfer the grid values to the MANO hand mesh giving us approximate force measurements. We build a semi-automatic pipeline to obtain accurate hand-object reconstructions detailed later [18].

We compare our method with a strong baseline: the SOFA physics simulator from EgoPHI [2] on the same 3D hand-object meshes for EgoTactile [8]. Our method outperforms SOFA across all the metrics, notably by +0.12 points on Pearson’s correlation. We also demonstrate generalization to other datasets like DexYCB [5] and in-the-wild samples from EPIC-Contact [17]. We also demonstrate an extension of the recipe to video data with joint optimization over force profiles over time.

2 Method

Physical properties. We lean on the world knowledge of a Multimodal LLM (VLM) to estimate $\phi = \{m, \mu_s\}$, m is mass, μ_s friction coefficient, from a single image [16]. The VLM is prompted to estimate the object’s mass in kilograms and its Coulomb friction coefficient against dry human skin.

Gravity estimation. In some datasets, the camera is parallel to the ground plane [8] which makes $\vec{g}$ trivial. For in-the-wild samples [17], we need to estimate gravity explicitly. Let’s assume known camera intrinsics K . Suppose we have a line $\vec{l}$ in the 2D image plane, then, it back-projects to a plane in 3D with normal $\vec{n} = K^T \vec{l}$. If this line was parallel to gravity $\vec{g}$, then $\vec{g}$ must lie in this plane, i.e., $\vec{n}^T \cdot \vec{g} = 0$. Thus, each line gives us one linear constraint on $\vec{g}$. If we have many (≥ 3) such lines $\{\vec{n}_j\}$ we can solve for $\vec{g}$ using robust least squares. Since we only qualitatively evaluate on [5, 17], we manually annotate such 2D lines (e.g., vertical edges of a box) but this can be done automatically at a larger scale by first detecting lines [19]. Then, either (i) using RANSAC to filter the set of vertical lines, (ii) using semantic labels (e.g., table edges). We leave this to future work.

Force estimation. Given a posed MANO hand mesh $\mathcal{H}$, a posed object mesh $\mathcal{O}$ (both in metres in a shared camera frame), object mass m , friction coef. μ_s , and a unit gravity direction $\hat{g}$, we estimate a per-vertex force magnitude field on the hand [1]. We pick a subset of K_a anchors within a distance threshold of $\mathcal{O}$. For each active contact k , let p_k be the closest point on the object, n_k the inward object normal at p_k , and $(t_{x,k}, t_{y,k})$ an orthonormal tangent basis. If fewer than three anchors are in range, we relax the threshold; if none remain, we return zero force. Let $F_k \in \mathbb{R}^3$ be the force applied *on the object* at p_k , written in the local contact frame as

$$F_k = f_{n,k} n_k + f_{t_x,k} t_{x,k} + f_{t_y,k} t_{y,k}, \quad f_{n,k} \geq 0, \quad \|(f_{t_x,k}, f_{t_y,k})\|_2 \leq \mu f_{n,k}. \quad (1)$$

Static equilibrium of the free body $\mathcal{O}$ requires

$$\sum_k F_k + mg\hat{g} = 0, \quad \sum_k (p_k - c_{\text{com}}) \times F_k = 0, \quad (2)$$

where c_{com} is the object centre of mass and $g=9.81 \text{ m s}^{-2}$. Among solutions of this underdetermined system we take the minimum-norm force field

$$\min_{\{F_k\}} \sum_k \|F_k\|_2^2 \quad \text{s.t. equilibrium and Coulomb cones}, \quad (3)$$

which is a second-order cone program (SOCP). Force/torque balance is linear in F_k ; the cones make the program an SOCP rather than an linear programming. If the hard program is infeasible (we cannot find a solution that satisfies all criteria), then, we add a small slack to force-torque equations and minimize the slack. The SOCP yields forces only at the K active anchors. We Gaussian-splat $|F_k|$ onto the 778 MANO vertices to obtain a dense magnitude map used for evaluation.

3 Experiments

3.1 Quantitative evaluation on EgoTactile

EgoTactile [8] provides grid-based pressure annotations which we normalize to $[0, 1]$ since we cannot compare with raw sensor units [11]. Zeng et al. [8]’s clever trick is to make the participant pick up the same object with both hands with tactile sensor gloves only on one hand while the video is recorded with the other bare hand. We use the bare-hand videos for 3D reconstruction of hand and object as described below [18].

1. **Initial hand and object mesh:** We segment the hand and object using SAM3 [20]. In each video, we pick a frame where the hand is not yet in touch with the object. We feed this frame to SAM-3D [21] to construct a high-fidelity mesh of the object. We also feed this frame to a frontier VLM (e.g., Claude Sonnet 5.5) and prompt it to estimate m, μ_s . Then, we pick a key frame where the object rests static in the hand. In that frame, we use WiLoR [22] to initialize the MANO hand parameters.
2. **Scale estimation:** We use MoGe-V2 [23] to get the depth map. Then, for both hand/object, we lift the 2D mask to 3D with the depth map and separately tune s_o, s_h and translation by fitting the hand/object mesh to its point cloud.
3. **Joint optimization:** Then, for the key frame, we jointly optimize the hand and object parameters in a curriculum fashion starting with refining scale, then translation, and then pose. We use standard mask re-projection loss and a penalty on intersection.
4. **Manual refinement:** We observe that Steps 1-3 do not yield clean, accurate contact points between hand and object. Since force estimation depends critically on reasonable contact points, we manually adjust the scale, translation and global rotation of the object on an interactive 3D platform. More details in Supp.

In total, we obtain 3D hand + object annotations for $N = 206$ frames in EgoTactile spanning $O = 25$ different objects. We then evaluate our method and compare it with SOFA physics-based simulator used in EgoPHI [2]. Both methods produce $V \times 3$ force vectors whose magnitudes are V -dim vectors. The ground truth has force magnitudes which we normalize to $[0, 1]$. We compare the true and predicted magnitude vectors with four similarity measures in Table 1. Our method clearly outperforms SOFA across all four metrics. Fig. 3 shows qualitative results for 3D as well as force estimation. Our method matches ground truth better - sometimes even producing more plausible force map than the approximate ground truth (e.g., see the base of the thumb in second row).

Evaluation of mass estimation on DexYCB [5]. Since EgoTactile does not have ground truth object physical parameters $\phi = \{m, \mu_s\}$, we evaluate the VLM for ϕ estimation on DexYCB [5]

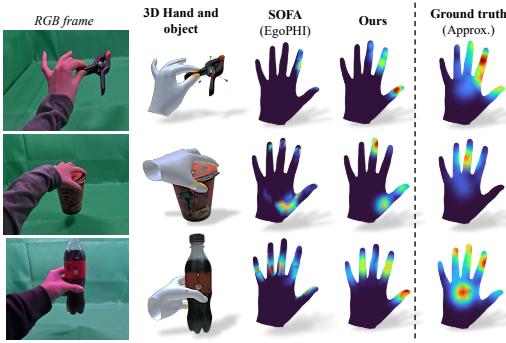


Figure 3: **EgoTactile**. Our method produces more realistic force heatmaps than the SOFA simulator, sometimes even outperforming the approximate ground truth (e.g., base of the thumb in the second row).

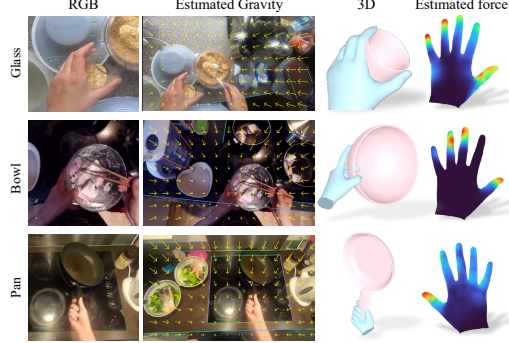


Figure 4: **In-the-wild generalization**. We demonstrate strong generalization to EPIC-Contact, using an additional heuristic to estimate gravity under unconstrained camera orientation.

which has ground truth. We experiment with Qwen 3.5 9B, Gemini 3.5 Flash and Claude Sonnet 5.5 to keep costs minimal. Claude performs the best and we use it as the default VLM for our task.

Table 1: Comparison of methods across force prediction metrics. Our method outperforms the SOFA [2] force simulator across all metrics.

Method	Pearson’s ρ ↑	Spearman’s ρ ↑	Cosine sim. ↑
Random	0.000	0.000	0.35±0.02
SOFA [2]	0.407	0.209	0.43±0.12
Ours	0.523	0.286	0.52±0.07

Table 2: Image-only estimation of mass m (in grams) and hand-object friction μ on the 21 DexYCB objects with meshes.

Model	Mass m		Friction μ	
	MAE	RMSE	MAE	RMSE
Qwen 3.5-9B	131.1	182.1	0.088	0.123
Gemini 3.5 Flash	69.6	130.4	0.093	0.111
Claude Sonnet 5.5	60.3	85.2	0.080	0.098

3.2 Qualitative generalization to other datasets

Results on samples from EPIC-Contact are shown in Fig. 4. The dataset comes with 3D annotations. Given that and our estimated $\vec{g}$ and $\phi = \{m, \mu_s\}$, our force estimation looks accurate.

Extension to video data (on DexYCB). We extend our method to consider temporal sequence of 3D poses (e.g., in case of object manipulation). We run a joint optimization process over $\mathbf{F}_t$ over all time t and include temporal smoothing regularization. An example result is shown in Fig. 5. As the hand lifts up the object, forces (particularly on the thumb) increase consistently.

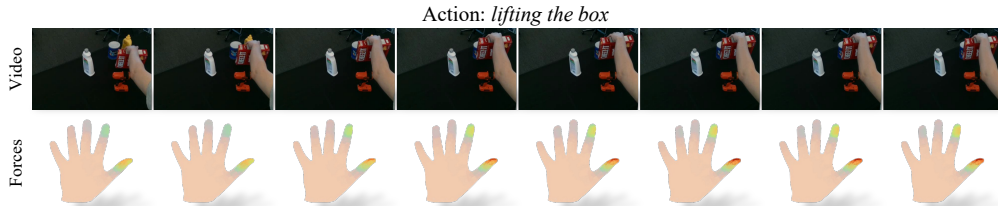


Figure 5: **Joint force optimization on temporal data**. As the hand lifts up the object, forces (particularly on the thumb) increase consistently.

4 Conclusion

We presented a simple method to simulate realistic force distributions on a hand mesh when holding an object static. We pose this as an optimization problem via Newton’s equations. This requires object properties (estimated by a VLM) and gravity (geometric heuristics). Our method outperforms the SOFA simulator on EgoTactile. We also annotate EgoTactile with 3D hand-object reconstruction which we will release publicly. Finally, our method demonstrates commendable generalization to in-the-wild samples. We envision this as a stepping stone towards applications like: (a) joint reconstruction + force estimation, (b) affordance, (c) mechanically-plausible manipulation, etc..

References

- [1] T.-H. Pham, N. Kyriazis, A. A. Argyros, and A. Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2883–2896, 2017.
- [2] A. Ilic, R. Schuchert, Y. Jiang, and C. Holz. Egophi: Estimating 3d hand-object contact and force from egocentric vision. In *European Conference on Computer Vision*, pages 410–428. Springer, 2026.
- [3] A. Sartison, D. Mironov, K. Youcef-Toumi, and D. Tsetserukou. Finger grip force estimation from video using two stream approach. *arXiv preprint arXiv:1803.01630*, 2018.
- [4] Y. Zhu, M. Hao, X. Zhu, Q. Bateux, A. Wong, and A. M. Dollar. Forces for free: Vision-based contact force estimation with a compliant hand. *Science Robotics*, 10(103):eadq5046, 2025.
- [5] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049. IEEE, 2021.
- [6] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954. IEEE, 2023.
- [7] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10118–10128. IEEE, 2021.
- [8] Y. Zeng, Y. Shi, T. Tan, X. Li, Y. Qin, Z. Lu, W. Yang, J.-H. Xue, and Q. Liao. Egotac-tile: Learning grasp pressure for everyday objects from egocentric video. *arXiv preprint arXiv:2606.09243*, 2026.
- [9] E. Dessalene, B. He, M. Maynard, Y. Tussa, P. Mantripragada, Y. Karabatis, N. Roy, and Y. Aloimonos. Feel (force-enhanced egocentric learning): A dataset for physical action understanding. In *European Conference on Computer Vision*, pages 546–563. Springer, 2026.
- [10] Y. R. Song, J. Li, R. Fu, D. Murphy, K. Zhou, R. Shiv, Y. Li, H. Xiong, C. E. Owens, Y. Du, et al. Opentouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025.
- [11] W. Zhang, C. Yuan, Z. Zhang, Z. Cheng, and Y. Gao. Egotac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026.
- [12] Y. Kanno. *Nonsmooth mechanics and convex optimization*. Crc Press, 2011.
- [13] L. Yang, X. Zhan, K. Li, W. Xu, J. Li, and C. Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11077–11086. IEEE, 2021.
- [14] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, Nov. 2017. URL <http://doi.acm.org/10.1145/3130800.3130883>.
- [15] W. Chow, J. Mao, B. Li, D. Seita, V. Campagnolo Guizilini, and Y. Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, volume 2025, pages 97959–98108, 2025.
- [16] W. Xie, M. Valentini, J. Lavering, and N. Correll. Deligrasp: Inferring object properties with llms for adaptive grasp policies. *arXiv preprint arXiv:2403.07832*, 2024.

- [17] S. Bansal, Z. Zhu, S. Tripathi, J. Zhao, M. J. Black, and D. Damen. Towards in-the-wild egocentric 3d hand-object pose estimation. In *European Conference on Computer Vision*, pages 21–39. Springer, 2026.
- [18] J.-C. Shi, B. Ye, T. Liu, X. Liu, Y. Xu, J. He, Z. Li, H. Chen, and C. Shen. Agile: Hand-object interaction reconstruction from video via agentic generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2026.
- [19] R. Pautrat, D. Barath, V. Larsson, M. R. Oswald, and M. Pollefeys. Deeplsd: Line segment detection and refinement with deep image gradients. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17327–17336. IEEE, 2023.
- [20] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris Coll-Vinent, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pages 138846–138923, 2026.
- [21] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7220–7232, 2026.
- [22] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12254. IEEE, 2025.
- [23] R. Wang, S. Xu, Y. Dong, Y. Deng, J. Xiang, Z. Lv, G. Sun, X. Tong, and J. Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026.