DexMV: Imitation Learning for Dexterous Manipulation from Human Videos

Yuzhe Qin * Yueh-Hua Wu* Shaowei Liu* Hanwen Jiang* Ruihan Yang Yang Fu
Xiaolong Wang
UC San Diego

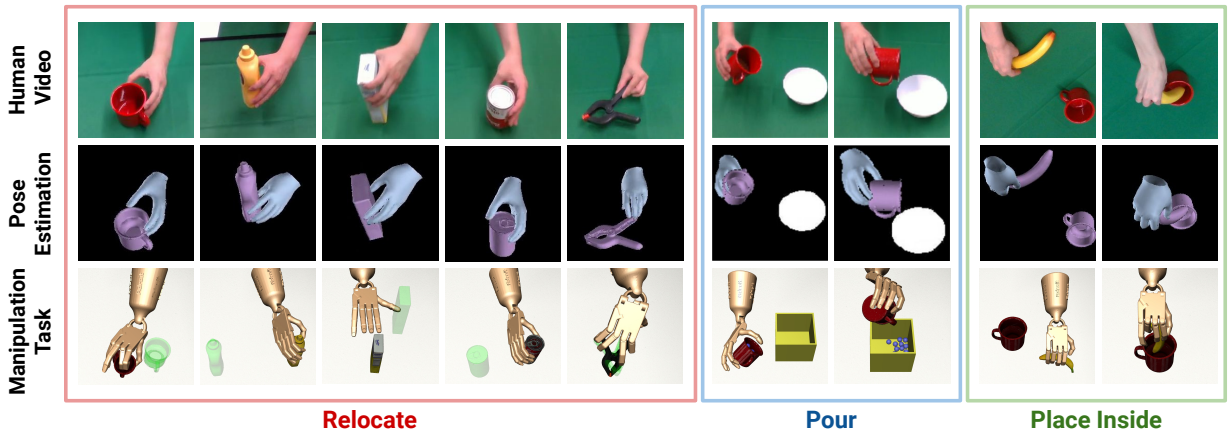


Figure 1: We propose to perform imitation learning for dexterous manipulation from human demonstration videos. We record human videos on manipulation tasks (1st row) and perform 3D hand-object pose estimations from the videos (2nd row) for constructing the demonstrations. We have a paired simulation system providing the same dexterous manipulation tasks for the multi-finger robot hand (3rd row), including the *relocate*, *pour*, and *place inside* tasks.

Abstract

*While significant progress has been made on understanding hand-object interactions in computer vision, it is still very challenging for robots to perform complex dexterous manipulation. In this paper, we propose a new platform and pipeline, DexMV (**Dex**terous **M**anipulation from **V**ideos), for imitation learning to bridge the gap between computer vision and robot learning. We design a platform with: (i) a simulation system for complex dexterous manipulation tasks with a multi-finger robot hand and (ii) a computer vision system to record large-scale demonstrations of a human hand conducting the same tasks. In our new pipeline, we extract 3D hand and object poses from the videos, and convert them to robot demonstrations via motion retargeting. We then apply and compare multiple imitation learning algorithms with the demonstrations. We show that the demonstrations can indeed improve robot learning by a large margin and solve the complex tasks. Project page with video: https://yzqin.github.io/dexmv.*

1. Introduction

Dexterous manipulation of objects is the primary means for humans to interact with the physical world. Humans perform dexterous manipulation in everyday tasks with diverse objects. To understand these tasks, in computer vision, there is significant progress on 3D hand-object pose estimation [3, 21] and affordance reasoning [1, 19]. While computer vision techniques have greatly advanced, it is still very challenging to equip robots with human-like dexterity. Recently, there has been a lot of effort on using reinforcement learning (RL) for dexterous manipulation with an anthropomorphic robot hand [10]. However, given the high Degree-of-Freedom joints and nonlinear tendon-based actuation of the multi-finger robot hand, it requires a *large amount* of training data with RL. Robot hands trained using only RL will also adopt *unnatural* behavior. Given these challenges, can we leverage humans' experience in the interaction with the physical world to guide robots, with the help of computer vision techniques?

One promising avenue is imitation learning from human demonstrations [13, 15]. In this paper, we propose **a new platform and a new imitation learning pipeline** for complex and generalizable dexterous manipulation, namely DexMV (**Dex**terous **M**anipuulation from **V**ideos). We introduce new tasks with the multi-finger robot hand (Adroit Robotic Hand [8]) on diverse objects in simulation. We collect real human hand videos performing the same tasks as demonstrations. By using human videos instead of VR, it

* Equal Contribution

largely reduces the cost for data collection and allows humans to perform more *complex and diverse* tasks. While the video demonstrations might not be optimal for perfect imitation (e.g., behavior cloning) to learn successful policies, the diverse dataset is beneficial for augmenting the training data for RL, which can learn from both successful and unsuccessful trials.

Our DexMV platform contains a paired systems with: (i) A computer vision system which records the videos of human hand performing dexterous manipulation tasks (1st row in Figure 1); (ii) A physical simulation system which provides the interactive environments for dexterous manipulation tasks with a multi-finger robot hand (3rd row in Figure 1). The two systems are aligned with the same manipulation tasks. With this platform, our goal is to bridge 3D vision and robotic dexterous manipulation via a new imitation learning pipeline.

Our DexMV pipeline contains three stages. First, given the recorded videos from our computer vision system, we extract the 3D hand-object poses from the videos (2nd row in Figure 1). Unlike previous imitation learning studies with 2-DoF grippers [22, 18], we need the human video to guide the 30-DoF robot hand to move each finger in 3D space. Parsing the 3D structure provides critical and necessary information. Second, we perform motion retargeting which converts the 3D human hand trajectories to robot hand trajectories. An optimization-based approach is proposed to align human-robot hands under kinematic constraints. Third, given the robot demonstrations, we perform imitation learning in the simulation tasks. We investigate algorithms which augment RL objectives with state-only [11] and state-action [5, 13] demonstrations.

We experiment with three types of challenging tasks with the YCB objects [2]. In our experiments, we benchmark different imitation learning algorithms and show human demonstrations improve dexterous hand manipulation by a large margin. We hope our new platform and new pipeline open up opportunities for research that connects imitation learning and 3D vision.

2. DexMV Platform

As shown in Figure 2, the DexMV platform is composed of a computer vision system to collect the videos of human perform dexterous manipulation task and a simulation system to provide the interactive environments for the same tasks with multi-finger robot hand. We will talk about the subsystem in the following paragraph.

2.1. Computer Vision System.

The computer vision system is used to collect human demonstration videos on interacting with diverse real object. In this system, we build a cubic frame (35 inch^3) and attach two RealSense D435 cameras on the top front and top left of the frame. The manipulation videos will be recorded using the two cameras as shown on the top left of Figure 2.

2.2. Simulation System.

Our simulation system is built on MuJoCo [20] with the Adroit Hand [8]. We design multiple dexterous manipulation tasks aligned with human demonstrations. As shown in the bottom row of Figure 2, we perform imitation learning by augmenting RL with the demonstrations from the computer vision system. Once the goal-conditioned policy is trained, it can be tested on achieving different manipulation goals. We will introduce the imitation learning algorithms in Section 3.3.

3. DexMV Pipeline

To bring the computer vision system and simulation system in DexMV platform, we propose a novel pipeline called DexMV pipeline. DexMV pipeline takes as input the human manipulation video collected in the computer vision system, and learn dexterous manipulation skills for a multi-finger robot. The DexMV pipeline contains three stage: (i) 3D hand-object pose estimation from videos, described in Section 3.1. (ii) Hand motion retargeting to convert human hand motion into robot motor command, described in Section 3.2. (iii) Imitation learning in the simulation environment given robot demonstration from last stages. It will be described in Section 3.3.

3.1. Hand-Object Pose Estimation

Object Pose Estimation: For each frame t in the video, we use the trained model from PVN3D [4] to detect objects and predict their 6-DoF poses. By taking both of the RGB image and the point clouds deprojected from the depth image as inputs, the model first estimates the instance segmentation mask. With dense voting on the segmented point clouds, the model then predicts the 3D location of predefined object keypoints.

Hand Pose Estimation We utilize the parametric MANO model [14] to represent the hand in a differentiable manner. Given a video, we use the off-the-shelf skin segmentation [7] and hand detection [17] methods to obtain a hand mask for every frame. We use the trained models in [9] to predict the 2D hand joints and the MANO parameters shape and pose parameters for every frame in the captured sequence using the RGB inputs. We estimate the root joint using the center of the depth image masked by segmentation.

3.2. Hand Motion Retargeting

To use the human demonstrations in our simulator, we need to convert the human hand motion to robot hand motion. Hand motion retargeting is used to map the observed human hand pose to the robot hand joints. The initial guess

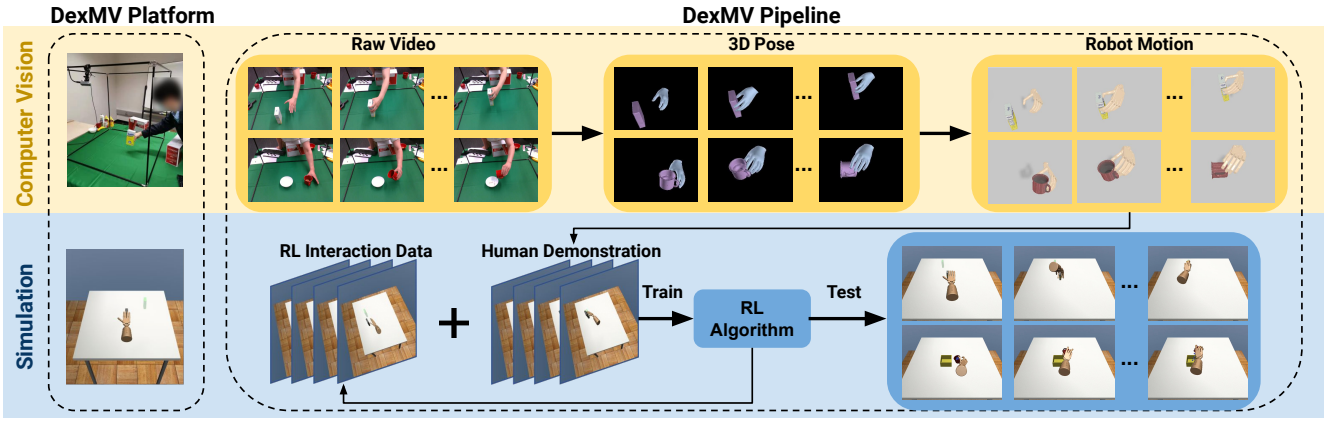


Figure 2: **DexMV platform and pipeline overview.** Our platform is composed of a computer vision system (colored with yellow) and a simulation system (colored with blue). The goal is to learn dexterous manipulation skills in our platform with DexMV pipeline.

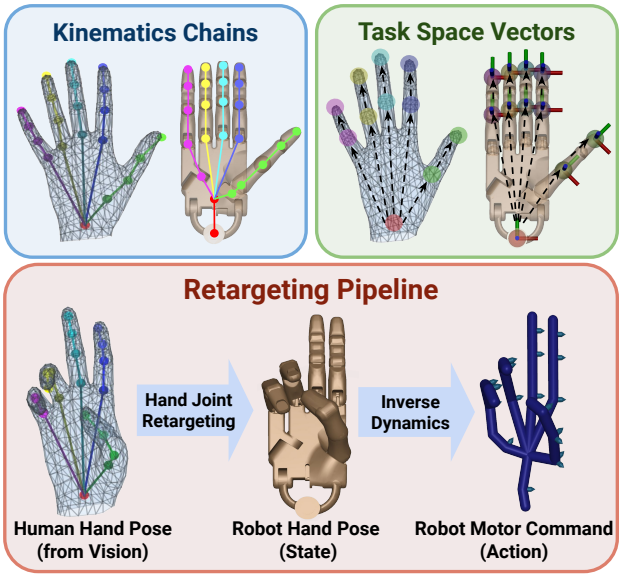


Figure 3: **Top row**: Kinematic Chains and Task Space Vectors (TSV) of human and robot hand. **Bottom row**: Two steps of hand motion retargeting.

of the robot hand pose is computed via a fitness function, which converts the human hand pose to robot joint configuration using a linear mapping. Given this initial guess, we perform another optimization approach via task space vectors. During optimization, we minimize the difference of ten task space vectors: vectors from wrist proximal phalanx plus vectors from proximal phalanx to tip for all five fingers.

3.3. Imitation Learning

State-Action Imitation Learning. We will introduce two algorithms. The first one is the Generative Adversarial Imitation Learning (GAIL) [6], which is the SOTA IL method that performs occupancy measure matching to learn parameterized policy. The key idea behind GAIL is that it uses generative adversarial training to estimate the distance and minimize it alternatively.

The second algorithm is Demo Augmented Policy Gradient (DAPG) [12]. It combines learning from demonstration and policy optimization for better sample complexity and resulting policies.

State-Only Imitation Learning. While state-action imitation approaches are shown to be effective, the action information computed from our motion retargeting approach might not be ideal. Thus we also investigate the demonstration with State-Only Imitation Learning (SOIL) [11], which extends DAPG to the state-only imitation setting and addresses the challenge by employing an inverse dynamic model h_ϕ optimizing the objective,

4. Experiment

We investigate the empirical performance of several imitation learning methods on dexterous manipulation tasks with DexMV. Specifically, the tasks are to relocate five different objects, pour and rearrange.

Implementation details. We parameterized the policy and value function in the RL methods with two separate 2-layer MLPs and Trust Region Policy Optimization (TRPO) [16] backbone. For each update iteration, we collect 200 trajectories from the environments to estimate the policy gradient and update both policy and value networks. In the following experiments, the performance is evaluated with three individual random seeds and the seeds are the same across all comparisons.

For the state-action imitation baselines that require action information such as GAIL+ and DAPG, we use inverse dynamic APIs provided by MuJoCo[20] to compute the robot action.

Methods for comparison. In the main comparison, we adopt SOIL, GAIL+, and DAPG as introduced in Section 3.3 to incorporate the demonstrations and compare the

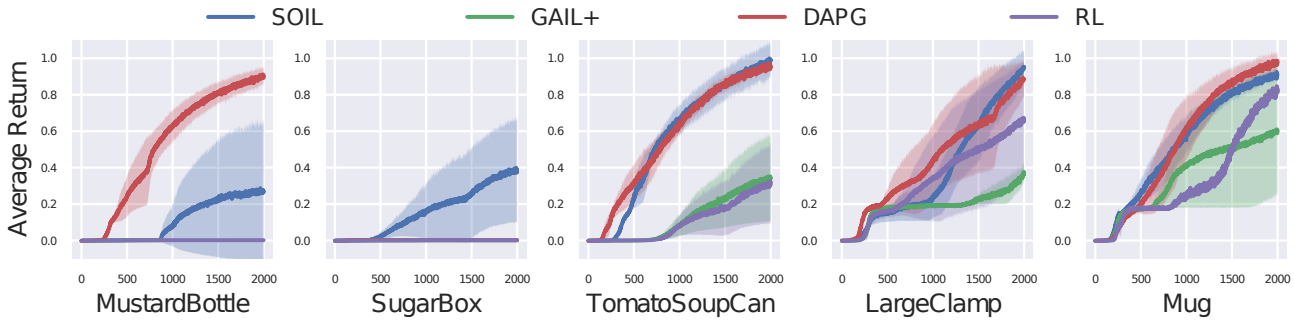
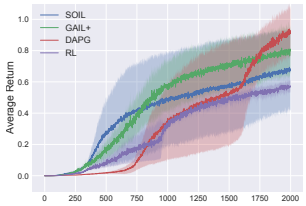


Figure 4: Learning curves of the four methods on the relocate task with respect to five different objects. The x-axis is training iterations. The shaded area indicates standard error and the performance is evaluated with three individual random seeds.

Model	Task - Relocate				
	Mustard	Sugar Box	Tomato Soup	Clamp	Mug
SOIL	0.33 ± 0.42	$\mathbf{0.67 \pm 0.47}$	0.98 ± 0.02	0.89 ± 0.15	0.71 ± 0.35
GAIL+	0.06 ± 0.01	0.00 ± 0.00	0.66 ± 0.47	0.52 ± 0.39	0.53 ± 0.37
DAPG	$\mathbf{0.93 \pm 0.05}$	0.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
RL	0.06 ± 0.01	0.00 ± 0.00	0.67 ± 0.47	0.51 ± 0.37	0.49 ± 0.36

Table 1: Success rate of the evaluated methods on the relocate task with five different objects. The success is defined based on the distance between object and target. The performance is evaluated via 100 trials for three seeds.



Model	Success (%)
SOIL	3.5 ± 3.3
GAIL+	3.4 ± 2.5
DAPG	$\mathbf{27.2 \pm 18.4}$
RL	1.3 ± 0.7

Figure 5: *Pour*. Left: learning curves; Right: table of the success rate, based on the percentage of water particles poured into the container.

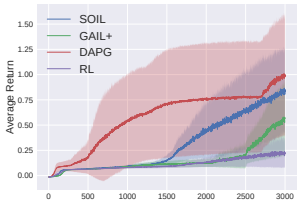
results with a pure RL algorithm. For fair comparison, we use TRPO as the policy gradient backbone to update the agent policy for all four approaches.

4.1. Experiments with Relocate

To better understand the performance of imitation learning from human demonstration, we evaluate the four methods: SOIL, GAIL, DAPG and RL on the relocate tasks. The results is presented in terms of *success rate* in Table 1 and training curve in Figure 4. The x-axis is the update iterations during training and the average-return of y-axis is normalized with the same threshold for all five relocate tasks across different seeds. A trial is counted as success only when the final position of objects after 200 iteration is within 0.1 unit length to the specified target.

4.2. Experiments with Pour

The *Pour* task involves a sequence of dexterous manipulations: reaching the mug, holding the mug, stably moving the mug, and pouring water into the container without



Model	Inside Score
SOIL	27.9 ± 26.5
GAIL+	16.0 ± 14.2
DAPG	$\mathbf{31.3 \pm 30.0}$
RL	3.2 ± 5.6

Figure 6: *Place Inside*. Left: learning curves; Right: table of the inside score, based on the volume of banana inside of the mug.

squirting. We report our results in Figure 5. We observe that DAPG converges to a good policy with much fewer iterations: 27.2% of the particles are poured into the container on average. Without the demonstrations, pure RL will only have a very small chance to pour few particles into the container.

4.3. Experiments with rearrange

The *Place Inside* task requires the robot hand to first pick up a banana, rotate it to the appropriate orientation, and place it inside the mug. To measure the performance, we define a metric *Inside Score*, which is computed based on the volume percentage of the banana inside of the mug. Since the banana is longer than the mug, we normalize the score by dividing it with 78.19% (the largest possible volume percentage inside the mug). We present our results in Figure 6, we find that DAPG outperforms other approaches whereas RL hardly learns to manipulate the object.

5. Conclusion

To the best of our knowledge, DexMV is the first work to provide a platform on computer vision/simulation systems and a pipeline on learning dexterous manipulation tasks from human videos. We benchmark multiple imitation learning algorithms and show how 3D pose estimation can affect imitation learning. We hope this provides new research opportunities for both robot learning and computer vision.

References

[1] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8709–8719, 2019. 1

[2] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv preprint arXiv:1502.03143*, 2015. 2

[3] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 1

[4] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and J. Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11629–11638, 2020. 2

[5] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NeurIPS*, 2016. 2

[6] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NeurIPS*, pages 4565–4573, 2016. 3

[7] Shah Mostafa Khaled, Md Saiful Islam, Md Golam Rabbani, Mirza Rehenuma Tabassum, Alim Ul Gias, Md Mostafa Kamal, Hossain Muhammad Muctadir, Asif Khan Shakir, Asif Imran, and Saiful Islam. Combinatorial color space models for skin detection in sub-continental human images. In *International Visual Informatics Conference*, pages 532–542. Springer, 2009. 2

[8] Vikash Kumar, Zhe Xu, and Emanuel Todorov. Fast, strong and compliant pneumatic actuation for dexterous tendon-driven hands. In *2013 IEEE international conference on robotics and automation*, pages 1512–1519. IEEE, 2013. 1, 2

[9] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 2

[10] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafał Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *arXiv*, 2018. 1

[11] Ilija Radosavovic, Xiaolong Wang, Lerrel Pinto, and Jitendra Malik. State-only imitation learning for dexterous manipulation. *IROS*, 2021. 2, 3

[12] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017. 3

[13] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. 2018. 1, 2

[14] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ToG*, 36(6):245, 2017. 2

[15] Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. *arXiv preprint arXiv:2011.06507*, 2020. 1

[16] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015. 3

[17] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2

[18] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *Robotics and Automation Letters*, 2020. 2

[19] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, pages 581–600. Springer, 2020. 1

[20] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 2, 3

[21] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *ArXiv*, abs/1711.00199, 2018. 1

[22] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. *arXiv e-prints*, pages arXiv–2008, 2020. 2