Automatic Data Generation for SORNet: PROPS Relation Dataset

1st Jace Aldrich
Robotics Department
University of Michigan
Ann Arbor, MI, USA
jacealdr@umich.edu

2nd Ariana Verges Alicea
Robotics Department
University of Michigan
Ann Arbor, MI, USA
alarian@umich.edu

3rd Hannah Ho
Robotics Department
University of Michigan
Ann Arbor, MI, USA
hdho@umich.edu

***Abstract*—SORNet (Spatial Object-Centric Representation Network) is a network architecture that takes an RBG image with several canonical object views and outputs object-centric embeddings. The authors of SORnet trained and tested on their custom Leonardo and Kitchen datasets, as well as the CLEVR dataset. We expanded SORnet's capability by training it on PROPS Dataset, which was extensively used throughout this course. Training SORNet with PROPS dataset allow us to test its capabilities to a real-world dataset in order to better understand how it performs in real-life applications. Our Dataset and its code can be found publicly at https://github.com/Jaldrich2426/PROPS-Relation-Dataset, and a modification of the SORNet code to run it can be found at https://github.com/Jaldrich2426/sornet**

I. Introduction

There are a plethora of applications for robots that can perform sequential tasks that involve manipulating objects around them. These tasks can range from object assembly to organizing and sorting to packing to much more. However, in order to perform these tasks, robots need a way to recognize the orientation of objects in the world frame and in relation to each other.

Existing methods to address this issue exist, yet it is difficult to derive precise estimations from unprocessed data. For example, Model-based Sequential Manipulation [1], [2], [3], [4], [5] attempts this estimation but experiences limitations based on the ability of the state estimator due to it outputting explicit object states. SORNet (Spatial Object-Centric Representation Network) [6], a neural network backbone, was proposed to be a more powerful solution to this problem.

SORNet is based on a Vision Transformer Model [7], where it learns object-centric representations from RGB images. In addition to encoding an image, the model encodes example patches defining an object, referred to as canonical object views, such that the network can be trained on relationships between them. Specifically, after the encoding portion of the model, SORNet maintains a series of readout networks, which accept object embeddings to output predictions of their relationships. These readout networks are trained on logical predicates such as "top_is_clear", "left_of", and "behind", either defining an object's properties or relations to another object. By training on large sets of simulated data, SORNet was able to achieve state-of-the-art classification accuracy in these predicates.

One of SORNet's limitations is the lack of training in real-world data. SORNet provides accurate results in certain training datasets like Leonardo or CLEVR (Compositional Language and Elementary Visual Reasoning). However, due to a lack of background noise, it provides significantly more accurate outcomes than what may be seen in a real-world setting. A second limitation is that it requires a high supervision burden of labeling object relationships explicitly.

Our primary contribution was to train SORNet on PROPS Dataset (Progress Robot Object Perception Samples) and collect data on its accuracy and performance. Since PROPS images come from real-world settings, they offer a clearer picture of SORNet's potential performance in real-world scenarios. In doing so, we generated a framework that is easily applicable to other datasets with pose data. Lastly, the framework automatically generates relationships for objects in a scene given their pose data, reducing the burden of labeling.

II. Related Work

Model-based Sequential Manipulation: A majority of the work [1], [2], [3], [4], [5] utilizes a sequential, two-step pipeline approach. The initial phase derives explicit object states(e.g., bounding boxes or 6D poses), while the second phase plans actions to achieve some goal state, given the object states. These model-based systems are powerful at reasoning and apply to many different tasks with various goal conditions. Nevertheless, their effectiveness is contingent upon the proficiency of the state estimator, which is often lacking.

End-to-end Manipulation: Knowing explicit object states is not necessary for manipulation [8], [9], [10]. Motor controls can be learned directly from raw sensor inputs, such as RGB imagery and joint encoder data, thereby bypassing the stage of object state estimation. End-to-end methods leverage powerful neural network backbones that are able to extract low-level embedding vectors from high-dimensional images and directly optimize for downstream tasks. Although these techniques avoid the reliance on object models and explicit states, they do not have the notion of objects at all. As a result, they may lack sophisticated reasoning capabilities and could be confined to simpler scenarios involving only one or two objects and tasks in a relatively short time.

Learning Spatial Relations: In the field of 3D vision, methods such as [11], [12], [13] have been developed to predict both discrete and continuous pairwise object relations, given 3D inputs such as point clouds or voxels representations. These methods, however, typically operate under the assumption that the scene is fully observed and the objects are segmented with identifiable features. In contrast, SORNet's approach does not make any assumptions regarding the observability of the objects and does not require pre-processing of the sensor data.

Visual Reasoning: SORNet's architecture demonstrates the capability to address spatial-reasoning tasks for novel object instances without the need for separate segmentation or object detection components. SORNet focuses on a relatively complex manipulation task domain involving a manipulator in the observations

III. Algorithmic Extension

Our algorithmic extension focuses on enhancing SORNet's current capabilities to work with real data. To achieve this, we developed a class-based framework that enables the transformation of any dataset comprising scene images, identifiable objects within those scenes, the three-dimensional coordinates of each object's center, and the ability to generate canonical object views into a format compatible with SORNet's infrastructure. A base class is provided to handle the majority of the functions involving objects and predicates. It is intended to be extended for individual datasets to assimilate the actual scene images with the corresponding object information and locations.

To handle canonical views of real data, a selection of candidate images is automatically generated. A user will then manually select an arbitrary number of images to keep, indexing the "best" one as zero for validation purposes. A minimum of one image is needed as a canonical representation of the object, but having more canonical representations of the object will increase performance. During the evaluation phase, for each object depicted in the scene, an image is selected: a random image during the training phase, and the most suitable or "best" image when it comes time for evaluation. Examples of these manually selected views are depicted in Figure 1

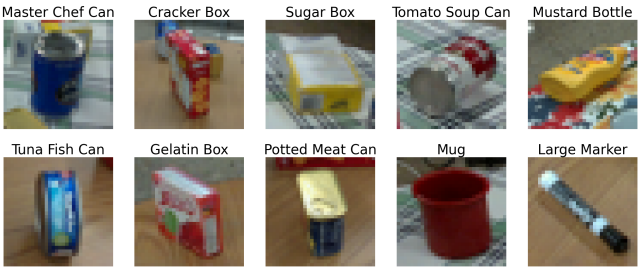

Fig. 1: PROPS Best Object Canonical Views

To obtain the relational data, images are parsed to obtain the three-dimensional coordinates of each object's center. These coordinates are used to infer information about the objects "left", "right", "back" and "front" relations relative to the camera's coordinate frame. Since datasets often augment their RGB images via a transform or some noise, these relations cannot be pre-computed and must be calculated at runtime. However, the operations are mostly binary and partially parallelizable, resulting in minimal performance decreases.

Additionally, our extension ensures consistent sizing of object views and images, enabling the extraction and resizing of image patches to meet the specific resolution input needs of SORnet. The extension seamlessly integrated with SORNet, enhancing its capacity to train and analyze across an extensive array of real-world data.

IV. Experiments and Results

To evaluate the dataset conversion framework and its ability to hold up to real-world data, SORNet was fully retrained for both the CLEVR dataset and PROPS dataset. Additionally, as an example of this paper's contribution, it was trained on the PROPS Pose dataset, with the resulting dataset named the PROPS Relation Dataset. An evaluation was performed across both models' final results and learning characteristics to address SORNet's ability to handle real-world data training.

A. Experimental Setup

To facilitate training on the PROPS dataset, the aforementioned dataset class was instantiated with logical predicates calculated and saved per image, and formatted to retain object naming conventions for readability. The class, along with PROPS example usage and sample-generated data, are made public at [14], and example usage with SORNet is shown at [15]. Although the framework is easily overloaded to other datasets with object center data, it was only evaluated on the PROPS dataset for brevity. To demonstrate its utility, predicates for whether objects were left, right, behind, or in front of other objects were all automatically generated through pose data.

To provide a fair comparison, the model for each dataset was trained on near-identical hyper-parameters, with exceptions only relating to dataset size and machine multi-threading capabilities. Since the PROPS dataset has images at a higher resolution (640 by 480 pixels compared to 480 by 320 pixels), a model was trained at each resolution. However, there were significantly fewer images in the PROPS dataset, so its model was trained for 80 epochs as opposed to CLEVR's 40. Additionally, images were sampled from the training set in identical batch sizes - 32 images at a time. After Each epoch, the models were evaluated against their respective validation sets, recording the accuracies depicted in Figures 2 and 3. Lastly, attempts were made to fine-tune the resulting model from CLEVR onto the PROPS dataset, however, the initial accuracy was on par with random guessing, resulting in only complete training being used.

B. Results

Overall, SORNet demonstrated capabilities to learn real-world data, but could not share an inference with a separate

set of simulated objects. As demonstrated by Figures 2-6 and Table I, each model individually achieved over 99% training and validation accuracy once converged. Due to the sharp contrast in objects, the model sets were incompatible with each other, scoring around 50% in each logical predicate, or effectively behaving randomly. Relative to the number of batches until convergence, it can be noted that the full-size PROPS dataset took roughly 200 batches (12-13 epochs) before any significant accuracy improvements, while the CLEVR set had a noticeably less drastic curve, improving within a single epoch (2187 batches). Of note as well, the downscaled version took twice as long to achieve significant improvements, demonstrating the level at which the model can learn the fine details in the objects.

The behavior differences in CLEVR and PROPS could in part be due to PROPS's smaller dataset, but is likely also due to the inherently larger amount of noise associated with real-world datasets. In particular, even the canonical object views sometimes had snippets of other objects in them, making learning an accurate embedding vector per object harder for the network. The network also converged in many fewer iterations, by nearly an order of magnitude, demonstrating that SORNet can scale to smaller object datasets relatively easily, even when there are fewer training images. Regardless, given enough time, the network prevailed, generating high-quality results.

Additionally, concerning the objects in the PROPS dataset itself, SORNet performed comparably across the board. Analyzing Table I, complete accuracy averages for relations involving the sugar box were the best at 99.36%, followed by the gelatin box, with the large marker the worst at 98.98%, barely beating out the master chef can at 98.99%. When analyzing their respective best canonical views in Figure 1, the master chef can's low accuracy could potentially be attributed to the presence of multiple objects at the same size of the can, as opposed to the smaller objects in the cracker box canonical view, or the other remaining views that are nearly free from other objects. The large marker is also quite small relative to the other objects, potentially introducing some errors as well. On the other side, the gelatin and sugar boxes have a very well-defined canonical object view, driving the importance of the quality of canonical object views.

V. Conclusions

To summarize, SORnet (Spatial Object-Centric Representation Network) demonstrated performance in learning object-centric representations from RGB images and has shown strong performance on datasets such as Leonardo, Kitchen, and CLEVR. However, the real-world application of SORNet was uncertain due to its initial training on simulated datasets, which lack details available in natural environments. Our work addresses this gap by training SORNet on a more challenging and noisy dataset, the PROPS (Progress Robot Object Perception Samples) dataset. SORNet demonstrated the ability to learn real-world data, but could not share an inference with a separate set of simulated objects. We demonstrate that SORnet

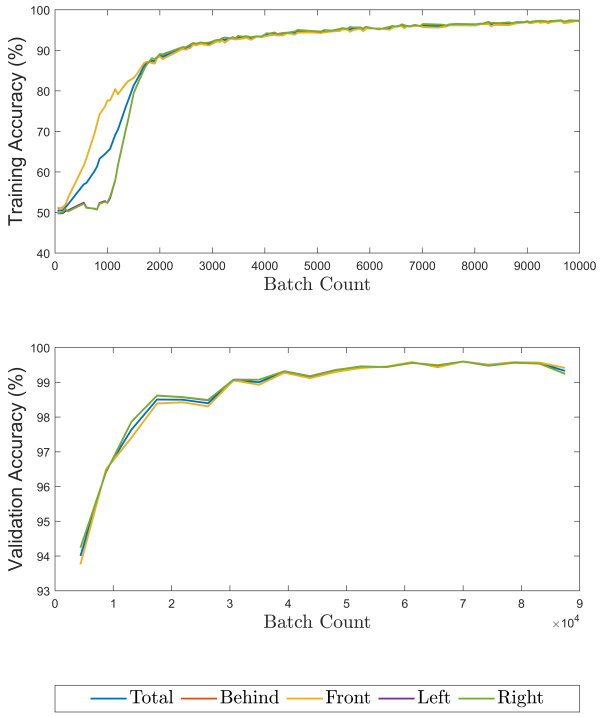

Fig. 2: CLEVR Dataset Results. Only the first 10,000 batches are shown for training to highlight early trends

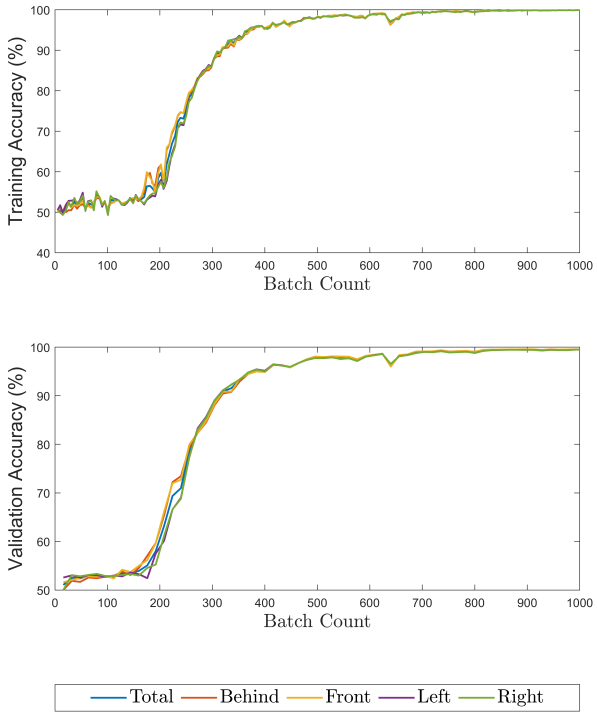

Fig. 3: PROPS Full Resolution Dataset Results

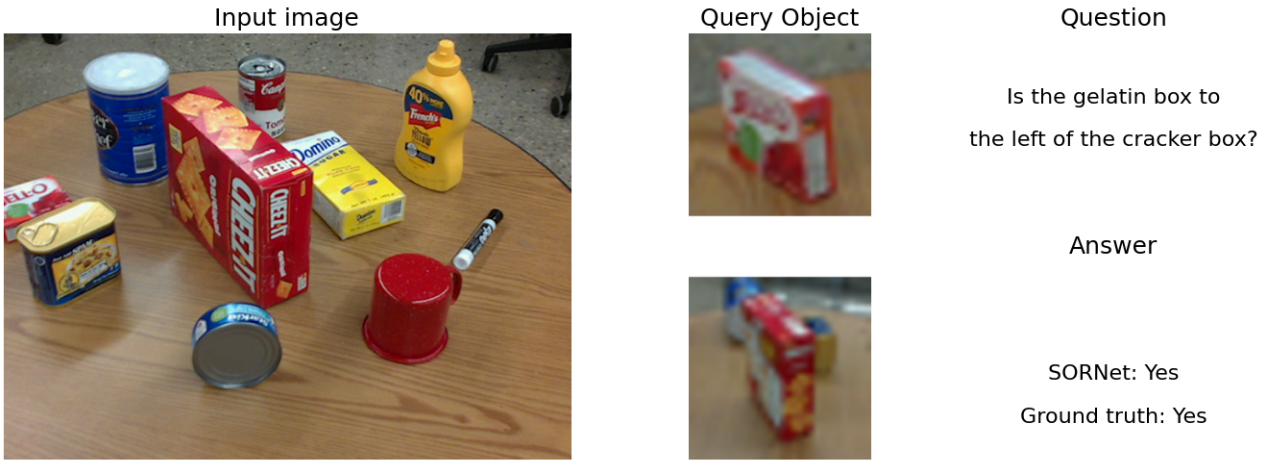

Fig. 4: PROPS Box Item Query

Fig. 5: PROPS Canned Item Query

	Master Chef Can	Cracker Box	Sugar Box	Tomato Soup Can	Mustard Bottle	Tuna Fish Can	Gelatin Box	Potted Meat Can	Mug	Large Marker	Average
Master Chef Can	-	99.30	99.77	98.80	98.90	98.77	98.65	99.20	99.15	98.85	99.04
Cracker Box	99.10	-	99.37	99.80	99.20	99.39	98.54	98.70	99.55	98.3000	99.11
Sugar Box	99.20	99.14	-	99.09	99.37	98.89	99.75	99.32	99.54	99.20	99.28
Tomato Soup Can	98.40	99.65	99.26	-	99.40	98.87	98.86	99.60	99.00	99.15	99.13
Mustard Bottle	99.30	98.90	99.26	99.90	-	98.87	99.68	98.95	98.55	98.95	99.15
Tuna Fish Can	98.98	99.28	99.41	98.98	97.95	-	99.11	99.13	98.98	99.28	99.01
Gelatin Box	99.19	99.40	99.88	99.51	99.89	99.33	-	98.81	99.78	99.03	99.43
Potted Meat Can	99.20	98.70	99.03	99.75	98.30	99.38	98.81	-	98.90	98.20	98.92
Mug	98.80	99.45	99.49	98.80	98.70	98.92	99.51	99.65	-	99.45	99.20
Large Marker	98.30	98.10	99.43	99.20	98.95	99.23	99.24	99.20	99.55	-	99.03
Average	98.94	99.10	99.43	99.31	98.96	99.08	99.13	99.17	99.22	98.93	99.13
Complete Average	98.99	99.10	99.36	99.22	99.06	99.04	99.28	99.05	99.21	98.98	

TABLE I: Full Size PROPS Data Validation Accuracy Percentages for all Relationships. The row is object 1 in the relationship, the column is object 2 in the relationship. The complete average is the average over the object's row and column, as SORNet treats the first and second patches differently.

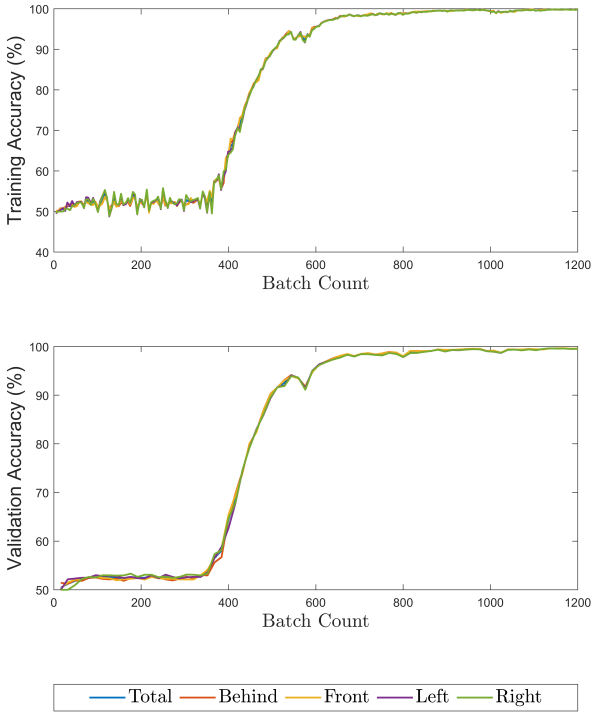

Fig. 6: PROPS Downsized Resolution Dataset Results

was able to output highly accurate predictions given enough training time, across all objects within the dataset. One of the limitations of SORNet we identified, but did not address, includes its inability to handle conflicting answers to a query (multiple of the same object in different locations). Regardless, the PROPS Relation Dataset and our modified SORNet code contribute to the community, offering opportunities for further research and refinement of SORNet's capabilities. Future work could focus on addressing the limitations related to dataset compatibility, applying our framework to other datasets, or exploring the effects of incremental learning or transfer learning techniques to bridge the gap between simulated and real-world data. The progress made on SORNet with the PROPS Dataset is a promising step towards more complex robotic perception systems that can interact with a dynamically changing real world.

References

[1] R. E. Fikes and N. J. Nilsson, "Strips: A new approach to the application of theorem proving to problem solving," *Artificial Intelligence*, vol. 2, no. 3, pp. 189–208, 1971. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370271900105

[2] M. Fox and D. Long, "PDDL2.1: an extension to PDDL for expressing temporal planning domains," *CoRR*, vol. abs/1106.4561, 2011. [Online]. Available: http://arxiv.org/abs/1106.4561

[3] T. Ribeaud and C. Z. Sprenger, "Behavior trees based flexible task planner built on ros2 framework." IEEE Press, 2022. [Online]. Available: https://doi.org/10.1109/ETFA52439.2022.9921683

[4] C. Paxton, N. D. Ratliff, C. Eppner, and D. Fox, "Representing robot task plans as robust logical-dynamical systems," *CoRR*, vol. abs/1908.01896, 2019. [Online]. Available: http://arxiv.org/abs/1908.01896

[5] Z. Sui, L. Xiang, O. C. Jenkins, and K. Desingh, "Goal-directed robot manipulation through axiomatic scene estimation," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 86–104, 2017. [Online]. Available: https://doi.org/10.1177/0278364916683444

[6] W. Yuan, C. Paxton, K. Desingh, and D. Fox, "Sornet: Spatial object-centric representations for sequential manipulation," in *5th Annual Conference on Robot Learning*. PMLR, 2021, pp. 148–157.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[8] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," 2016.

[9] A. Ghadirzadeh, A. Maki, D. Kragic, and M. Björkman, "Deep predictive policy training using reinforcement learning," 2017.

[10] Y. Duan, M. Andrychowicz, B. C. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," 2017.

[11] S. Fichtl, A. McManus, W. Mustafa, D. Kraft, N. Krüger, and F. Guerin, "Learning spatial relationships from 3d vision using histograms," in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation*. United States: IEEE, 2014, pp. 501–508, 2014 IEEE International Conference on Robotics and Automation, ICRA 2014 ; Conference date: 31-05-2014 Through 05-06-2014. [Online]. Available: https://ewh.ieee.org/soc/ras/conf/fullysponsored/icra/2014/www6.cityu.edu.hk/icra2014/index.html

[12] B. Rosman and S. Ramamoorthy, "Learning spatial relationships between objects," *The International Journal of Robotics Research*, vol. 30, pp. 1328 – 1342, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:17126420

[13] M. Sharma and O. Kroemer, "Relational learning for skill preconditions," 2020.

[14] A. V.-A. Jace Aldrich and H. Ho, "Props relation dataset," https://github.com/Jaldrich2426/PROPS-Relation-Dataset, 2024.

[15] ——, "Sornet," https://github.com/Jaldrich2426/sornet, 2024.