2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

CabiNet: Scaling Neural Collision Detection for Object Rearrangement with Procedural Scene Generation

Adithyavairavan Murali Arsalan Mousavian Clemens Eppner Adam Fishman Dieter Fox
NVIDIA
{admurali, amousavian, ceppner, afishman, dieterf}@nvidia.com

https://cabinet-object-rearrangement.github.io

Fig. 1: CabiNet is able to *(right)* perform complex rearrangement tasks in novel, cluttered scenes on the real robot from just partial point cloud observations without object or environment models. The model is trained with over 650K procedurally generated synthetic scenes *(left)*.

***Abstract*— We address the important problem of generalizing robotic rearrangement to clutter without any explicit object models. We first generate over 650K cluttered scenes—orders of magnitude more than prior work—in diverse everyday environments, such as cabinets and shelves. We render synthetic partial point clouds from this data and use it to train our CabiNet model architecture. CabiNet is a collision model that accepts object and scene point clouds, captured from a single-view depth observation, and predicts collisions for $SE(3)$ object poses in the scene. Our representation has a fast inference speed of 7μs/query with nearly 20% higher performance than baseline approaches in challenging environments. We use this collision model in conjunction with a Model Predictive Path Integral (MPPI) planner to generate collision-free trajectories for picking and placing in clutter. CabiNet also predicts waypoints, computed from the scene's signed distance field (SDF), that allows the robot to navigate tight spaces during rearrangement. This improves rearrangement performance by nearly 35% compared to baselines. We systematically evaluate our approach, procedurally generate simulated experiments, and demonstrate that our approach directly transfers to the real world, despite training exclusively in simulation. Supplementary material and videos of robot experiments in completely unknown scenes are available at: cabinet-object-rearrangement.github.io.**

I. Introduction

Object rearrangement is an important challenge in robotic manipulation [1, 2]. It requires the skills of picking, placing and generating complex collision-free motions in a cluttered environment. The Task-And-Motion-Planning (TAMP) [3, 4] and Embodied AI literature [5–8] have reported impressive results of rearrangement in complex human environments like kitchens. But, they are largely limited to simulation [2, 5, 6] or make the strong assumption of state estimation of the environment [9] and objects [4] in the real world. Recent neural rearrangement methods [10–14] generalize from sensed observations, without requiring state information, and have been demonstrated in the real world. Yet, they are limited to tabletop scenes [10, 11] or require expensive data collection on the real robot (1.5 years on 13 robots in the case of [14]). Overall, there are no rearrangement systems that generalizes to novel challenging environments and works out-of-the-box with minimal engineering effort for each new scene type [15]. The cost of system integration, a major task of which is environment modelling, comes up to 3X of the price of the robot itself [16]. In this work, we aim to learn a single representation, trained over 650K scenes in simulation, that enables rearrangement in diverse unknown environments.

One of the primary challenges in achieving generalization in object rearrangement is the limited availability of rearrangement datasets. Large data has been the driving force behind the success of visual learning [17] and large language models [18]. There have been some recent attempts at large-scale learning in simulation, such as the Habitat [6] and ManipulaTHOR [5] efforts. However, these actions abstract and are not realistic. For example, objects simply stick to the gripper based on proximity, thereby sidestepping the complex dynamics of pick-and-place that are explicitly addressed in prior work in the robotic grasping literature [11, 19–21]. Additionally, the policies learned in these simulated environments are unproven

979-8-3503-2365-8/23/$31.00 ©2023 IEEE

in the real world. To facilitate performance in a physical system, our approach is conditioned on 3D point clouds instead of the RGBD representation commonly used in [5, 6]. Prior work [11, 22] has demonstrated that point clouds are an effective representation to transfer from simulation-based training to real world observations.

In robotic rearrangement, a fundamental component for planning is collision detection with an unknown environment. Classical TAMP methods typically rely on the complete geometric model of the scene for planning [2, 23] in the form of a triangular mesh or Signed Distance Field (SDF). Visual reconstruction systems such as SLAM [24], KinectFusion [25] and more recently NERF [26] are needed to generate a geometric model of the scene. Each system has its drawbacks, which may include a long start-up time [25, 26], multi-view requirements [24, 26], costly updates in dynamic scenes [24–26], or poor generalization [26]. Instead of explicitly reconstructing the scene for a traditional collision checker, recent neural methods speed up collision detection with learning [27, 28] and generalize to partial real-world observations [11, 22]. [11] proposed the CollisionNet model to predict collisions from scene point clouds for 6-DOF grasping. [22] extended this to an architecture that allowed fast collision checking from point clouds, enabling fast sampling-based placing in cluttered tabletop scenes. In this work, we extend this 3D implicit representation to scale to multiple cluttered environments, which we call CabiNet and learn a SDF-based waypoint sampler from this 3D representation. We then apply a Model Predictive Path Integral (MPPI) [29] algorithm on the GPU to use our CabiNet model to generate pick-and-place motion trajectories in clutter. In summary, our contributions are as follows:

- Scaling up neural collision checking by 30X compared to prior work [22], training over nearly 60 billion collision queries. We also learn from over 650K cluttered scenes generated procedurally, which is six orders of magnitude more scene data than prior work on learning rearrangement in simulation [5]. We train a implicit 3D scene encoder CabiNet from this dataset.
- We demonstrate that CabiNet achieves fast collision detection inference of around 7μs/query and is 19.7% mAP higher than baselines when tested on 2.5 million queries in five diverse sets of environments. Using the same CabiNet encoding, we learn a scene SDF-based waypoint sampler and show that it is crucial for transitioning between pick and place actions.
- We demonstrate sim2real transfer for our model on completely unknown scenes and objects in the real world.

II. Related Work

Collision Detection from Point Clouds: There are a variety of options to check for collision with known object meshes or fully visible point clouds. One can use computational geometry libraries [23] if the object mesh is known. Alternatively, one can voxelize or spherize [30] the point clouds and formulate the collision checking problem as evaluating whether any of the elements is in collision with robot links. However, voxel-based approaches suffer from occlusion which is mitigated through use of multiple views. This unfortunately constraints the robot workspace or requires mapping of the the environments which needs to be dynamically updated as objects are moving around. SceneCollisionNet [22] frames the collision checking problem as a hybrid of classical voxel based method and data driven methods. It encodes the scene to coarse voxels where each voxel is represented by a deep embeddding vector. Each collision query is defined as a pair of query object and scene point clouds. Collision checking is done with a binary classifier which takes as input the scene voxel embedding, object embedding, and the relative $\mathrm{SE}(3)$ transformation to that voxel. SceneCollisionNet was trained on only the table top scenes. In this paper, we build on top of SceneCollisionNet. By scaling up the training data to go beyond table top settings, we observe a boost in performance and generalization across variety of different scenes.

Neural Rearrangement Planning: Traditionally, solutions to object rearrangement have been dominated by model-based methods, such as TAMP [2] which use an explicit 3D world representation estimated from sensor observations. More recently, there has been an increase in learning-based vision-centric rearrangement approaches [1], although the majority assumes simplified action spaces that abstract away the actual grasping and placing motion and often focus on navigation. In [31] a learned visual state estimator is combined with a Monte-Carlo tree search planner, to efficiently solve planar tabletop rearrangements. When goals are provided as target images, transporter networks [32], their equivariant version [33] or goal-conditioned transporter networks [34] can be used for pick-and-place. The problem of matching object instances between goal and initial image can also be simplified with vision-language models [35]. Closest to our approach for rearrangement are NeRP [10] and IFOR [12].

Large-scale Procedural Scene Generation: Probabilistic models for indoor scene generation have been originally developed in computer graphics [36–38]. Apart from appealing visually, simulating scenes physically requires more care when arranging objects. As a result most data available for learning robot manipulation in simulation is limited to a fixed number of artist designed scenes: iGibson [8], Meta-World [39], RLBench [40], Sapien [7], Habitat [6], or AI2 ManipulaTHOR [5]. Those scenes mostly consist of assets and arrangements from datasets such as PartNet [41], Replica-CAD [42], and 3D-Front [43]. More recently, ProcTHOR [44] has shown to procedurally generate large amounts of entire apartment layouts with room-specific object arrangements. Nevertheless, our use case focuses on smaller-scale clutter for which ProcTHOR's scenes are not dense enough. We will present our procedural scene generation pipeline next.

III. Procedural Data Generation

Synthetic Scene Generation: We procedurally generate synthetic data in simulation. To generate our cluttered scenes, we first assemble a set of environment assets $\mathcal{E}$, object assets $\mathcal{O}$ and fixed robot manipulator $\mathcal{R}$ (Franka Panda in our case). We have a probabilistic grammar [45] P which dictates how

the assets can be organized into random scene graphs $\mathcal{S} \sim P(\mathcal{E}, \mathcal{O}, \mathcal{R})$. This grammar is composed of the following key components: 1) sampling potential supports surfaces γ in an environment asset to place objects 2) rejection sampling to sequentially place objects on these surfaces without colliding with the scene and 3) fixing the robot base in a region where there is sufficient intersection over union (IoU > 0.8) between the workspace of the robot (approximated by a cuboid volume) and γ. Once the scene S is generated, collision queries are sampled with free-floating object meshes (computed in a straight-line trajectory) and the scene. The synthetic point clouds X are rendered online during training.

Dataset: Our dataset of object assets for training comes from ACRONYM [46], that contains wide range of object geometries from 262 categories as well as high-quality $SE(3)$ grasps which we use for picking objects. We split the dataset for training and testing. Fig 1 depicts examples of our environment assets, which we chose from common categories such as shelves, cubby, cabinet, drawers and table. All the assets are procedurally generated with the exception of shelves. For shelves, we aggregate the shelf categories from ShapeNetCore[47] and filter assets which cannot be made watertight or if proper support surfaces cannot be extracted. Nonetheless, the object placements on all the environments, including the shelves, are procedurally generated. We include more examples of scenes in the appendix. For testing, we only consider assets from the shelf environment dataset and the objects are from a novel dataset [20] unseen during training.

Rearrangement Problem Generation: A valid rearrangement problem needs a scene, a target object that the robot needs to grasp and the placement shelf that the robot needs to place the object. Given a scene, a target object is sampled if there exists a set of ground truth grasps associated with that object where they do not collide with the environment and have valid collision free inverse kinematic configuration. Once the target object is sampled, we check if a collision free placement pose for the target object exists within the placement shelf. To make sure that our rearrangement problems are challenging, the placement shelf is going to be different from the shelf that has the pick object. A problem is chosen if it passes both stages of sampling target object and also having a valid placement location. Overall, this process has a success rate of 20.8% in finding successful problems. An example of rearrangement problem is shown in Fig 3.

IV. Neural Rearrangement Planning

Our approach and model architecture is summarized in Fig 4. We first learn a implicit 3D encoding of the scene point cloud. We use the encoded scene feature along with learned object features for fast point-cloud based collision detection with CabiNet. The same scene feature is then used for predicting waypoints for the rearrangement task with a simple feedforward network. Both these models are then used to generate robot trajectories for object rearrangement with a Model Predictive Path Integral (MPPI) policy [29].

Collision Prediction: This is a learned collision model that accepts as input the scene point cloud X_S and the object point cloud X_O. The point cloud is encoded with voxelization and 3D convolution layers $\Psi_S = Enc(X_S)$. This is followed by a MLP binary classifier $c = g_\theta(\Psi_S, \Psi_O, T_{O \to S})$ that predicts if the object collide with the scene, where $T_{O \to S}$ is the relative transformation between the object and the scene and Ψ_O are the object features encoded with PointNet++ [48] layers akin to [22]. We adopt the model architecture of prior work [22] but it was only trained on table top scenes which results in poor generalization to other type of scenes such as shelves. We showed that by scaling up the training data to more diverse set of environments the model generalizes to different type of scenes. CabiNet is trained with binary cross entropy loss and with SGD with constant learning rate. Overall, we train CabiNet for two weeks, considering over 650K scenes and for 60 billion query pairs. We also modified the architecture of [22] by increasing the voxel sizes.

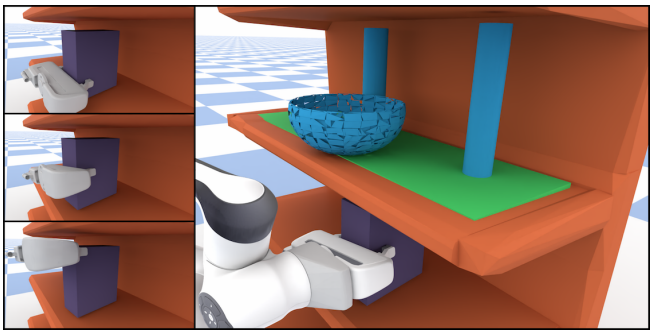

Fig. 3: An example of a procedurally generated CabiNet rearrangement scene. The target object (here in purple) is chosen if it has a selection of valid collision-free grasp poses. The green region represents the placement shelf, which is chosen if a) it is different from the shelf the object originates from and b) has a valid placement pose for the target object.

Waypoint Prediction: Object rearrangement in more constrained environments such as shelves imposes new challenges. Some of the approaches that work quite reliably in table top settings fail in navigating the tight spaces between shelves. One such approach is the work of [22] that finds the collision free path by rolling out multiple trajectories in configuration space (C-space) between the current configuration of the robot and the closest goal $\mathcal{G}$ in the C-space. These rollouts are sampled around a straight line that connects the current robot configuration to $\mathcal{G}$. Given the nominal line between the robot configuration and $\mathcal{G}$, different lines are sampled where the slope of the lines are gaussian distribution centered at the slope of nominal trajectory with a predefined variance. Each rollout is trimmed at the point of collision using SceneCollisionNet and ranked based on the distance of the rollout's final point to $\mathcal{G}$. The success of this approach hinges on the quality of the sampled trajectory. The simple sampling explained above, works quite well for table top scenes where there is more free space. However, it fails to sample promising trajectories for shelves and more constrained environments, such as when the robot needs to move from one compartment of the shelf to another. The majority of the sampled trajectories around the nominal

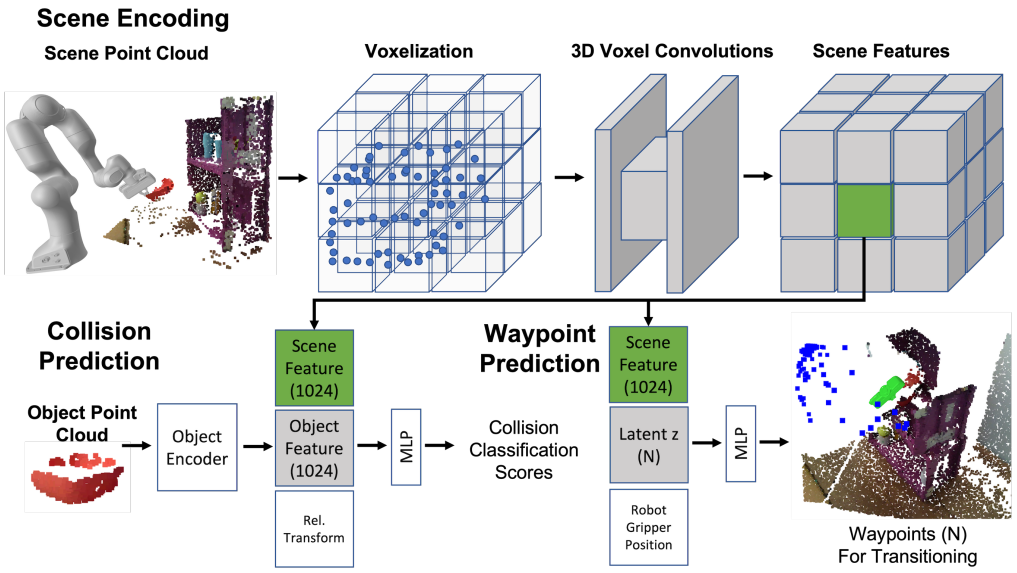

Fig. 4: Our CabiNet architecture first encodes the scene point cloud with voxelization and 3D convolutions, shown in the top. The robot is only used for visualization and the robot point cloud is removed from the scene in practice. The scene features are then used with the object features to predict scene-object collision queries. We also predict waypoints (points colored in blue) for rearrangement, conditioned on latent vector z and the current gripper position (shown in green).

trajectory would go through the divider between shelves.

To address this shortcoming, we propose to use CabiNet to sample waypoints with a larger signed distance value. Given the current end-effector position of the robot and the scene point cloud, it samples goals that gets the robot out of tight spaces. More formally, these waypoints $w \in \mathbb{R}^3$ are defined as to be in the set $\{w|\tau_{min} \leq SDF(S,w) \leq \tau_{max} \cap \|w - p_{gripper}\| \leq D\}$, where S is the scene mesh, $p_{gripper}$ is the end-effector position. The CabiNet waypoint sampler is modelled as a conditional generator $\hat{w} = f_\theta(\Psi_S, p_{gripper}, z)$ as in Generative Adversarial Networks [49]. Instead of using adversarial training, we train the sampler with Implicit Maximum Likelihood Estimation (IMLE) [50] which attempts to make each generated sample similar to a ground truth sample. We empirically found that making the loss bidirectional improved the generated samples - enforcing that ground truth samples are also similar to the set of nearest predicted samples. Let $z_1, \ldots, z_m \sim \mathcal{N}(0, I)$ denote randomly sampled latent input noise vectors and w_i the ground truth waypoints. The IMLE loss is as follows:

$$\mathcal{L}_{IMLE} = \frac{1}{n}\sum_{i=1}^{n} \min_j |\hat{w}_j - w_i| + \frac{1}{m}\sum_{j=1}^{m} \min_i |\hat{w}_j - w_i| \quad (1)$$

In our experiment, we let $\tau_{min} = 0.40$, $\tau_{max} = 0.45$, $D = 0.40$ and both m and n are 70. We let z have two dimensions and train with SGD with a constant learning rate.

Object Rearrangement: We use contact graspnet [21] to sample 6-DOF grasps for the picking action. For placing objects, we first sample potential object positions based on the scene point cloud and the support surface. The placement orientation is set to the current object pose while it is being grasped. The poses are filtered by CabiNet for collisions with the environment and followed by whether a kinematic solution can be found for the manipulator.

V. Experimental Evaluation

A. Evaluation on Collision Benchmark

We evaluate CabiNet on a collision benchmark against four baseline point cloud-based collision detection algorithms. We want to emphasize that our setting only requires a point cloud observation from a single view. We sample synthetic scene/object point cloud pair where the objects move in 64 linear trajectories in a scene. The collision ground truth information is computed with FCL [23] in simulation. For each experiment, we have a balanced set of 256K collision and collision-free queries for a total of 512K queries/experiment. We evaluated on five environments (1000 scenes each) from Fig 1 and the results are averaged across them in Table I.

Quantitative Metrics: We report the following 1) mean Average Precision (mAP) score for the classifier, averaged across the five environment test sets 2) collision prediction accuracy and 3) time/query in s. Our baselines are as follows:

- **SceneCollisionNet** [22]: We directly evaluated the pretrained model from prior work. This approach was just trained on a single environment (Tabletop) with a fixed robot-to-tabletop transformation, and directly infers collision from point clouds without any preprocessing.
- **Occupancy Mapping** [51]: This is one of the most commonly used geometric collision checking heuristic representation used in the robotics community [51–53]. We use the open source implementation from Open3D [54] to convert the sensed point cloud to a occupancy map. Voxels are specified to be of $1cm$ in size and are labelled to be either collision free or occupied.
- **Marching Cubes + FCL** [23]: We first convert the scene and object point clouds to a mesh with the marching

TABLE I: Results on Collision Benchmark

Collision Model	mAP	Accuracy (%)	Time/Query (μs)
CabiNet (Ours)	**0.971**	**89.0**	**6.41 ± 3.58**
SceneCollisionNet [22]	0.706	69.9	7.03 ± 3.89
OccupancyMap [51]	0.732	74.1	174.5 ± 61.9
MC + FCL [23]	0.767	78.8	27.5 ± 6.82
MC + SDF [55]	0.773	80.0	168.6 ± 46.4

TABLE II: CabiNet Generalization to Environments

Train Set	Test Set (AP)					
	Tabletop	Shelf	Cubby	Drawers	Cabinet	**mAP**
Tabletop	**0.989**	0.910	0.855	0.861	0.855	0.894
Shelf	0.924	**0.985**	**0.978**	0.956	0.972	0.963
Cubby	0.930	0.974	0.977	0.961	0.990	0.966
Drawers	0.923	0.856	0.848	**0.972**	0.914	0.903
Cabinet	0.924	0.971	0.961	0.961	**0.990**	0.961
All Envs	0.971	0.969	0.965	0.971	0.979	**0.971**

cubes algorithm and use FCL to compute the collision between the scene and object meshes. We parallelize this baseline across 10 processes for a fairer comparison.

- **Marching Cubes + SDF** [55]: The scene point cloud is first converted to a mesh and we fit a SDF to it using the GPU implementation from [55]. Each point in the object point cloud is computed for its SDF value from the scene. If any of the points have a negative distance (due to a penetration), the entire scene/object point cloud is considered to be in collision.

Baseline Comparisons: Overall CabiNet outperforms the baselines methods in terms of both accuracy and inference speed. It has the highest mAP across the five environment test sets. It is nearly 24% and 15% higher mAP and accuracy respectively compared to OccupancyMap [51] which is a popular method in the community, while being nearly 25x faster with a 7μs inference time. OccupancyMap performance has direct correlation with the coverage of point cloud over occluded part of the scenes. The more occluded areas in the scene, the less accurate it becomes. CabiNet, on the other hand, does not suffer from the occlusion issue since it is trained with single camera and has been learned to extrapolate to occluded parts in order to solve collision queries. CabiNet also generalizes to more diverse environments and point data compared to the pretrained SceneCollisionNet [22] which shows the importance of training on diverse set of scenes and objects. Our approach is also about 4X faster than the parallelized FCL baseline with a 20.4% higher mAP score.

Ablation on Environments: We show in Table II that CabiNet generalizes to diverse environments by training with more in-distribution data. We notice that the model generalizes to similar environments even without any training, such as cabinets and shelves. The model trained on all the environments performed the best on all the test sets.

B. Object Rearrangement Evaluation in Simulation

We evaluate our collision model in simulated rearrangement trials in IssacGym [59] as shown in Table III. To focus more on the rearrangement aspect of the problem, we used the ground truth grasp poses from [20]. We adopt a standard state-machine for rearrangement tasks used in prior works [10, 12, 60]. The objects are chosen from bowl, box, and cylinder categories of [20] and are held out from training data. For the environment assets, we use the seven shelves (from ShapeNet) in our test set to construct 30 scenes. We only use one fixed scene camera for all experiments and each scene gets two rearrangement trials, for a total of 60 experiments for each method.

Quantitative Metrics: We report three metrics on this task: 1) *overall success rate* is the success rate for the whole pick and place operation where the robot picks the target object and place it in the designated shelf without any failures. 2) *individual success rate* is the success rate of each state given the number of times the policy state machine reaches to each particular state and 3) *total time* taken for each experiment. There are three stages in rearrangement: pick, transitioning to a placing pose and place. A pick is considered a success if the object is grasped after lifting the object from the support surface. The same is true for transition. The placement is considered a success if the object is detected to be resting on the place support surface, regardless of its final orientation. We emphasize that rearrangement is a long-horizon task and conditional success rates for each state are based on the performance of the previous state. As a result, even if the performance of individual state success rates are high, errors accumulate over time leading to a lower overall success rate.

Collision Representation for Planning: We compared to Occupancy Mapping since it a popular heuristic collision representation used in the community[51]. The performance of all planners using this collision model significantly deteriorated compared to our CabiNet model. This shows the benefit of data driven approaches where they reason beyond the part of the point cloud that is visible and implicitly reason about occlusions as well. Occupancy maps is also significantly slower, increasing the rearrangement time by about 80% for the MPPI planner using the CabiNet waypoint sampler.

Comparison to Global Planning pipeline: We compare to the standard off-the-shelf motion planning pipeline used in the community with a Occupancy Mapping [51] representation. Specifically, we compare to the state-of-the-art configuration space planner [56], which is an almost-surely asymptotically optimal planner. We also compare to RRTConnect[58], which commonly is used to find feasible, though not necessarily optimal, paths. We give a timeout of 60s to find a solution for both planners. After planning, we apply spline-based, collision-aware trajectory smoothing [57] to the solutions. If the planner fails to find a valid solution, we simply execute the greedy solution to the goal to continue with the rearrangement process. Overall, the MPPI planner with our CabiNet model outperforms RRTConnect and AIT* by about 10% and 15% respectively and is also significantly faster.

Importance of Waypoints: We demonstrate that learning waypoints are crucial for rearrangement to navigate out of tight spaces. We compare to a common heuristic used in the motion generation literature to move robots out of tight spaces [61], which is to move the gripper in the reverse of the approach direction. We noticed that this approach, while

TABLE III: Simulated Rearrangement Experiments

Planner	Collision Model	Waypoint Type		
		CabiNet (Ours)	Reverse Approach	No Waypoints
MPPI	CabiNet (Ours)	**36.6%/159s**	16.7%/158s	4.3%/201s
MPPI	OccupancyMap [51]	15.0%/289s	11.0%/234s	7.5%/307s
AIT* [56, 57]	OccupancyMap [51]	21.0%/808s	18.5%/380s	5.0%/451s
RRTConnect [58]	OccupancyMap [51]	26.4%/357s	18.1%/380s	5.8%/389s

TABLE IV: Success Rate by States

Waypoint Strategy	States			
	Pick	Transition	Place	Overall
CabiNet (Ours)	**61.0%**	**80.0%**	**75.0%**	**36.6%**
Reverse Approach	54.8%	43.5%	70.0%	16.7%
No Waypoints	60.9%	21.4%	33.3%	4.3%

proficient for primitive shapes (e.g. cylinders) it does not scale to more complex shapes like bowls, which have more complicated 6-DOF grasps. Hence, retracting in the reverse of the approach direction with an object in hand, the grasped object could collide with neighbouring support surfaces while transitioning from the pick to the place poses. As shown in Table IV, our CabiNet waypoint sampler improves the transition success rate by nearly 60% compared to when having no waypoints and about 35% when compared to the strategy of retracting in the reverse of the approach direction.

C. Real Robot Experiments

We run experiments to show that our model transfers to a real robot despite only being trained in simulation.

Hardware Setup: Experiments are done on a 7-DOF Franka Panda Robot with a parallel-jaw gripper. We want to emphasize that the CabiNet model is robot-agnostic and is not conceptually limited to the franka robot. The system is equipped with two cameras: 1) Wrist mounted camera, which is an Intel Realsense D415 RGB-D camera, that is used for grasping 2) External camera, which is an Intel L515 RGB-D camera, that is used for generating the point cloud for CabiNet. Grasps are generated using Contact-GraspNet [21] and placement shelf is manually annotated for each problem by labeling the region that belongs to the desired placement shelf. We use the model from [62] for instance segmentation. The user selects the target object by clicking on the external camera image. Upon user selection of target object, the robot takes a closer look at the object and find the object in the wrist camera through relative camera pose between wrist camera and external camera. This step is crucial for grasp success since the wrist camera provides denser points on the object and it mitigates the effect of calibration imperfections. CabiNet has access to only the external camera point cloud and the inference is run on NVIDA Titan RTX gpu.

Experiment Setup: We test our approach in novel environments, with unseen shelf and objects assets in unknown poses. We experiment with three objects from different object categories and two tasks. In the vertical transport task the robot has to pick an object from the top shelf and place it in the bottom one or vice versa. Similarly for horizontal transport task the shelf compartments are horizontally next to each other. For each task-object pair, we attempt four trials, two of

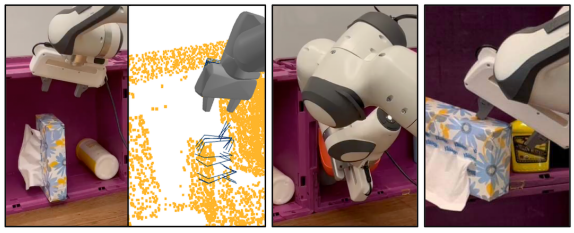

Fig. 5: Examples of failure cases, left: roof partially occluded leading to collision with wrist camera during grasping, middle: grasped object collided with barrier, right: pick failure.

TABLE V: Real Robot Experiments

Task	Object Category			
	Bowl	Box	Bottle	Overall
Vertical Transport	75.0%	50.0%	100.0%	75.0%
Horizontal Transport	50.0%	25.0%	100.0%	58.0%

which go from one support compartment to the next and two in the opposite direction. In total we have 24 experiments and each task-object pair has a unique environment, where the object has to be picked and placed amidst clutter.

Discussion: Results are reported in Table V and the vertical and horizontal transfer tasks have 75% and 58% success rates respectively. There were 2/24 pick failures due to incorrect grasps. One attempt failed during placing and 5/24 attempts failed when transitioning from pick to place states. The horizontal transfer task was relatively more challenging due to two reasons - the transitioning was over a longer distance leading to a greater chance of collision and the leftmost shelf compartment was not entirely visible from our static scene camera. Three failures were specifically due to occlusion and the CabiNet model not seeing enough of the leftmost cubby geometry from our camera setup, as shown in Fig 5. Videos of real robot execution are included in the supplementary.

VI. Conclusion

We present an effort in scaling up neural rearrangement in clutter. We train our CabiNet model to predict collisions and motion waypoints from point cloud observations. It outperforms baseline approaches in terms of collision predicted, simulated experiments and also transfers well to real world clutter despite being only trained in simulation. A limitation of our architecture is that the 3D voxelization enforces queries to be within the model workspace which can sometimes be out of bounds during manipulation. Potential extensions could explore new architectures leveraging recent learned motion policies [63] that are faster than traditional planners, along with CabiNet. Supplementary videos and material are available at: https://cabinet-object-rearrangement.github.io.

References

[1] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, M. Savva, and H. Su, "Rearrangement: A challenge for embodied ai." 2020. [Online]. Available: https://arxiv.org/abs/2011.01975

[2] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual Review of Control, Robotics, and Autonomous Systems*, 2021.

[3] S. Srivastava, E. Fang, L. Riano, R. Chitnis, S. Russell, and P. Abbeel, "Combined task and motion planning through an extensible planner-independent interface layer," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[4] C. R. Garrett, C. Paxton, T. Lozano-Pérez, L. P. Kaelbling, and D. Fox, "Online replanning in belief space for partially observable task and motion problems," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[5] K. Ehsani, W. Han, A. Herrasti, E. VanderBilt, L. Weihs, E. Kolve, A. Kembhavi, and R. Mottaghi, "Manipulathor: A framework for visual object manipulation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[7] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su, "SAPIEN: A simulated part-based interactive environment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. E. Vainio, C. Gokmen, G. Dharan, T. Jain, A. Kurenkov, K. Liu, H. Gweon, J. Wu, L. Fei-Fei, and S. Savarese, "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," in *Conference on Robot Learning*, 2021. [Online]. Available: https://openreview.net/forum?id=2uGN5jNJROR

[9] K. Wada, S. James, and A. J. Davison, "ReorientBot: Learning object reorientation for specific-posed placement," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[10] A. H. Qureshi, A. Mousavian, C. Paxton, M. Yip, and D. Fox, "Nerp: Neural rearrangement planning for unknown objects," *RSS*, 2021.

[11] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[12] A. Goyal, A. Mousavian, C. Paxton, Y.-W. Chao, B. Okorn, J. Deng, and D. Fox, "Ifor: Iterative flow minimization for robotic object rearrangement," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *RSS*, 2017.

[14] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," in *arXiv preprint arXiv:2212.06817*, 2022.

[15] J. Horst, J. Marvel, and E. Messina, "Best practices for the integration of collaborative robots into workcells within small and medium-sized manufacturing operations," *NIST Advanced Manufacturing Series*, vol. 100, no. 41, 2021.

[16] M. Belanger-Barrette, "What is an average price for a collaborative robot?" 2021. [Online]. Available: https://blog.robotiq.com/what-is-the-price-of-collaborative-robots

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. U. andNeil Houlsby, "Attention is all you need," in *In International Conference on Learning Representations*, 2021.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *arXiv:2103.00020*, 2021.

[19] J. Mahler and K. Goldberg, "Learning deep policies for robot bin picking by simulating robust grasping sequences," *Conference on Robot Learning*, 2017.

[20] A. Mousavian, C. Eppner, and D. Fox, "6-DOF GraspNet: Variational grasp generation for object manipulation," *International Conference on Computer Vision*, 2019.

[21] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," 2021.

[22] M. Danielczuk, A. Mousavian, C. Eppner, and D. Fox, "Object rearrangement using learned implicit collision functions," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6010–6017, 2021.

[23] J. Pan, S. Chitta, and D. Manocha, "Fcl: A general purpose library for collision and proximity queries," *2012 IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[24] M. Klingensmith, S. Sirinivasa, and M. Kaess, "Articulated robot motion for simultaneous localization and mapping (arm-slam)," in *Robotics and Automation Letters*. IEEE, 2016.

[25] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," *2011 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.

[26] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision (ECCV)*, 2020.

[27] J. C. Kew, B. Ichter, M. Bandari, T.-W. E. Lee, and A. Faust, "Neural collision clearance estimator for batched motion planning," 2021.

[28] N. Das and M. Yip, "Learning-based proxy collision detection for robot motion planning applications," *Transactions on Robotics*, 2021.

[29] A. A. G. Williams and E. A. Theodorou, "Model predictive path integral control: From theory to parallel computation," *Journal of Guidance, Control, and Dynamics*, 2017.

[30] P. M. Hubbard, "Approximating polyhedra with spheres for time-critical collision detection," *ACM Transactions on Graphics (TOG)*, vol. 15, no. 3, pp. 179–210, 1996.

[31] Y. Labbé, S. Zagoruyko, I. Kalevatykh, I. Laptev, J. Carpentier, M. Aubry, and J. Sivic, "Monte-carlo tree search for efficient visually guided rearrangement planning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3715–3722, 2020.

[32] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," *Conference on Robot Learning*, 2020.

[33] H. Huang, D. Wang, R. Walter, and R. Platt, "Equivariant transporter network," *arXiv preprint arXiv:2202.09400*, 2022.

[34] H. Wu, J. Ye, X. Meng, C. Paxton, and G. Chirikjian, "Transporters with visual foresight for solving unseen rearrangement tasks," *arXiv preprint arXiv:2202.10765*, 2022.

[35] W. Goodwin, S. Vaze, I. Havoutis, and I. Posner, "Semantically grounded object matching for robust robotic scene rearrangement," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 138–11 144.

[36] M. Fisher, D. Ritchie, M. Savva, T. Funkhouser, and P. Hanrahan, "Example-based synthesis of 3d object arrangements," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, pp. 1–11, 2012.

[37] L. Majerowicz, A. Shamir, A. Sheffer, and H. H. Hoos, "Filling your shelves: Synthesizing diverse style-preserving artifact arrangements," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 11, pp. 1507–1518, 2013.

[38] L.-F. Yu, S.-K. Yeung, and D. Terzopoulos, "The clutterpalette: An interactive tool for detailing indoor scenes," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 2, pp. 1138–1148, 2015.

[39] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on Robot Learning (CoRL)*, 2019.

[40] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, 2020.

[41] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[42] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[43] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao *et al.*, "3d-front: 3d furnished rooms with layouts and semantics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 933–10 942.

[44] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi *et al.*, "Procthor: Large-scale embodied ai using procedural generation," *arXiv preprint arXiv:2206.06994*, 2022.

[45] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler, "Meta-sim: Learning to generate synthetic datasets," in *International Conference on Computer Vision (ICCV)*, 2019.

[46] C. Eppner, A. Mousavian, and D. Fox, "ACRONYM: A large-scale grasp dataset based on simulation," in *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.

[47] A. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository." *Technical report, Stanford University — Princeton University — Toyota Technological Institute at Chicago*, 2015.

[48] H. S. Charles R Qi, Li Yi and L. J. Guibas, "Pointnet++:

Deep hierarchical feature learning on point sets in a metric space." *Neural Information Processing Systems (NeurIPS)*, 2017.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Neural Information Processing Systems*, 2014.

[50] K. Li and J. Malik, "Implicit maximum likelihood estimation," *arXiv preprint arXiv:1809.09087*, 2018.

[51] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[52] C. Chamzas, C. Quintero-Pena, Z. Kingston, A. Orthey, D. Rakita, M. Gleicher, M. Toussaint, and L. E. Kavraki, "Motionbenchmaker: A tool to generate and benchmark motion planning datasets," in *IEEE Robotics and Automation Letters*. IEEE, 2021.

[53] M. B. C. S. A. Hornung, K. M. Wurm and W. Burgard, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," in *Autonomous Robots*, 2013.

[54] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[55] C. Fuji Tsang, M. Shugrina, J. F. Lafleche, T. Takikawa, J. Wang, C. Loop, W. Chen, K. M. Jatavallabhula, E. Smith, A. Rozantsev, O. Perel, T. Shen, J. Gao, S. Fidler, G. State, J. Gorski, T. Xiang, J. Li, M. Li, and R. Lebaredian, "Kaolin: A pytorch library for accelerating 3d deep learning research," https://github.com/NVIDIAGameWorks/kaolin, 2022.

[56] J. D. Gammell, S. S. Srinivasa, and T. D. Barfoot, "Batch informed trees (bit*): Sampling-based optimal planning via the heuristically guided search of implicit random geometric graphs," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3067–3074, 2015.

[57] K. K. Hauser and V. Ng-Thow-Hing, "Fast smoothing of manipulator trajectories using optimal bounded-acceleration shortcuts," *2010 IEEE International Conference on Robotics and Automation*, pp. 2493–2498, 2010.

[58] J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2000.

[59] Y. G. M. L. K. S. M. M. D. H. N. R. A. A. A. H. G. S. Viktor Makoviychuk, Lukasz Wawrzyniak, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv:2108.10470*, 2021.

[60] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[61] K. Van Wyk, M. Xie, A. Li, M. A. Rana, B. Babich, B. Peele, Q. Wan, I. Akinola, B. Sundaralingam, D. Fox, B. Boots, and N. D. Ratliff, "Geometric fabrics: Generalizing classical mechanics to capture the physics of behavior," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3202–3209, 2022.

[62] A. Mousavian, L. Manuelli, B. Okorn, Y. Xiang, and C. E. an Dieter Fox, "Objectseeker: A unified framework for one-shot object detection, tracking, and instance segmentation of everyday objects," in *arXiv*, 2023.

[63] A. Fishman, A. Murali, C. Eppner, B. Peele, B. Boots, and D. Fox, "Motion policy networks," *Conference on Robot Learning (CoRL)*, 2022.