Augmenting Simulation Data with Sensor Effects for Improved Domain Transfer

Adam J. Berlier[1 [0000−0001−9657−5148]], Anjali Bhatt[1 [0000−0003−0740−0560]], and Cynthia Matuszek[1 [0000−0003−1383−8120]]

University of Maryland Baltimore County, 1000 Hilltop Rd, Baltimore, MD 21250, USA

Abstract. Simulation provides vast benefits for the field of robotics and Human-Robot Interaction (HRI). This study investigates how sensor effects seen in the real domain can be modeled in simulation and what role they play in effective Sim2Real domain transfer for learned perception models. The study considers introducing naive noise approaches such as additive Gaussian and salt and pepper noise as well as data-driven sensor effects models into simulation for representing Microsoft Kinect sensor capabilities and phenomena seen on real world systems. This study quantifies the benefit of multiple approaches to modeling sensor effects in simulation for Sim2Real domain transfer by their object classification improvements in the real domain. User studies are conducted to address hypotheses by training grounded language models in each of the sensor effects modeling cases and evaluated on the robot's interaction capabilities in the real domain. In addition to grounded language performance metrics, user study evaluation includes surveys on the human participant's assessment of the robot's capabilities in the real domain. Results from this pilot study show benefits to modeling sensor noise in simulation for Sim2Real domain transfer. This study also begins to explore the effects that such models have on human-robot interactions.

Keywords: Sim2Real, robotics, virtual reality, human-robot interaction

1 Introduction

The field of robotics continues to benefit from advances in machine learning to better understand the world they operate in. These data-driven approaches are showing promising results for developing effective robot control policies as well as machine perception. However, machine learning approaches require large amounts of data to train generalized models that represent their operational environment. The best-performing machine learning models benefit from massive databases of well-curated data. These resources are typically not widely available and protected by corporate intellectual property or licensing agreements. Representative datasets in the robotics domain are especially difficult to come

by. Due to demanding cost, time, operator expertise, and the number of moving parts required to conduct a robotics experiment, most datasets are sparse, limiting machine learning approaches. This becomes increasingly difficult for the field of Human-Robot Interaction (HRI).

Introducing humans into experiments requires additional overhead time and can be difficult to find schedule and coordinate. To address these challenges, many researchers use simulation to train their robots. In simulation, experiments are less expensive, require less setup and tear-down, can be run faster than real time, do not experience mechanical failures, and can be bound to isolated experiments for improved reproducibility. Simulation experiments are more accessible and carry less risk of robot failure which may lead to the destruction of property and persons. It provides significant benefits for the field of robotics and Human-Robot Interaction (HRI). With the ultimate goal of learning perception models from the simulation that can successfully be deployed on real robots, the Sim2Real domain transfer problem takes advantage of these benefits. The simulation also provides a more efficient process for data collection. Experiments can be run in parallel, faster than real-time, and require less effort to assemble and disassemble. Some human populations are less at risk of interacting with research robots in simulation compared to the real world. This became especially true during the COVID-19 pandemic stay-at-home orders that prevented subjects from physically entering research labs.

Even when working in simulation, robotics researchers are pursuing the end goal of enabling robots to successfully perform tasks in the real world. However, there are significant drawbacks to simulation as well, and directly applying trained models in simulation to embodied robots performing in the real world is showing less than desired results. The Sim2Real domain transfer problem is driven by simulation missing detailed complexities of the real world. This challenge of domain adaptation has led many researchers to leverage advances in computer graphics to make simulated scenes more photo-realistic. Researchers have also improved domain transfer by simulating realistic sensor effects experienced in the physical world. Some approaches investigate signal noise and well-characterized physics models of sensor effects, while others resort to data-driven approaches that collect large amounts of measurements from real-world sensors and generate learned sensor models for mapping clean simulation data to data that represents learned characteristics in real measurements.

The simulation environment being used for our work is Robot Interaction in Virtual Reality (RIVR) with a Clearpath Robotics Husky UGV using a Microsoft Kinect as the primary sensor [5]. Each method is applied in simulation and compared using them as training data for an object classifier, along with two baseline cases. The first baseline case is trained in the real domain as an expected upper bound best case and the second is trained in simulation without sensor effects being modeled as an expected lower bound worst case. The expectation is that by modeling these sensor effects in simulation, object classification performance in

the real domain will improve compared to the lower bound. This paper presents a comparison of evaluation metrics across three differing approaches for modeling sensor effects for improved Sim2Real domain transfer. The three approaches considered are 1) statistical signal noise approaches, 2) physics-informed sensor effects models specific to each modality, and 3) data-driven models for adding sensor effects that are learned from real-world measurements. Evaluation metrics are compared for performance on a grounded-language task using the Grounded Language Dataset (GoLD) [6]. GoLD contains measurements of common household objects using a Kinect sensor that measures three optical modalities: RGB image, depth image, and raw point clouds. GoLD also includes object-associated descriptions in multiple formats: text, speech (audio), and speech transcriptions. The Kinect sensor is used to generate this data and measurements are collected in a laboratory setting. Since GOLD contains only real-world measurements of objects and their associated language descriptions, this work will also collect simulated sensor data for each implemented sensor effects method applied to a virtual Kinect sensor and conduct human trials for language descriptions of simulated objects.

2 Related Works

Since the Microsoft Kinect is a popular commercial-off-the-shelf sensor used by robotics researchers, many investigations have been conducted to understand real-world sensor effects experienced by this specific device. Khoshelham et al. investigate Kinect sensor quality, discussing the calibration process and analyzing sensor resolution and quality [7]. They construct a mathematical model of depth measurements and present a theoretical error analysis to provide insight into how sensor accuracy is influenced by the effects it is exposed to. Since internal operations of the sensor are protected by the manufacturer, these studies provide deeper insight into sensor design. Researchers have also completed an investigation toward understanding the Kinect's internal workings by conducting error analysis and building a representative model from those observations. A few major works also consider physical phenomena in sensor modeling in a different environment. Farrell et al. develop MATLAB sensor models that specify sensor properties and successfully predict sensor performance in natural scenes with high dynamic range or low light levels, capturing spectral sensitivity and electrical properties including dark current, read noise, dark signal non-uniformity, and photoreceptor non-uniformity estimated from a set of calibration measurements [4]. However, Konnik et al. investigate training data augmentation for CCD and CMOS sensors by closely modeling physical phenomena experienced by these sensors including voltage-to-voltage, voltage-to-electrons, and analog-to-digital converter non-linearities [8].

Our variant of the Husky robot uses the Kinect as its primary sensor. Research to understand the Kinect specifically greatly benefits the development of the RIVR simulation models of this sensor package [5]. Few studies highlight interesting insights for building geometric models of sensor performance.

Smisek et al. investigate Kinect performance errors, diving deeper into Time Of Flight (3D-TOF) [13]. They studied sensor performance, highlighting what sort of degradation can be applied for a more realistic representation in simulation. Clouet et al. propose a novel geometric representation of sensor noise propagation from raw acquisition through spectral reconstruction and color correction [3]. They focus on characterizing the existing noise of RGB and multispectral images to effectively mitigate sensor noise and develop more accurate sensor noise models. Nguyen et al. build on the previously quantified Kinect axial noise distributions as a function of distance to the observed surface [11]. They expand this work by quantifying a novel Kinect noise model having both axial and lateral components. The new model is a statistical, data-driven approach that derives parametric models as a function of both distance and angle to the observed surface.

Learning from effects seen specifically in our operating environment will likely lead to improved Sim2Real domain transfer. Many previous studies conducted have shown value in Sim2Real domain transfer, as well as tracking, filtering, and estimation. Sweeney et al. investigate a data-driven approach leveraging a convolutional neural network (CNN) to predict which pixels of a simulated noise-free depth image will be no-depth-return pixels [14]. They use noise-free simulated depth images and noisy real-world depth image pairs as labeled examples to train the network for adding no-depth-return pixels to simulated images. They focus on no-depth-return pixels because they believe that this is the most disruptive sensor effect experienced in the depth modality. Some have investigated data-driven approaches to generate realistic simulated sensor models. Other approaches considered implementing both commonly assumed image noise filters and high-fidelity physics models of sensor phenomena. Liu et al. develop a method for inferring an image noise level function from an image [10]. Image noise is not simply additive, but a function of lighting conditions, object colors, and object textures. Their noise level function is built from piece-wise smooth priors, likelihood models, and Bayesian MAP inference. This work will help inform possible methods for more realistic simulated images calibrated with a real image as a low overhead one-shot method.

Of the sensor modalities used by our Husky robot, depth is particularly subject to effects disrupting information provided by the sensor. This makes work focusing on depth especially important to our work. Landau et al. focus on the contribution of IR sensor noise in estimating depth [9]. They propose a high-fidelity Kinect IR and depth image predictor and their work develops a simulator that models the physics of the transmitter/receiver system, unique IR dot pattern, disparity/depth processing technology, random intensity speckle, and IR noise in the detectors. Carlson et al. propose an automated data augmentation pipeline to vary chromatic aberration, blur, exposure, noise, and color temperature sensor effects for RGB images [1]. They showed that training on synthetic data generalizes to improved performance in the real domain for object detection and segmentation tasks [2]. We will be leveraging each of these

advances to benefit Sim2Real domain transfer as we implement our variations of these methods into the RIVR simulation and use prior work as a baseline for our results [5].

3 Approach

To investigate each of the three approaches and compare their performance for Sim2Real domain transfer, a simulated dataset with associated sensor effects, a real-world dataset, and a learning task are developed. The Human-Robot Interaction in Virtual Reality (RIVR) Simulation will be used to simulate realistic environments that the robot can learn in [5]. A sensor effects ROS package was developed to manage multiple implementations of sensor effects models in simulation. An extension of GoLD will be used as the real-world dataset [6]. The learning task will be a grounded language task in which robots learn to associate human language descriptions of objects in the simulated scene with perceived measurements from the Kinect sensor. Lastly, performance metrics, such as Mean Reciprocal Rank (MRR) and the accuracy of each model trained using the three different sensor effects methods in the simulation are evaluated on the real-world dataset to characterize how much each approach improves Sim2Real domain transfer.

3.1 Simulator

The RIVR simulation is a 3-D environment built in the Unity game development engine that is compatible with virtual reality headsets. Intending to provide robots with a diverse set of virtual environments in which they can interact, RIVR currently provides three different scenes: an apartment, a hospital room, and a maker space. The Unity simulation is built with the ROS URDF plugin for adding models of physical robots to the scene. This simulation is used to simulate realistic environments in which the robot and a human participant using

Fig. 1: RIVR Simulation Apartment Scene

a VR headset can interact. ROS bag files are used to collect all sensor data from an HRI training session, as described in the following user studies section. An example of the apartment scene is shown in Fig. 1.

3.2 Sensor Effects Modeling

A ROS sensor effects node is built into the simulation system that reads the raw simulation sensor measurements, absent of realistic sensor effects, and filters them through the sensor effects node as shown in Fig. 2. The output is published on a new topic that introduces realistic sensor effects into the measurements. The robot subscribes to these new topics for realistic perception in simulation. This approach reduces the effort required to introduce new sensor effects to the real-time raw measurements during an experiment or apply them to recorded data post-experiment.

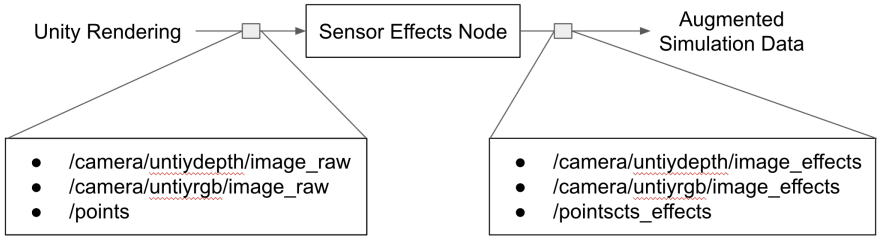
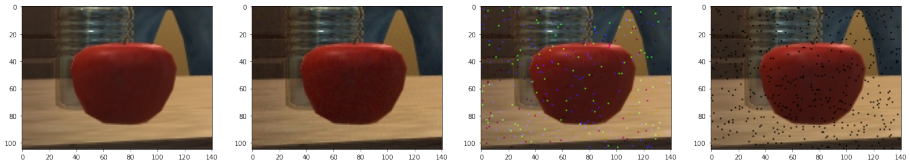


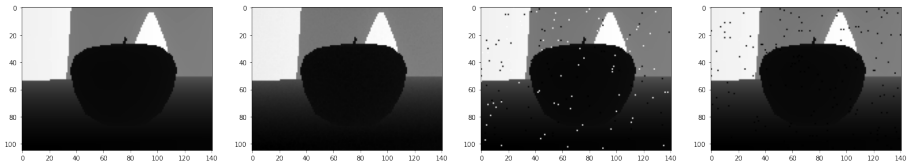
Fig. 2: Sensor Effects Node Dataflow

Naive Noise Models We investigate three primary naive noise model approaches. The first is additive Gaussian noise. This approach adds Gaussian noise to the image-based user-defined mean and variance parameters. Further work should consider a formal approach to selecting the best mean and variance in this approach. For this experiment, we chose a zero mean and variance of 10. The second naive noise approach is random salt and pepper noise. This approach randomly adds white and black pixels to the image for a user-defined percentage of pixels and a ratio of white and black pixels, salt-to-pepper ratio. Further work should consider a formal approach to selecting pixel percentage and salt to pepper ratio in this approach. For this experiment, we chose 10 percent of pixels and a uniform salt-to-pepper ratio of 0.5. The last naive noise approach is random pixel dropout. This approach is the same as the salt and pepper noise approach with a salt vs pepper ratio of zero adding only pepper noise to the image. A comparison of each approach can be seen in Fig. 3 and Fig. 4 applied to color and depth respectively.



(a) Original (b) Additive Gaussian Noise (c) Salt and Pepper Noise (d) Dropout Noise

Fig. 3: Naive Noise on Color Images



(a) Original (b) Additive Gaussian Noise (c) Salt and Pepper Noise (d) Dropout Noise

Fig. 4: Naive Noise on Depth Images

Data Driven Models Data-driven models for sensor effects are a way to use the real dataset for training a model to accurately replicate the desired real images in simulation. This paper compares two data-driven approaches presented in recent publications. The first approach addresses accurately modeling monochrome depth images. In this data-driven approach Sweeney et al. leverage a convolutional neural network (CNN) to predict which pixels of a simulated noise-free depth image will be no-depth-return pixels [14]. They use noise-free simulated depth images and noisy real-world depth image pairs as labeled examples to train the network for adding no-depth-return pixels to the simulated images. They focus on no-depth-return pixels because they believe that this is the most disruptive sensor effect experienced in the depth modality. The second approach addresses accurately modeling sensor effects in color images. Carlson et al. address this problem from a desire to develop more realistic data augmentation for training autonomous cars in video games that are capable of transferring what they learned to the real world [1]. They propose an automated data augmentation pipeline to vary chromatic aberration, blur, exposure, noise, and color temperature sensor effects for RGB images. Such sensor effects are well-known characteristics of operational sensors. This work learns to tune these features for an accurate representation of the target sensor operating in the real world. Carlson et al. also develop a process that uses a generative augmentation network to learn a transfer function for sensor effects observed in real domain images transferred to synthetic domain images. They show that training on synthetic data generalizes to improved performance in the real domain for object detection and segmentation tasks [2]. Each of these approaches is implemented for comparison.

Trained models provided by the authors are fine-tuned on GoLD data with a portion of data held out for testing.

3.3 Dataset

A training dataset is constructed from simulated measurements of Unity assets that represent household objects in GoLD. Each of the previously described sensor effects modeling approaches is also added to the simulation dataset. An extended version of GoLD is also used as a training dataset. This dataset consists of five perception modalities for grounded language learning on a variety of household objects: RGB image, depth image, typed text descriptions, spoken descriptions, and Google ASR. Since this dataset was collected with a different sensor than is modeled in RIVR, GoLD is extended by adding RGB and depth images for a subset of the same household objects as captured on the Husky Kinect sensor in the real world. A subset of the real dataset will be held out for testing learned models.

3.4 Grounded Language Learning

One element of learning physically embodied, or *grounded*, language is associating natural language words with perceptual inputs, such as imagery. A grounded language learning task is used to evaluate transfer learning performance afforded by each sensor effects implementation. Each of the visual perception modalities in the training set: real, raw simulation, and simulation with sensor effects models, are used in conjunction with language descriptions of household objects to train a grounded language model. This grounded language model is then deployed on the Husky robot in the real world. Users interact with the robot by asking it to hand them objects that they see on the table. The models' performance will be evaluated during this interaction using Mean Reciprocal Rank (MRR) and accuracy. This process is visualized in Fig. 5. RIVR takes the raw Unity rendering for both color and depth and adds the desired sensor effects models with the sensor effects ROS node, which publishes simulation data augmented with sensor effects. This data is then used to train the grounded language model by Richards et al. [12]. GoLD is used for testing the learned model during the training process. The learned model is then deployed on the Husky robot for a live HRI study where the user speaks to the robot, that speech is transcribed to text, and the three modalities of speech transcription, RGB image, and depth image are all provided to the grounded language model. The robot will finally display a cropped image of the object that it has selected to be the most aligned with the spoken language. The separation in performance across sensor effects implementations provides insight into understanding the extent to which certain sensor effects modeling improves Sim2Real domain transfer.

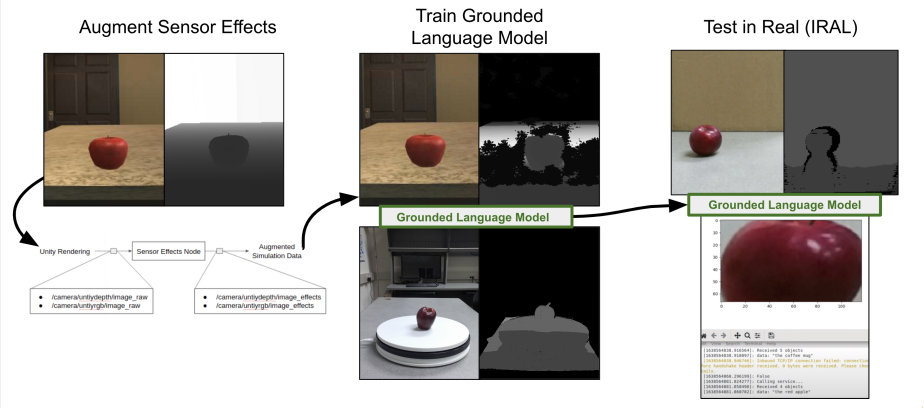


Fig. 5: Grounded Language Learning with Sensor Effects

3.5 User Studies

The participants for this user study were 15 computer science and engineering students. Each of the participants was asked to sign a self-declaration form for any COVID-19 symptoms, including their travel during the last 14 days, or if they were in contact with a COVID-19 positive patient. The participants also directly interacted with the Husky Robot, which is equipped with a Kinect camera and one Kinova Jaco arm. The robot was operating with two different models, one trained on raw simulation data and another model trained with sensor effects. Before starting the experiment, each participant was asked to fill out a pre-questionnaire.

Two models were deployed: a model with raw simulated data categorized as Model A, and a model with sensor effects as Model B. The order in which the models were introduced to participants was different for each participant, and instructors reminded each participant to note the sequence of models they encountered. The participants were instructed with two sets of similar tasks, to be performed with 5 objects from GoLD placed on a table, while interacting with both models. The task was to ask the robot to pick up an item of their choice from the table and wait for 10 seconds for the robot to respond. The robot was expected to choose the object from the table that most aligned with the user's spoken request.

Each experiment took roughly 30 minutes, including consent forms, instructions, and post-task questionnaires, as well as interacting with the robot as it was deployed with each grounded language model. The order of models was randomized in order to prevent any bias in the perceived performance of the robot between the grounded language models used.

This user study addresses three primary hypotheses:

Hypothesis 1 Sensor effects in simulation during grounded language learning will improve the robot's understanding of real-time human requests when evaluated in the physical world.

Hypothesis 2 Participants will subjectively rate their perception of the robot's performance to be better when using the grounded language model that was trained in simulation using sensor effects modeling compared to the grounded language model trained in simulation without using sensor effects modeling.

Hypothesis 3 Participants will prefer to work with a robot using the grounded language model that was trained in simulation using sensor effects modeling compared to the grounded language model trained in simulation without using sensor effects modeling.

At the beginning of the experiment, participants were given detailed instructions about what they will experience during the study. Each of the participants was provided a short description of the task with a sample interaction of a moderator instructing the robot, followed by the robot's response being demonstrated. They were instructed to describe each item once for each model. The experiment started with the robot asking the participant which item they would like to pick up, and then participants describe an item they see on the table through spoken language. The participants were asked to wait for the robot to complete the required task for at least 10 seconds. After the robot has detected and selected an item, the complete process is reiterated after 15 seconds.

Example: Robot: "Which item would you like me to pick?" Participant: "Where is the apple?" The participant will wait while the robot processes and then moves its arm to indicate the object described. The robot will continue to ask 5 more such commands.

After 5 such commands, the experimenter asks: "That's five instructions. Would you be willing to do five more? Any answer is fine." Participant: "I think I'll stop now."

The participants continued with the experiment and were given a short break of 1-2 minutes before continuing the experiment by interacting with the second grounded language model. All participants were asked to describe the items in their own words as before.

Post-experiment, every participant was asked to submit a survey based on their encounter with the robot. The survey consists of 5-point Likert scale agree/disagree questions comparing the results of the two models they interacted with. They were also asked about suggested feedback for the robot, and

if they were able to detect any dissimilarity between the two models they interacted with. The experimenter tracked what the participant asked for, how often the robot crashed for each user, and noted the robot's predicted item. The instructor remained in the room and made sure not to engage with the experiment except to restart the robot in case of crashes, tracking of correct/incorrect indications made by the robot, and keeping notes if the user asks to stop after 5 trials. Based on the participant's survey and the robot's performance, the participant's perception of the robot's performance, level of comfort, and level of willingness to use the robot are evaluated. All measurements were collected both using the grounded language model that was trained in simulation using sensor effects modeling and the grounded language model trained in simulation without using sensor effects modeling.

4 Results

Two sets of results are presented. First, results are presented on the performance of the language grounding model itself. This compares a variety of sensor effects modeling approaches and their performance on both GoLD and during the live user studies. Second, results are presented on the data collected from the users in both the pre-study and post-study questionnaires. The user study results analysis will also cover some observations collected by the experimenters.

4.1 Model Performance

Results for the model performance are presented in two ways. First, accuracy and mean reciprocal rate (MRR) are presented for each combination of sensor effects models in Table 1. The MRR metric measures how high the correct response is in a ranked list of alternatives. The average of the reciprocal rank $\frac{1}{n_i}$ across all instances is evaluated, where the reciprocal rank is the inverse of the rank at which the specific value was found.

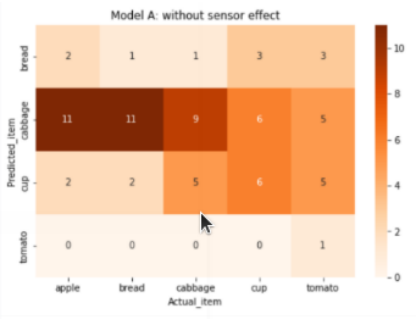
$$MRR = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{n_i} \tag{1}$$

The hypothesis is that by adding realistic sensor effects models, the grounded language model would learn a better model for associating the transcripts of spoken language to the objects perceived in the scene. This hypothesis is supported. In this study, it was shown that using simple Gaussian noise for color and NDP for depth provides the best performance on GoLD. Given this information, the same sensor effects model combination was used for the user studies.

Sensor Effects Model		**Test Data**	**Evaluation Metrics**	
Color	***Depth***		***Accuracy***	***MRR***
None	None	GoLD	0.1997	0.4551
None	Dropout	GoLD	0.1976	0.4525
None	SNP	GoLD	0.2050	0.4602
None	Gaussian	GoLD	0.1988	0.4547
None	NDP	GoLD	0.1953	0.4504
Gaussian	None	GoLD	0.1945	0.4531
Gaussian	Dropout	GoLD	0.2062	0.4639
Gaussian	SNP	GoLD	0.2050	0.4605
Gaussian	Gaussian	GoLD	0.2085	0.4635
Gaussian	NDP	GoLD	**0.2134**	**0.4665**
All	All	GoLD	0.2092	0.4658

Table 1: Sensor Effects Models Performance on GoLD

In the user study, participants interact with both models: Model A, which is trained on simulation data without sensor effects modeled, and Model B, which is trained on the best previously performing combination of sensor effects models. Measurements were collected from these interactions and evaluated for quantitative performance on real-world data. Fig. 6 presents the performance metrics for model A. Overall accuracy is 0.2192 and the macro average f1-score is 0.1378. This performance is better than random given that there are 5 classes to choose from. Model A does seem to be slightly biased toward predicting "cabbage" and "cup." Fig. 7 presents the performance metrics for model B. Overall accuracy is 0.2761 and the macro average F1-score is 0.2576. Model B seems to have a slight bias toward "tomato." The main takeaway is that Model B, which incorporates sensor effects models, performs nearly twice as well in the real-world user studies as Model A which does not have sensor effects modeled in the simulation training data. As for the biases, it is hard to tell with such a small sample of user study test cases how biased these models really are. A larger study should be conducted to ensure these trends are maintained.



Classification report for Model A

	precision	recall	f1-score	support
apple	0.000000	0.000000	0.000000	15.000000
bread	0.100000	0.071429	0.083333	14.000000
cabbage	0.214286	0.600000	0.315789	15.000000
cup	0.250000	0.357143	0.294118	15.000000
tomato	1.000000	0.071429	0.133333	14.000000
accuracy	0.219178	0.219178	0.219178	0.219178
macro avg	0.260714	0.183333	0.137762	73.000000
weighted avg	0.302935	0.219178	0.162847	73.000000

Fig. 6: Confusion Matrix for Model A (No Sensor Effects)

Classification report for Model B

	precision	recall	f1-score	support
apple	0.375000	0.200000	0.260870	15.00000
bread	0.222222	0.142857	0.173913	14.00000
cabbage	0.272727	0.200000	0.230769	15.00000
cup	0.190476	0.266667	0.222222	15.00000
tomato	0.320000	0.533333	0.400000	15.00000
accuracy	0.270270	0.270270	0.270270	0.27027
macro avg	0.276085	0.268571	0.257555	74.00000
weighted avg	0.276813	0.270270	0.258685	74.00000

Fig. 7: Confusion Matrix for Model B (Sensor Effects using Gaussian for Color and NDP for Depth)

4.2 User Study

The experiment started with a set of a pre-task questionnaires in which participants were shown a video of the RIVR simulation; they responded positively that the RIVR simulation seems realistic. This trend held in the post-task questionnaire, but was less significant since a couple of participants changed their opinions after working with the real robot. Post-task questionnaire results showed that participants preferred working with Model B and found it to be more successful model.

One observation from the pre-task questionnaire form was regarding participants' experience with robots and machine learning. During experiments, the robot sometimes crashed due to audio failure or the arm took longer to operate than expected. Interestingly, the participants with minimal background in robotics and machine learning were frustrated by the response time of the robot and its failure modes. The other group with sufficient background in these fields patiently waited for the robot to respond even after a few failed attempts.

Bias in the model is another intriguing observation of this user study. The instructor made no comments on the user's response when the robot did not

detect the correct object. However, some participants observed this behavior from the robot and commented on model A being biased towards "cabbage" and model B being biased towards "tomato," which was later verified with the confusion matrix. Note that the item "apple" was never recognized by model A, while model B gave a more accurate prediction.

5 Conclusion

This work serves as a promising pilot study with intriguing results. This pilot study looked at results from a small group of participants that were a population of convenience. Both quantitative evaluation metrics and qualitative observations suggest all three hypotheses were valid. Additional studies that improve available data and participant's data collection will provide more statistically relevant results. During user studies on the live robot, challenges were presented with data rate processing that stressed the robot's computational capabilities. More work can be done on the robot to make it more effective and robust during user studies. This will help get more participants through the study. Overall, results suggest improvements and the need for a larger study to provide more statistically significant results.

6 Future Work

Future work will further develop this method and address some concerns with the approach presented in this paper. The study will also expand the number of trials to provide more statistically significant results. The current approach models the Kinect 2 sensor in simulation and runs the Kinect 2 on the live robot for the user study. However, GoLD, which was used for the testing portion of the grounded language model, was collected with the updated Azure Kinect sensor. Future work will be updating the robot to be equipped with the Azure Kinect and the simulation will also model the Azure Kinect and its related sensor noise. This is expected to have an impact on results. In future work we will also be padding the cropped image in the real user studies with white pixels to maintain proper aspect ratios. Currently, all images are being resized to the same aspect ratio, which can warp information in the image and should be avoided in future work. A larger domain randomization study is also of interest. Combining all approaches to model sensor noise in the training data will provide more instances to train over and may increase robustness in a variety of real-world scenarios. Bias was also shown to be a potential issue in this experiment. There has been much research on this topic that could be leveraged to negate any issues of bias from impacting this experiment. A larger study could be an opportunity to better understand what biases are present and how to negate them. Lastly, an approach to one-shot calibration of the various method using a single datum collected from the initial scene the robot is presented with may help improve autonomy and performance.

References

1. Carlson, A., Skinner, K.A., Vasudevan, R., Johnson-Roberson, M.: Modeling camera effects to improve visual learning from synthetic data. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (September 2018)
2. Carlson, A., Skinner, K.A., Vasudevan, R., Johnson-Roberson, M.: Sensor transfer: Learning optimal sensor effect image augmentation for sim-to-real domain adaptation. IEEE Robotics and Automation Letters **4**(3), 2431–2438 (2019). https://doi.org/10.1109/LRA.2019.2896470
3. Clouet, A., Vaillant, J., Alleysson, D.: The geometry of noise in color and spectral image sensors. Sensors **20**(16) (2020), https://www.mdpi.com/1424-8220/20/16/4487
4. Farrell, J., Okincha, M., Parmar, M.: Sensor calibration and simulation. In: DiCarlo, J.M., Rodricks, B.G. (eds.) Digital Photography IV. vol. 6817, pp. 249 – 257. International Society for Optics and Photonics, SPIE (2008)
5. Higgins, P., Gaoussou Youssouf Kebe, K.D., Don Engel, F.F., Matuszek, C.: Towards Making Virtual Human-Robot Interaction a Reality. In: 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions (VAM-HRI) (March 2021)
6. Kebe, G.Y., Higgins, P., Jenkins, P., Darvish, K., Sachdeva, R., Barron, R., Winder, J., Engel, D., Raff, E., Ferraro, F., Matuszek, C.: A spoken language dataset of descriptions for speech-based grounded language learning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
7. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors **12**(2), 1437–1454 (2012)
8. Konnik, M., Welsh, J.: High-level numerical simulations of noise in ccd and cmos photosensors: review and tutorial (2014)
9. Landau, M.J., Choo, B.Y., Beling, P.A.: Simulating kinect infrared and depth images. IEEE Transactions on Cybernetics **46**(12), 3018–3031 (2016)
10. Liu, C., Freeman, W., Szeliski, R., Kang, S.B.: Noise estimation from a single image. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1, pp. 901–908 (2006)
11. Nguyen, C.V., Izadi, S., Lovell, D.: Modeling kinect sensor noise for improved 3d reconstruction and tracking. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission. pp. 524–530 (2012)
12. Richards, L.E., Nguyen, A., Darvish, K., Raff, E., Matuszek, C.: A manifold alignment approach to grounded language learning. In: Unpublished Proceedings of the 8th Northeast Robotics Colloquium (2019)
13. Smisek, J., Jancosek, M., Pajdla, T.: 3D with Kinect, pp. 3–25. Springer London, London (2013)
14. Sweeney, C., Izatt, G., Tedrake, R.: A supervised approach to predicting noise in depth images. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 796–802 (2019). https://doi.org/10.1109/ICRA.2019.8793820