

Synthetically Trained Neural Networks for Learning Human-Readable Plans from Real-World Demonstrations

Jonathan Tremblay Thang To Artem Molchanov[†] Stephen Tyree Jan Kautz Stan Birchfield

Abstract— We present a system to infer and execute a human-readable program from a real-world demonstration. The system consists of a series of neural networks to perform perception, program generation, and program execution. Leveraging convolutional pose machines, the perception network reliably detects the bounding cuboids of objects in real images even when severely occluded, after training only on synthetic images using domain randomization. To increase the applicability of the perception network to new scenarios, the network is formulated to predict in image space rather than in world space. Additional networks detect relationships between objects, generate plans, and determine actions to reproduce a real-world demonstration. The networks are trained entirely in simulation, and the system is tested in the real world on the pick-and-place problem of stacking colored cubes using a Baxter robot.

I. INTRODUCTION

In order for robots to perform useful tasks in real-world settings, it must be easy to *communicate* the task to the robot; this includes both the desired end result and any hints as to the best means to achieve that result. In addition, the robot must be able to perform the task *robustly* with respect to changes in the state of the world, uncertainty in sensory input, and imprecision in control output.

Teaching a robot by demonstration is a powerful approach to solve these problems. With demonstrations, a user can communicate a task to the robot and provide clues as to how to best perform the task. Ideally, only a single demonstration should be needed to show the robot how to do a new task.

Unfortunately, a fundamental limitation of demonstrations is that they are concrete. If someone pours water into a glass, the intent of the demonstration remains ambiguous. Should the robot also pour water? If so, then into which glass? Should it also pour water into an adjacent mug? When should it do so? How much water should it pour? What should it do if there is no water? And so forth. Concrete actions themselves are insufficient to answer such questions. Rather, abstract concepts must be inferred from the actions.

We believe that language, with its ability to capture abstract universal concepts, is a natural solution to this problem of ambiguity. By inferring a human-readable description of the task from the demonstration, such a system allows the user to debug the output and verify whether the demonstration was interpreted correctly by the system. A human-readable description of the task can then be edited by the user to fix any errors in the interpretation. Such a description also

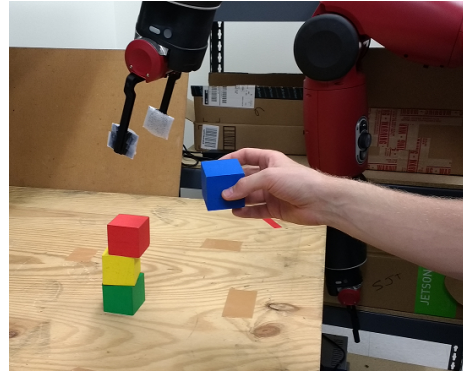


Fig. 1. In this work, a human stacks colored cubes either vertically or in a pyramid. A sequence of neural networks learns a human-readable program to be executed by a robot to recreate the demonstration.

provides qualitative insight into the expected ability of the system to leverage previous experience on novel tasks and scenarios.

In this paper we take a step in this direction by proposing a system that learns a human-readable program from a single demonstration in the real world. The learned program can then be executed in the environment with different initial conditions. The program is learned by a series of neural networks that are trained entirely in simulation, thus yielding inexpensive training data. To make the problem tractable, we focus in this work on stacking colored cubes either vertically or in pyramids, as illustrated in Fig. 1. Our system contains the following contributions:

- The system learns from a single demonstration in the real world. We believe that real-world demonstrations are more natural, being applicable to a wider set of scenarios due to the reduced system complexity required, as compared to AR/VR systems (e.g., [7]).
- The system generates human-readable plans. As demonstrated in [8], this enables the resulting plan to be verified by a human user before execution.
- The paper introduces image-centric domain randomization for training the perception network. In contrast with a world-centric approach (e.g., [27]), an image-centric network makes fewer assumptions about the camera’s position within the environment or the presence and visibility of fixed objects (such as a table), and is therefore portable to new situations without retraining.¹

The authors are affiliated with NVIDIA. Email: {jtremblay, thangt, styree, jkautz, sbirchfield}@nvidia.com

[†]Work was performed as an intern with NVIDIA. Also affiliated with the University of Southern California. Email: molchano@usc.edu

¹Recalibrating to determine a camera’s exterior orientation is arguably easier than creating a virtual environment to match a new actual environment, generating new training images, and retraining a network.

II. PREVIOUS WORK

Our work draws inspiration from recent work on one-shot imitation learning. Duan et al. [7], for example, use simulation to train a network by watching a user demonstration and replicating it with a robot. The method leverages a special neural network architecture that extensively uses soft-attention in combination with memory. During an extensive training phase in a simulated environment, the network learns to correctly repeat a demonstrated block stacking task. The complexity of the architecture, in particular the attention and memory mechanisms, support robustness when repeating the demonstration, e.g., allowing the robot to repeat a failed step. However, the intermediate representations are not designed for interpretability. As argued in [11], the ability to generate human interpretable representations is crucial for modularity and stronger generalization, and thus it is a main focus of our work.

Another closely related work by Denil et al. [5] learns programmable agents capable of executing readable programs. They consider reinforcement learning in the context of a simulated manipulator reaching task. This work draws parallels to the third component of our architecture (the execution network), which translates a human-readable plan into a closed-loop sequence of robotic actions. Further, our approach of decomposing the system is similar in spirit to the modular networks of [6].

These prior approaches operate on a low-dimensional representation of the objects in the environment and train in simulation. Like Duan et al. [7], we acquire a label-free low-dimensional representation of the world by leveraging simulation-to-reality transfer. We use the simple but effective principle of domain randomization [27] for transferring a representation learned entirely in simulation. This approach has been successfully applied in several robotic learning applications, including the aforementioned work in demonstration learning [7], as well as grasp prediction [27] and visuomotor control [16]. Building on prior work, we acquire a more detailed description of the objects in a scene using object part inference inspired by Cao et al. [3], allowing the extraction of interpretable intermediate representations and inference of additional object parameters, such as orientation. Further, we make predictions in image space, so that robust transfer to the real world requires only determining the camera’s extrinsic parameters, rather than needing to develop a simulated world to match the real environment for training.

Related work in imitation learning trains agents via demonstrations. These methods typically focus on learning a single complex task, e.g., steering a car based on human demonstrations, instead of learning how to perform one-shot replication in a multi-task scenario, e.g., repeating a specific demonstrated block stacking sequence. Behavior cloning [22], [23], [24] treats learning from demonstration as a supervised learning problem, teaching an agent to exactly replicate the behavior of an expert by learning a function from the observed state to the next expert action. This approach may suffer as errors accumulate in the agent’s actions

leading eventually to states not encountered by the expert. Inverse reinforcement learning [15], [19], [1] mitigates this problem by estimating a reward function to explain the behavior of the expert and training a policy with the learned reward mapping. It typically requires running an expensive reinforcement learning training step in the inner loop of optimization or, alternatively, applying generative adversarial networks [14], [13] or guided cost learning [10].

The conjunction of language and vision for environment understanding has a long history. Early work by Winograd [30] explored the use of language for a human to guide and interpret interactions between a computerized agent and a simulated 3D environment. Models can be learned to perform automatic image captioning [28], video captioning [25], visual question answering [12], and understanding of and reasoning about visual relationships [18], [21], [17]—all interacting with a visual scene in natural language.

Recent work has studied the grounding of natural language instructions in the context of robotics [20]. Natural language utterances and the robot’s visual observations are grounded in a symbolic space and the associated grounding graph, allowing the robot to infer the specific actions required to follow a subsequent verbal command.

Neural task programming (NTP) [31], a concurrent work to ours, achieves one-shot imitation learning with an on-line hierarchical task decomposition. An RNN-based NTP model processes a demonstration to predict the next sub-task (e.g., “pick-and-place”) and a relevant sub-sequence of the demonstration, which are recursively input to the NTP model. The base of the hierarchy is made up of primitive sub-tasks (e.g., “grip”, “move”, or “release”), and recursive sub-task prediction is made with the current observed state as input, allowing closed loop control. Like our work, the NTP model provides a human-readable program, but unlike our approach the NTP program is produced *during* execution, not before.

III. METHOD

An overview of our system is shown in Fig. 2. A camera acquires a live video feed of a scene, and the positions and relationships of objects in the scene are inferred in real time by a pair of neural networks. The resulting percepts are fed to another network that generates a plan to explain how to recreate those percepts. Finally, an execution network reads the plan and generates actions for the robot, taking into account the current state of the world in order to ensure robustness to external disturbances.

A. Perception networks

Given a single image, our perception networks infer the locations of objects in the scene and their relationships. These networks perform object detection with pose estimation, as well as relationship inference.

1) *Image-centric domain randomization*: Each object of interest in this work is modeled by its bounding cuboid consisting of up to seven visible vertices and one hidden vertex. Rather than directly mapping from images to object

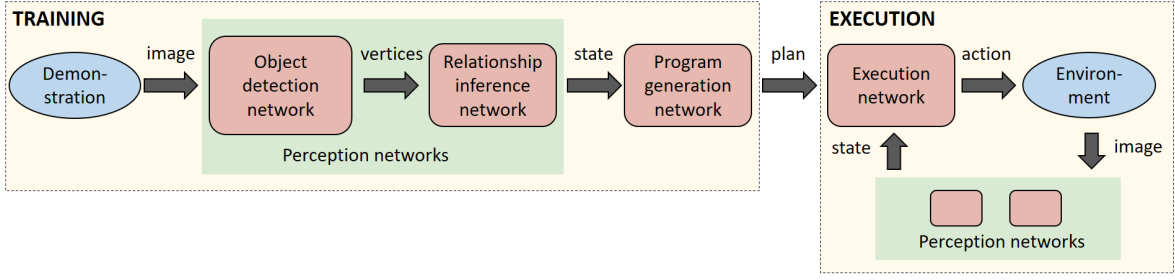


Fig. 2. System overview. As a user performs a demonstration in the real world, the perception networks detect the pose of the objects in the image domain, as well as their relationships. These percepts are fed to a network that generates a plan, which can then be implemented by the execution network.

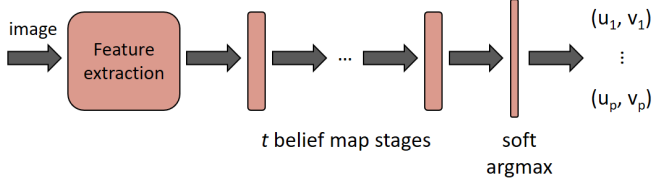


Fig. 3. Architecture of the perception network for detecting a single object from a $400 \times 400 \times 3$ RGB input image. Features are extracted using the first ten layers of VGG-19, followed by a series of belief map tensors that leverage increasingly larger amounts of context. Each $50 \times 50 \times p$ belief map tensor contains one belief map for each of the $p = 7$ vertices of the bounding cuboid. Finally, a soft argmax operation outputs the (u, v) image coordinates of each vertex. Multiple objects are detected by running multiple networks simultaneously.

world coordinates, the network outputs values in the image coordinate system. This makes the system robust to changes in camera position and orientation, as well as making it independent of the contents of the background of the scene (e.g., it does not need to see a table of a particular size, as in [27]). Using image coordinates also makes the results easier to visualize.

The network architecture is illustrated in Fig. 3. Feature extraction consists of the first ten layers of VGG-19 [26], pre-trained on Imagenet [4]. Inspired by convolutional pose machines [29], [3], the output of these layers are fed into a series of t stages. Each stage is a series of convolutional / ReLU layers (the same number of layers, type of layers, stride, and padding as in [3]) with weights that are learned during training. These stages output belief maps for each vertex, with increasingly larger receptive fields to capture more of the surrounding context and resolve ambiguity. These stages are depicted in Fig. 4 for the prediction of a single vertex.

We use the L_2 loss between the predicted belief maps and the ground truth of the training data. Specifically, the loss function for stage i is

$$f_i = \sum_{j=1}^p \left[\sum_{u,v} \left\| B_j^i(u, v) - \hat{B}_j(u, v) \right\|_2 \right], \quad (1)$$

where \hat{B}_j is a Gaussian-smoothed ($\sigma = 3.4$ pixels in the 400×400 image) ground truth belief map for vertex $j \in \{1..p\}$ and B_j^i is the neural network output for the belief map at stage $i \in \{1..t\}$ for vertex j , where $p = 7$ is the number

of vertices estimated. The total loss is the sum of the losses of the individual stages: $f = \sum_{i=1}^t f_i$. This approach of applying the loss at each stage (also known as *intermediate supervision* [29]) avoids the vanishing gradient problem by restoring the gradients at each stage.

Each belief map is treated as a probability mass function for the location of a vertex in the image domain. To obtain image coordinates, soft argmax is applied along the rows and columns of the final belief map.

Examples of cuboid object detection are shown in Fig. 5. The image-centric representation makes it easy to visualize whether the object has been detected accurately. In addition, detecting the individual vertices yields a rich representation to facilitate estimation of the full pose of the object. By training on instances of occluded objects, the network learns to detect the object even when it is severely occluded, as shown in the figure.

The network is trained entirely on simulated data. Colored cubes (5 cm per side) were placed within a virtual environment consisting of various distractor objects in front of a background. Images were generated by randomly varying distractors, texture, background, camera pose, lighting, and noise. For each color, 12k synthetic images were collected for training.

2) *Relationship inference*: After objects have been detected, their relationships are inferred. This is accomplished via a fully connected neural network with 28 inputs,² two outputs, and three hidden layers with 100 units in each. The inputs to the network are the image coordinates of the vertices of two detected cuboids, and the output is an unnormalized distribution over the symbols ABOVE and LEFT; values below a threshold yield a third symbol, NONE. This set is rich enough to allow the system to build not only stacks of cubes but also more complicated structures such as pyramids. With n detected objects, the pairwise network is run $n(n-1)$ times to generate the full state of the scene from a single image. The state is represented as an $n \times n \times k$ tensor (with the diagonal ignored), where $k = 3$ is the number of relationships considered.³ Note that in the case of a pyramid,

²That is, 2 objects, 7 vertices per object, and (u, v) coordinates for each vertex.

³Although such a representation could lead to self-contradiction (e.g., blue is above red, and red is simultaneously above blue), it also allows the possibility of discarding inferences with low confidence.

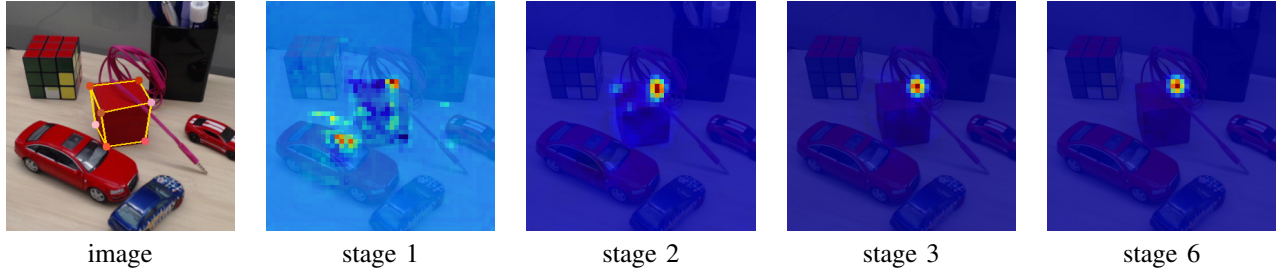


Fig. 4. An image with detected vertices overlaid, and belief maps B_j^t for stages $t \in \{1, 2, 3, 6\}$. Ambiguity in the early stages due to small receptive fields is resolved in later stages by taking into account more context.

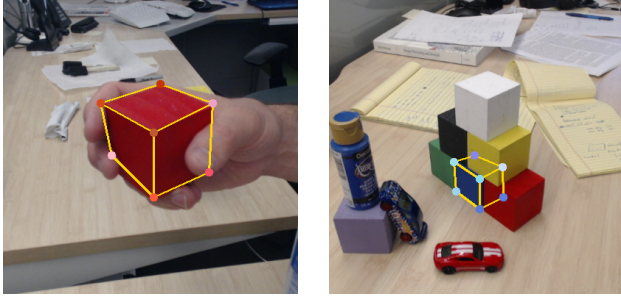


Fig. 5. Examples of object detection from image-centric domain randomization, showing the seven detected vertices. Shown are detections of a red cube (left) and blue cube (right), even when the latter is severely occluded.

an object is above two other objects, which is manifested by two elements of the tensor having roughly equal weight.⁴

The relationship network was trained on the vertex coordinates of the simulated data mentioned above with a cross-entropy loss function. To make the network more robust to real-world noisy inputs, these coordinates were perturbed by randomly swapping vertices (with probability 1%) to another vertex within the same cube and adding Gaussian noise to the observed vertex coordinates. Moreover, occluded vertices were randomly relocated (with probability 50%) using a uniform distribution within the convex hull of the occluder.

B. Program generation network

The purpose of the system is to learn a human-readable program from a real-world demonstration. While the camera watches the scene, an agent (such as a person) moves the blocks on the table to stack them vertically or in a pyramid structure. Multiple stacks are allowed, in which case the order of operations between the different stacks is arbitrary.

As the demonstration is being performed, the perception network detects the objects and their relationships. Once the demonstration is complete, the state tensor from the relationship inference is thresholded (at 0.5) to yield a set of discrete relationships between the objects. This tensor is sent to a program generation network which outputs a human-readable plan to execute.

Our framework assumes that the demonstration involves a sequence of pick-and-place operations. Each step of the program can therefore be represented as a $2 \times (n+1)$ binary

array of values indicating which of the n objects (or none) is the source (the object to be picked), and which of the n objects (or none) is the target (the object upon which the picked object is to be placed). Since with n objects, there are at most $n - 1$ steps in the program, the output of the program generation network is a $2 \times (n+1) \times (n-1)$ tensor of floating-point values which are converted to binary by applying argmax. The resulting binary tensor is then trivially converted to a human-readable program. For example, the following binary tensor

		<i>place</i>			
		red	green	blue	yellow
<i>pick</i>	red	-	10	00	00
	green	00	-	00	00
	blue	01	00	-	00
	yellow	00	00	00	-

would be translated as follows:

“Place the red cube on the green cube, then place the blue cube on the red cube.”

The network is implemented as a fully connected double-headed neural network with seven layers (one input, one output, and five hidden) in each of the two paths, with 1024 units in each hidden layer. The network is trained using an MSE loss function with data from simulated scenes generated by enumerating possible states and corresponding plans.

C. Execution network

Once a program has been generated, it could be executed in an open-loop fashion by sequentially performing each step. However, in order to allow recovery from manipulation mistakes, or to handle external disturbances, a network is used to guide execution. This execution network is fully connected with two inputs, one output, and five hidden layers with 128 units in each. It takes as input the program (represented as a $2 \times (n+1) \times (n-1)$ tensor) and the current state of the scene (represented as an $n \times n \times k$ tensor), and it returns the next action to take. This action is represented by a $2 \times n$ array of values indicating the source and target objects, along with a $1 \times (k-1)$ array indicating the relationship to achieve (e.g., place one object on top of another). The execution network is trained on synthetic data generated by randomly sampling among the possible programs and states, with an MSE loss function.

⁴To improve robustness, we trained a separate relationship inference network to handle this case.

D. Entire system

The different networks presented in this section are linked together to create a combined system for learning by demonstration. The human demonstrates a task once in the real world, from which the system infers a program. Once the demonstration is complete, the objects may be shuffled on the working surface to change the initial conditions, and the robot can then execute the program in a closed loop manner, correcting mistakes when operations fail or perturbations are introduced.

IV. EXPERIMENTAL RESULTS

In this section we report tests of the individual components of the system, followed by tests of the entire end-to-end system.

A. Cube detection

We begin by assessing how accurately the object detection network, trained using image-centric domain randomization, can detect the location and pose of colored cubes in real images. We collected a dataset of images of cubes in various arrangements on a table. Images were captured to test the effects of (a) cube color, (b) position of the cube on the table, (c) height of the cube above the table, (d) presence of other distractor cubes, (e) occlusion, and (f) position and orientation of the camera with respect to the cube. The images were manually labeled to indicate the vertex locations, and the output of the detection network was compared with these ground truth labels.

Several measures of accuracy could be used. For an interpretable measure of the accuracy with which points can be detected, either the mean absolute error (MAE) or root mean square (RMS) error are possible. For determining how likely a point is detected within some radius of the ground truth, variants of the PCP metric [9] such as PCK [32] or PCKh [2] are common choices.⁵

In this work, to measure the expected error in pixels, we use the MAE rather than RMS since it is less sensitive to outliers. To determine the probability that a vertex will be detected correctly, we use PCKh but modified to normalize the error by the size of the cube in the image (the square root of the number of pixels in the convex hull) instead of the size of the human head. We also introduce a third measure, MAEc, which is the MAE of the vertices that are correctly detected according to PCKh (thus ignoring outliers). All three measures can be expressed as

$$\frac{\sum_{i=1}^N \phi_i \delta_i}{\sum_{i=1}^N \psi_i}, \quad (2)$$

where N is the total number of vertices in all the images being used for evaluation, and δ_i , ϕ_i , and ψ_i are given by:

⁵In the context of human pose estimation, PCP is the percentage of correctly localized *parts* with respect to the length of the ground truth segment; PCK is the percentage of correctly localized *keypoints* with respect to the size of the ground truth bounding box of the person; PCKh is like PCK except with respect to the size of the head rather than the person to reduce the effects of body pose. In the latter two cases, the size of the bounding box is the maximum of length and width.



Fig. 6. One image from each of the six subsets of the data used in testing cube detection. In lexicographic order: isolated, cluttered, above table, occluded, stacked, and pyramid.

	δ_i	ϕ_i	ψ_i
MAE	d_i	1	1
PCKh	1	c_i	1
MAEc	d_i	c_i	c_i

where $c_i = 1$ if $(d_i/\sqrt{A}) \leq \epsilon$ or 0 otherwise, $d_i = \sqrt{(u_i - \hat{u}_i)^2 + (v_i - \hat{v}_i)^2}$ is the Euclidean distance in pixel space between the detected vertex and the ground truth, A is the number of pixels in the convex hull of the cube, and ϵ is a threshold.

The dataset of 192 images was split into six subsets, as illustrated in Fig. 6. Images containing an isolated cube on a table, clutter on the table near the cube (including cubes of other colors), a cube elevated above the table by resting on another object, partial occlusion of a cube, inclusion in a stack of cubes, and inclusion in a pyramid of cubes. The subsets consist of images with an equal number of annotations of red, green, blue, and yellow cubes, as well as a proportional number of stack sizes.

The results are shown in Table I where the false negative rate (FNR) is the percentage of cubes not detected. Over all six scenarios, the MAE was 5.8 pixels, and PCKh@0.5 was 90.0%. In other words, 90.0% of the vertices were detected within half the width of a square containing the same number of pixels as the convex hull of the ground truth cube. Ignoring the 10% outliers, the MAE was consistently 3.8 pixels (that is, MAEc= 3.8). This result is remarkable, because all errors were measured with respect to the 400×400 image input to the network, whereas the final stages of the network output belief maps of size 50×50 . As a result, this means that the MAE of the inliers was actually less than 0.5 pixels with respect to the belief maps. Of the 219 cubes in the dataset, 3 were not detected at all by the network. The rest of the numbers are with respect to the 216 detected cubes. Results are also fairly consistent across various bright colors, as shown in Table II, although we experienced more difficulty with black or white cubes.

Although for practical reasons this work focuses on perceiving and manipulating solid cubes, the perception network as described is applicable to any rigid real-world object that can be reasonably approximated by its 3D bounding cuboid. Shown in Fig. 7 is the output of the same network applied to a toy car, after training on simplistically simulated car

TABLE I

RESULTS OF THE PERCEPTION NETWORK ON THE SIX SCENARIOS.

	MAE (pixels)	PCKh@0.5 (%)	MAEc@0.5 (pixels)	FNR (#/cubes)
isolated	4.5	90.6%	3.9	0/20
cluttered	4.6	90.9%	4.0	2/24
above table	4.0	97.7%	3.7	1/20
occluded	9.5	83.6%	3.6	0/24
stacked	6.4	91.3%	3.7	0/96
pyramid	4.9	83.7%	3.8	0/35
all	5.8	90.0%	3.8	3/219

TABLE II

RESULTS OF THE PERCEPTION NETWORK BY CUBE COLOR.

	MAE (pixels)	PCKh@0.5 (%)	MAEc@0.5 (pixels)	FNR (#/cubes)
red	4.1	92.1%	3.6	2/55
green	4.8	86.7%	4.0	1/55
blue	5.8	93.2%	3.6	0/54
yellow	8.4	87.9%	3.8	0/55

images in the manner described above. Unlike colored cubes, for which each face has identical appearance, objects such as cars have distinctive sides (front / back / left / right / top / bottom). Although not obvious in the figure, the network is able to consistently maintain the identity of the vertices regardless of the relative roll, pitch, and yaw of the object with respect to the camera.

B. Relationship inference

We evaluated the performance of the relationship inference network using a hand-labeled dataset of scenes containing four cubes arranged using combinations of relationship types ({ABOVE, LEFT, ABOVE-PYRAMID, NONE}), with some images containing multiple relationships and others none at all. Across the 35 images in the dataset, there were 72 valid pairwise relationships among a total of 420 cube pairs under evaluation by the network. Cube vertices were detected using the cube detection networks discussed previously.

Results are shown in Table III. The false positive rate (FPR) indicates the rate at which non-adjacent cubes are

TABLE III

RELATIONSHIP INFERENCE PERFORMANCE WHEN TRAINED WITH DIFFERENT AUGMENTATION METHODS.

Augmentation	FPR	FNR
None	1.4%	22.2%
independent Gaussian ($\sigma = 10^{-3}$)	1.4%	22.2%
structured Gaussian ($\sigma = 5^{-4}$)	1.6%	20.8%
vertex confusion (1%)	1.7%	19.4%
vertex occlusion	1.3%	13.9%
all	1.9%	13.9%

predicted to be adjacent by the relationship network, while the false negative rate (FNR) indicates the rate at which valid relationships are missed.

The table evaluates four types of data augmentation, applied to the ground truth cube vertices in simulated training data: (1) independent Gaussian perturbation applied per vertex; (2) structured Gaussian perturbation applied per cube; (3) vertex confusion, whereby a vertex is randomly assigned to the location of another on the same cube; and (4) vertex occlusion, whereby vertices that overlap with a different cube are assigned a new random location with high variance.

Despite training only on simulated data, the relationship network transfers readily to the real world data. The best network achieves FPR = 1.3% with FNR = 13.9% when evaluated with a 0.5 confidence threshold. The most effective augmentation (in fact, the only one that makes improvements to FNR and FPR) is simulated vertex occlusion. The other augmentations may be less necessary given the variation in viewing angles, distances, and random cube rotations and alignments already present in the simulated data, combined with the reliability of the detection network.

C. Program generation & execution

The generalization performance of the program generation network was evaluated by testing its accuracy with respect to the percentage of all possible stacks used for training data.⁶ Fig. 8 shows the accuracy (number of times the network predicted the correct pick or place) versus training data size for several different networks. The smallest network (red dotted line), with 2 hidden layers and 128 units per hidden layer, is not even able to memorize the dataset (since it achieves just 83% accuracy even with 95% training data). The largest network (blue solid line), with 6 hidden layers and 1024 units per hidden layer, is not only able to memorize the dataset but also achieves reasonable generalization performance of 71% when trained on 80% of the dataset. Since $w = 512$ (not shown to avoid clutter) achieves similar results to $w = 128$, we conclude that 1024 units are necessary. However, although the number of hidden units is crucial to good performance, the number of hidden layers is less so; in fact, 4 layers outperforms 6.

Fig. 9 shows the program learned by the network to reproduce a three-cube pyramid. What is interesting about

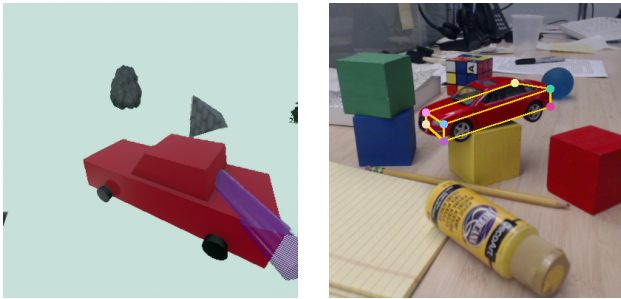


Fig. 7. Example of detecting the bounding cuboid of an object that does not have a simple geometric shape. Shown are a synthetic training image of a car (left), and a toy car detected in a real image (right) using the same method described for colored cubes.

⁶Since the percepts are discrete, all possible stacks can be enumerated.

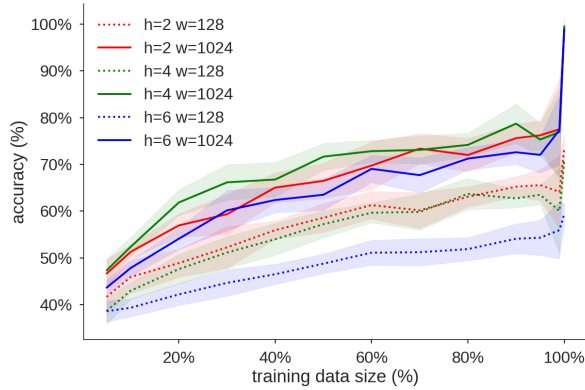


Fig. 8. Comparison of various neural network architecture sizes for predicting valid programs (h : number of hidden layers, w : number of units per hidden layer). The x -axis represents the percentage of data used during training (ranging from 1% to 95%), whereas the y -axis is the accuracy on the entire dataset. The best results are achieved with $h = 4$, $w = 1024$ (solid green line).

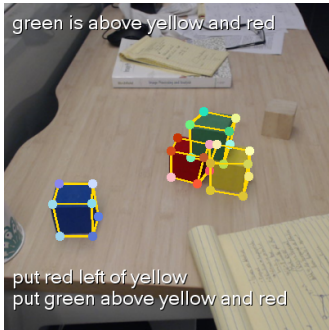


Fig. 9. The correct program (text at the bottom of the image) is learned in this case, even though the state is ambiguous (text at the top of the image) because it is missing the LEFT relationship.

this case is that the state is ambiguous since the relationship network failed to predict that the red cube is left of the yellow cube, perhaps due to the large gap between the two cubes. Nevertheless, the network has learned to infer that there should be a third relationship in the pyramid and returns a program that first creates a left relationship, before indicating how the green cube should be stacked. Of course, in this case the network does not have enough information to decide which of the two symmetric left relationships (“red left of yellow” or “yellow left of red”) is correct, since the coordinates of the cubes are not available. But the network’s guess indicates behavior that is more than mere memorization of the training data.

The execution network was tested in a similar manner by measuring the influence of withholding training data. Table IV presents the accuracy versus the amount of data withheld. The data consisted of approximately 36k data points from all possible arrangements of 2- to 6-cube stacks. Each result is the average over the five last epochs from five independently trained networks. When the network is trained on the whole collection, it achieves perfect accuracy, as expected, but it also works surprisingly well when only

TABLE IV
EXECUTION NETWORK PERFORMANCE (ON ENTIRE DATASET) AS A
FUNCTION OF THE AMOUNT OF DATA USED FOR TRAINING.

training data	5%	10%	50%	95%
accuracy	56%	85%	97%	98%

half the data is present for training, thus demonstrating generalization.

D. Entire system

The networks were connected and tested in our lab with a Baxter robot. For perception, we used an Intel RealSense SR300 camera, whose exterior orientation with respect to the robot was determined using a standard calibration approach. RGB images were used for perception, while depth images were used to determine the 3D position of the cubes with respect to the robot.

To test the system, a human stacked the cubes together, then pressed a button to indicate that the demonstration was complete. The system automatically inferred the positions of the cubes, their relationships, and the human-readable program. After the human rearranged the cubes and pressed another button, the robot then performed the pick-and-place steps output by the execution network. In addition to cubes, we also performed experiments involving a toy car (mentioned earlier) placed on a cube. Some examples are shown in Fig. 10.

Qualitatively, the system worked fairly reliably, but a few limitations are worth mentioning. First, in our setup the intersection of the robot arm’s working area and the camera’s field of view yielded an effective working area approximately 20 cm by 30 cm, which was rather small. Another challenge involved detecting the severely occluded cubes at the bottom of the stack; coupled with the fact that the execution network does not retain any memory of its previous actions, any such missed detections led to incorrect behavior in the execution. Obviously, open loop execution would solve this problem but at the expense of limiting generalizability of the system. Moreover, the open gripper was only 9 cm wide, leading to just a 2 cm tolerance for grasping the 5 cm cubes, which was close to the accuracy of our detection system; as a result, the gripper would occasionally fail to envelope the intended cube. Finally, even though all cubes were simply painted wooden blocks, the depth sensor of the RealSense camera in our lab struggled inexplicably to obtain meaningful measurements of the green cube at certain angles, thus further challenging the system.

V. CONCLUSION

We have presented a system to generate human-readable programs from a real-world demonstration. The system consists of a sequence of neural networks to perform tasks associated with perception, program generation, and program execution. For perception, we introduce image-centric domain randomization leveraging convolutional pose machines,

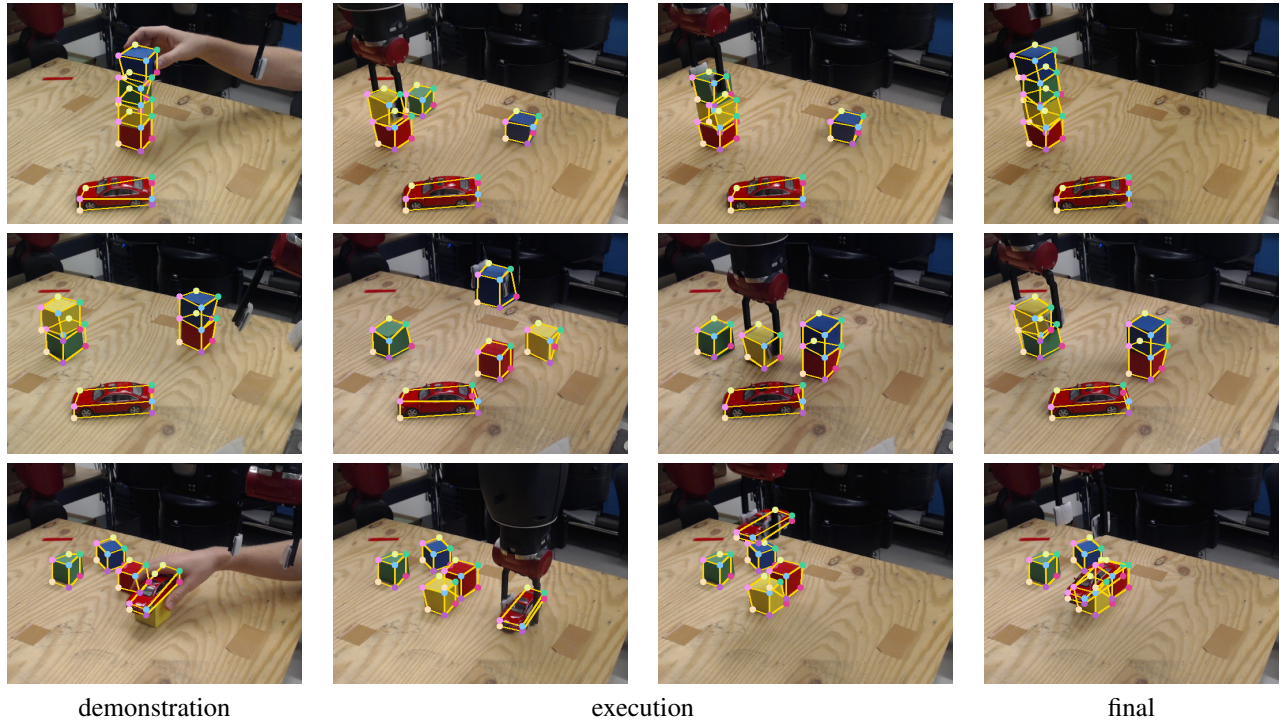


Fig. 10. Three demonstrations (left column), and snapshots of the robot executing the automatically generated program (remaining columns). From top to bottom, the programs are as follows: “Place yellow on red, green on yellow, and blue on green” (top row), “Place yellow on green, and blue on red” (middle row), and “Place car on yellow” (bottom row). In the second row, notice that the robot recovered from an error in initially misplacing the yellow cube.

which results in a vision-based network that can be applied to any camera, without assumptions about the camera pose or the presence of specific background features in the scene. For program generation and execution, we show that fully connected networks, despite their simplicity, generalize surprisingly well when considering relationships, states, and programs not encountered during training. Although training individual networks separately may be suboptimal compared with end-to-end training, it facilitates component-wise testing, interpretability, and modularity. There remain many issues to explore, such as increasing robustness of domain randomization to variations in lighting and color, incorporating context to better handle occlusion, leveraging past execution information to overcome limitations in sensing, and expanding the vocabulary of the human-readable programs.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful comments. Appreciation also goes to Omer Shapira, Mark Brophy, Hai Loc Lu, and Bryan Dudesch for sharing their expertise and insights regarding the design and implementation of the simulator used for domain randomization, as well as for many helpful discussions.

REFERENCES

- [1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning, in *Proceedings of the International Conference on Machine Learning (ICML)*, July 2004.
- [2] M. Andriluka, L. Pishchulin, P. Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis, in *CVPR*, June 2014.
- [3] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields, in *CVPR*, June 2017.
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. ImageNet: A large-scale hierarchical image database, in *CVPR*, 2009.
- [5] M. Denil, S. G. Colmenarejo, S. Cabi, D. Saxton, and N. de Freitas. Programmable agents, arXiv preprint, arXiv:1706.06383, 2017.
- [6] C. Devin, A. Gupta, T. Darrell, P. Abbeel, S. Levine. Learning modular neural network policies for multi-task and multi-robot transfer, in *ICRA*, 2017.
- [7] Y. Duan, M. Andrychowicz, B. C. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, W. Zaremba. One-shot imitation learning, in *NIPS*, Dec. 2017.
- [8] A. Feniello, H. Dang, and S. Birchfield. Program synthesis by examples for object repositioning tasks, in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Sept. 2014.
- [9] V. Ferrari, M. Marín-Jiménez, and Andrew Zisserman. Progressive search space reduction for human pose estimation, in *CVPR*, June 2008.
- [10] C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization, in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 49–58, June 2016.
- [11] M. Garnelo, K. Arulkumaran, and M. Shanahan. Towards deep symbolic reinforcement learning, in *NIPS Deep Reinforcement Learning Workshop*, 2016.
- [12] T. Gupta, K. Shih, S. Singh, D. Hoiem. Aligned image-word representations improve inductive transfer across vision-language tasks, in *ICCV* 2017.
- [13] K. Hausman, Y. Chebotar, S. Schaal, G. S. Sukhatme, and J. Lim. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets, in *NIPS*, 2017.
- [14] J. Ho and S. Ermon. Generative adversarial imitation learning, in *Advances in Neural Information Processing Systems (NIPS)*, pp. 4565–4573, 2016.
- [15] H. H. Hu, A. A. Gooch, W. B. Thompson, B. E. Smits, J. J. Rieser, and

- P. Shirley. Visual cues for imminent object contact in realistic virtual environment, in *Proceedings of the Conference on Visualization*, pp. 179–185, 2000.
- [16] S. James, A. Davison, and E. Johns. Transferring end-to-end visiomotor control from simulation to real world for a multi-stage task, in *Conference on Robot Learning (CoRL)*, 2017.
 - [17] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, in *CVPR*, 2017.
 - [18] A. Karpathy, L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions, in *CVPR*, June 2015.
 - [19] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning, in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pp. 663–670, June 2000.
 - [20] R. Paul, A. Barbu, S. Felshin, B. Katz, N. Roy. Temporal grounding graphs for language understanding with accrued visual-linguistic context, in *IJCAI*, 2017.
 - [21] J. Peyre, I. Laptev, C. Schmid, J. Sivic. Weakly-supervised learning of visual relations, in *ICCV*, 2017.
 - [22] D. Pomerleau. Efficient training of artificial neural networks for autonomous navigation, *Neural Computation*, 3(1):88–97, 1991.
 - [23] S. Ross and D. Bagnell. Efficient reductions for imitation learning, in *AISTATS*, pp. 661–668, 2010.
 - [24] S. Ross, G. J. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning, in *AISTATS*, pp. 627–635, April 2011.
 - [25] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, X. Xue. Weakly supervised dense video captioning, in *CVPR*, 2017.
 - [26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, in *ICLR*, 2015.
 - [27] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world, in *IROS*, 2017.
 - [28] O. Vinyals, A. Toshev, S. Bengio, D. Erhan. Show and tell: A neural image caption generator, in *CVPR*, June 2015.
 - [29] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh. Convolutional pose machines, in *CVPR*, 2016.
 - [30] T. Winograd. Procedures as a representation for data in a computer program for understanding natural language, MIT, No. MAC-TR-84, 1971.
 - [31] D. Xu, S. Nair, Y. Zhu, J. Gao, A. Garg, L. Fei-Fei, S. Savarese. Neural task programming: Learning to generalize across hierarchical tasks, in arXiv 1710.01813, 2017.
 - [32] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *PAMI*, 35(12):2878–2890, Dec. 2013.