Depth Estimation Through Translucent Surfaces

Siyu Dai*[1], Xibai Lou*[1], Petter Nilsson[1], Shantanu Thakar[1], Cassie Meeker[1], Ariel Gordon[1], Xiangxin Kong[1], Jenny Zhang[1], Benjamin Knoerlein[1], Ruguan Liu[1], Bhavana Chandrashekhar[1], Sisir Karumanchi[1]

***Abstract*— In this paper, we tackle the novel computer vision problem of depth estimation through a translucent barrier. This is an important problem for robotics when manipulating objects through plastic wrapping, or when predicting the depth of items behind a translucent barrier for manipulation. We propose two approaches for providing depth prediction models the ability to see through translucent barriers: removing translucent barriers through image inpainting before passing to standard depth prediction models as input, and directly training depth models with images with translucent barriers. We show that image inpainting allows standard learned monocular and stereo depth estimation models to achieve 3 cm MAE for predicting depth of shelved items behind plastic, whereas training with real images with translucent barriers allows them to achieve centimeter or sub-centimeter MAE. We demonstrate in real robot experiments that depth-aided space estimation allows the robot to place 46% additional items into shelves with translucent barriers. This paper also provides a publicly available dataset of objects occluded by translucent barriers in a tabletop environment and a shelf environment which will allow others to contribute to this novel problem that's critical for many robotic manipulation applications including suction gripping and item packing (available at `https://sites.google.com/view/vulcan-depth-estimation`).**

I. INTRODUCTION

Depth estimation is a task of key importance in robotics since it is a prerequisite to pose estimation, manipulation, and navigation tasks to name a few. Decades of research have been dedicated to this problem, and multiple approaches have been explored, including active sensing [1], [2], [3], hardware-enabled methods [4], [5], and deep learning methods [6], [7], [8], [9]. Notably, substantial effort [10], [11] has been dedicated to the challenging task of estimating depth information for transparent, translucent, and reflective surfaces, which are prevalent in everyday environments.

In this paper, we tackle a complementary problem: rather than estimating the depth *of* translucent surfaces, we predict the depth of objects *through* such surfaces for robotic manipulation tasks. This problem is of particular importance to real-world applications in warehouses and manufacture lines. For example, robots are tasked with storing billions of objects onto shelves that secure them by translucent barriers, shown in Fig. 1(a). Planning action primitives can be challenging since the depth sensor will only capture the translucent barrier as illustrated in Fig. 1(b - c). Our approach enables depth sensing over the barrier to estimate available volume more accurately than solely relying on 2D segmentation masks, as shown in Fig. 1(e) over (d). This enhancement allows

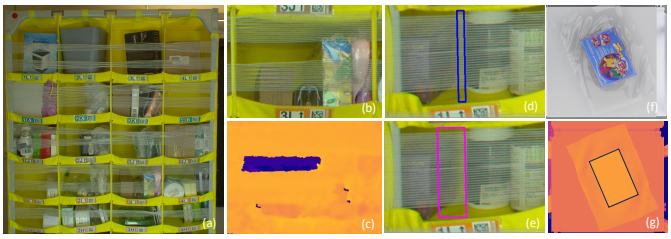


Fig. 1: (a) A shelf with translucent barriers (e.g. elastic bands) (b) A container in the shelf and (c) the corresponding depth from a depth sensor; (d) available space detected in a container through 2D occupancy and (e) available space detected through depth-based correction. Similarly, (f) an item on a conveyor with plastic wrapping and (g) the corresponding depth profile with the plastic wrapping.

objects previously considered oversized to be accommodated, which increases storage space utilization. In addition, motion primitives for barrier opening also requires object depth behind barriers to minimize item damage.

Another application is when the objects on factory conveyors are wrapped in soft plastic bags and robots are tasked with grasping them. Object depth through the plastic bag can provide crucial shape and dimension knowledge which aids grasp planning, as illustrated in Fig 1(f - g). In addition, depth through soft plastic also avoids inflate object dimension when the robot needs to pack them tightly in the storage space. Our study on 54076 unique real-world objects with or without plastic packaging showed by we can save 7.17% volume by ignoring soft transparent plastic.

To our knowledge, literature in this area is focused on estimating depth of the translucent surface and not through it [12]. TransProteus [13], [14] contributed relevant datasets in simulation but is limited to pure computer vision tasks. In this paper, we explore methods for tackling the latter problem for robotic manipulation, including learned monocular and stereo depth estimation through translucent barriers, augmenting barrier-less images with synthetic renderings of barriers, and generative methods that remove translucent barriers prior to depth estimation. We compare the performance of different depth estimation approaches and the impact on real robots, and demonstrated that depth-aided space estimation allows the robot to place 46% additional items into shelves with transluscent barriers. Our models were trained and evaluated on datasets of both real and simulation images with translucent barriers (Sec. III). This dataset will be made publicly available as part of the publication of this paper.

To summarize, the main contributions of our paper are:

- We introduce a new computer vision task of depth

*Equal contribution.

[1]All authors are with Amazon Robotics, Seattle, WA 98109, USA.

estimation through a translucent barrier.
- We release a dataset for training and evaluation of this task.
- We explore various learning-based methods, and compare their performance and trade-offs.

II. RELATED WORK

A. Monocular, Stereo and Multi-view Depth Estimation

Classical depth estimation approaches typically rely on finding corresponding pixels across multiple cameras or views and using the stereo or multi-view geometry to lift the pixels into 3D space using triangulation [15], [16], but often suffer in the presence of large uniform regions and occlusions. In recent years, classical approaches are generally outperformed by learned ones [17], [18], [19], [20], [21], [8], [22], especially in terms of the completeness metric, including on featureless surfaces [23]. Monocular depth estimation through deep learning has been particularly popular [24], [25], [26], [27], [28], [29], with approaches ranging from encoder-decoder architectures, to architectures that provide inductive bias, and to foundation models pretrained on large-scale unlabeled datasets. Neural Radiance Fields (NeRF) have shown the ability to estimate depth in a multi-view camera setup [30] and tackle translucent and reflective surfaces by baking their reflective properties into the reflectance model [31]. While a NeRF might be utilized to see through translucent surfaces, it generally requires a large number of cameras and training a model per scene, with generalization across scenes still being a major challenge [32].

Unlike the previous approaches that directly estimate the depth of uncovered surfaces, in our paper, we explore the new task of depth estimation through translucent barriers and evaluate several learned monocular and stereo-based approaches in the established literature on the new task.

B. Depth sensing

Multiple depth sensing technologies have been developed in the past decades: structured light [33], active stereo [34], Lidar [1] and depth from focus [35] to name a few. Each technology has its own strengths and weaknesses, but transparent and translucent surfaces have always been a major challenge for each one of them. A transparent surface reflects light, but the reflection pattern violates the assumptions of the sensor algorithm, such as a Lambertian surface in the case of active stereo or a single return in the case of lidar.

C. Reflection removal, defencing and inpainting

While estimating depth from images through obstacles is still an understudied domain, the removal of the obstacles themselves is a very active research area. Reflection removal from images captured through a window pane [36], removal of fences from imagery to render an unobstructed view of what's behind [37], and inpainting areas in the image that are obstructed by an obstacle [38], have been landmark problems in computer vision. Our paper leverages these methods for estimating depth by first rendering an unobstructed view of our scene and then applying standard depth estimation methods to the result. We compare this two-stage approach to alternatives that estimate depth through the obstacle directly.

D. Depth estimation of transparent and translucent surfaces

The limitations of depth sensors in capturing data on transparent and translucent surfaces have led to a recent shift towards deep learning-based approaches [39], but acquiring large-scale depth datasets for transparent objects is particularly challenging. With advances in photo-realistic simulation, many approaches [40], [41] are trained on simulated data and demonstrate impressive generalizations to real data. Datasets [13], [14] have also provided to predict depth through such surfaces, but they are limited to simulation environments and are focused solely on computer vision tasks. In contrast, this paper introduces a large-scale dataset containing both real and simulated translucent barriers with corresponding ground truth depth data. We evaluated several methods on this dataset for robotic manipulation tasks and demonstrated the effectiveness of the improved depth using a UR10-based robotic manipulation system that operates daily in real warehouses.

III. DATASET

In addition to a novel problem in the space of depth estimation, we are also releasing a dataset containing two different types of scenes that are commonly found in everyday scenarios (summary see Table I). The first scene type in our dataset has real-world and simulation shelf images broken down into individual containers, as shown in Fig. 2(d) and Fig. 3(d) respectively. The second scene is a tabletop scenario with items contained in plastic packaging, shown in Fig.4.

The real-world dataset has an RGB stereo image pair of shelves covered with translucent barriers as shown in Fig. 2(a) and (b), a depth map of the unobstructed scene as shown in Fig. 2(e), and corresponding RGB images of the unobstructed scene as shown in Fig. 2(f). This dataset includes 500 shelves with different shelf structures and 1500 unique objects randomly arranged in each image. The RGB images were collected with industrial RGB cameras, and the ground truth depth was collected with an active stereo sensor. The translucent barrier (i.e. elastic bands) was carefully attached after capturing images in the unobstructed scene to avoid moving the contents of the shelf during the unobstructed captures. Samples where the contents were accidentally moved during the attachment of the bands were algorithmically pruned out of the dataset. We also provide data with synthetically rendered translucent barriers on RGB images without barriers using methods described in Section IV-B.

The simulation images are generated with Isaac Sim [42], which is an application provided in Nvidia Omniverse [43], and includes 1987 unique 3D scanned objects. The simulated shelf dataset also includes RGB images with and without barriers and depth images without barriers collected with the same camera setup as the real shelf data, as shown in Fig. 3. The simulated tabletop dataset consists of 10k image sets with 6.7k objects wrapped in soft plastic and 3.3k objects wrapped in hard plastic. Using the Nvidia Omniverse Path-Tracing Renderer in Isaac Sim with 64 samples per pixel, RGB and depth ground truth data are generated from each

Scenario	Env.	Barrier	# of Images	Details
Real shelf	Real	No bands	25,832	RGB, depth w/o bands
	Real	Real bands	51,811	RGB, depth w/ bands
	Real	Rendered bands	24,136	RGB, depth w/o bands
Simulated shelf	Simulated	Simulated bands	322,584	RGB, depth w/ bands
	Simulated	No bands	322,584	RGB, depth w/o bands
Simulated tabletop	Simulated	Plastic packaging	10,000	RGB, depth w/ container
	Simulated	No plastic	10,000	RGB, depth w/o container

TABLE I: Datasets for predicting object depth/shape behind translucent barriers

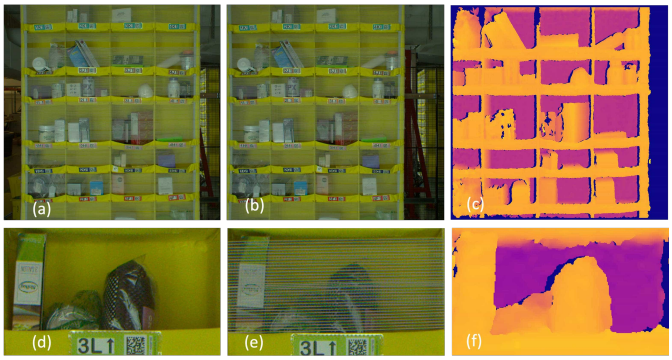


Fig. 2: (a) Left stereo image of the shelf with translucent barriers, (b) Right stereo with translucent barriers, (c) Depth image of the shelf without barrier (d) Container level image without translucent bands (e) Container level image with bands (f) Ground truth depth for the container without barrier

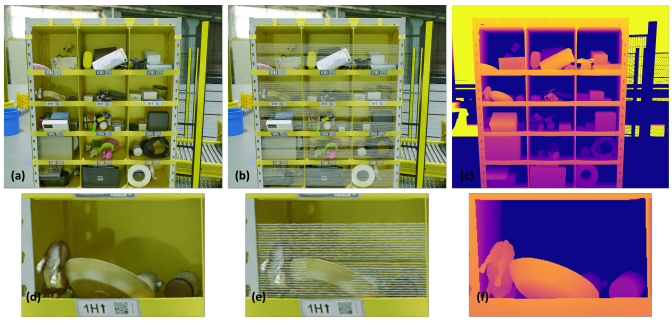


Fig. 3: (a) RGB without translucent barriers i.e elastic bands, (b) RGB with translucent elastic bands, (c) Visualization of depth without translucent elastic bands (depth values are clipped during visualization to better illustrate shelf depth). In addition to these, this dataset also contains RGB instance, semantic segmentation of items in the shelves. (d) container image without bands, (e) container image with bands (f) corresponding ground truth depth

camera with a resolution of 2448 x 2048. The soft plastic image sets contain ground truth data without transparent packaging, while the hard plastic image sets only contain data with transparent packaging. This is because the robot can't ignore hard plastic packaging during manipulation; hence, our depth model needs to learn not to see through hard plastic.

IV. APPROACH: DEPTH ESTIMATION THROUGH TRANSLUCENT BARRIERS

In this section, we present two different approaches for providing depth models with the ability to see through

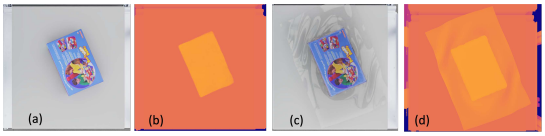


Fig. 4: (a) RGB without translucent barriers, (b) Depth image without barrier (c) RGB with translucent barriers, (d) Depth with the translucent barrier. In addition to these, this dataset also contains RGB from a stereo pair, the intrinsics and the semantic segmentation

translucent barriers: removing barriers through image inpainting before passing them to standard depth models as input, and directly training depth models on images with translucent barriers. In this paper, we refer to standard off-the-shelf depth models that are not trained to predict depth through translucent barriers as **Baseline Models**.

A. *Removal of the Translucent Barriers*

The first approach for seeing depth through translucent barriers is to conduct image inpainting to remove translucent barriers from the RGB images and then use Baseline Models to predict depth from inpainted images, as delineated in Fig. 5 (3). A GCANet [44], which is an end-to-end gated context aggregation network, is used to demonstrate the performance for image inpainting in this paper. The pre-trained GCANet lacks generalization ability towards different datasets with different RGB distributions, light settings, blurriness, resolution, and field of view. We augment our data with translucent barriers with different brightnesses (0.6–1.2x) and RGB distributions, where we multiply each RGB channel with a random number from a uniform distribution between (0.9–1.1). Moreover, we also augment the data for blurriness with a boxblur filter (kernel size 1 4), which makes the model more robust to different camera focus and resolution. Images that has been passed through the inpainting network will be the input image through Baseline Models to predict the depth of the items in the scene.

B. *Generating Data with Translucent Barriers*

We experimented with three different approaches to generate training data for models to predict depth without barriers using RGB images with barriers as input. The first method is to directly render photo-realistic plastic barriers in simulation environments to create RGB and depth image pairs with and without translucent barriers. The second method is to capture depth images without translucent barriers in the real environment and then add translucent barriers before capturing RGB images. As this method can be very time-consuming and it is nearly impossible to add or remove

the barrier without even slightly perturbing the scene, we also experimented with a third method that renders synthetic barriers on real unobstructed RGB images. This allows for more efficient data collection while still providing the model with in-distribution data that allows it to learn to see through translucent objects. To render synthetic barriers on real images, we collected images of translucent bands in front of a black background. These bands were overlaid on the unobstructed RGB images, using alpha blending. The location, orientation, hue, saturation, and lightness were randomly varied as a form of data augmentation.

In the shelf environment, we use the first method to generate simulation data with barriers as pre-training data and then fine-tune the model on real data with barriers generated with the second and third methods, referred to as the **Real Barrier Model** and **Rendered Barrier Model**, respectively. In the tabletop environment, we only use the first method and train the models on simulation data with barriers, referred to as **Model with Barrier**. These models are trained with the same training regimen as the baseline models.

V. APPLICATION: DEPTH-BASED PERSPECTIVE CORRECTION

Depth estimation through translucent barriers plays a critical role in our robotic manipulation systems operating daily in real-world warehouses, and one of the key application is perspective correction for accurate space estimation in shelves. We represent linear spaces in our containers as slots, and use a depth-based perspective correction method to transform 2D slots obtained from 2D object segmentation masks into accurate 3D linear distances. Examples of additional space discovered through perspective correction are shown in Fig. 6.

Our space estimation algorithm for containers occluded by translucent barriers relies on occupancy maps obtained by segmenting items behind barriers. Raw occupancy maps obtained from 2D RGB images are subject to perspective errors and inflated occupancy caused by item rotation. As our robotic manipulators are capable of consolidating spaces by sweeping, we want to modify the raw occupancy map to accurately represent the post-consolidation occupancy of items in the containers. We use a combination of 2 methods to achieve this: 1) occupancy correction through depth, illustrated in Fig. 7; 2) linear space (slot) correction through depth, illustrated in Fig. 8.

For easy identification of items, warehouses typically require items to be stored closer to the front of the container. Therefore, we can use depth maps to remove pixels on the occupancy map that are beyond a certain percentage of the container depth as they are usually pixels on the side of the item. In Fig. 7, we can see that additional linear space is identified after removing occupied pixels that are beyond 30% of the container depth.

To accurately represent the linear space (slots) in the 3D space, we utilize the geometric relationship between camera distance to shelf and edge depth of the slots to conduct perspective correction. As shown in Fig. 8, the corrected slot width can be calculated as $y_{true\_slot} = y_{estimated\_slot}/x_{cam\_to\_shelf\_dist} * x_{slot\_edge\_depth}$.

VI. EXPERIMENT RESULTS

Our proposed approaches can be applied to any depth prediction model architectures. Here we chose two common low-latency models to evaluate the performance across all different methods, including the baseline models and our proposed approaches: 1) a customized U-Net [45]-based monocular depth prediction model with residual connections and ResNet [46] blocks, and 2) a stereo depth prediction model adapted from CREStereo [47] by removing its iterative refinement mechanism, as it contradicts with our goal of ignoring the translucent barrier. We evaluate the performance of the Baseline Models as well as models that learned to see through translucent barriers described in Section IV on both the shelf domain and the tabletop domain.

A. Real Shelf Experiment

We evaluated each of our methodologies on an evaluation set with bands and on the same evaluation set without bands. Table II compares the MAE (m) of different models on the test set of real shelf images with and without translucent barriers. The baseline monocular and stereo models are also evaluated on inpainted images generated with the methods described in Section IV-A. Example images of predicted depth from different models are shown in Fig. 10. Fig. 9a provides error curves for the different models, where we show the percentage of pixels in the dataset that have an MAE above a given threshold. In these cases, lower curves are better.

Results in Table II show that both the baseline monocular and baseline stereo models have sub-centimeter accuracy when predicting depth on images without bands. However, when the bands are added to the scene, the MAE increases to above 5 cm for the monocular model and above 4 cm for the stereo model. This is expected for models that have never seen scenes with translucent barriers, and are therefore predicting the depth of these barriers instead of the depth of items behind the barriers. When evaluated on images where the translucent barriers were removed through image inpainting, the baseline monocular model's MAE decreased by 45%, and the baseline stereo model's MAE decreased by 6%. This shows that removing barriers through inpainting can improve the baseline monocular model's ability to see through barriers, but is not very effective for the baseline stereo model. Both models are not able to achieve sub-centimeter accuracy with inpainted images, showing that our inpainting method is not yet sufficient to recreate the no band conditions.

In contrast, the monocular and stereo models trained on the dataset with real translucent barriers have centimeter or sub-centimeter accuracy on images both with bands and without bands. Table II shows that the Real Barrier Monocular Model can predict depth through barrier with a 1 cm MAE, and the Real Barrier Stereo Model can achieve a 0.8 cm MAE. Despite being trained on RGB images with barriers, these

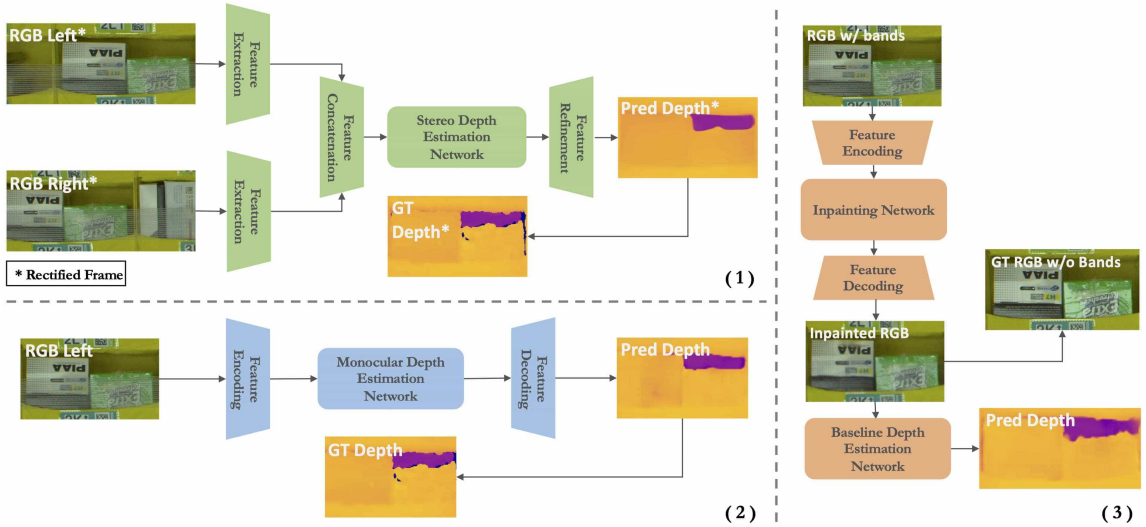


Fig. 5: Depth estimation over the translucent barrier pipelines of each approach. Learned stereo, monocular, and inpainting methods are shown in (1), (2), and (3), respectively.

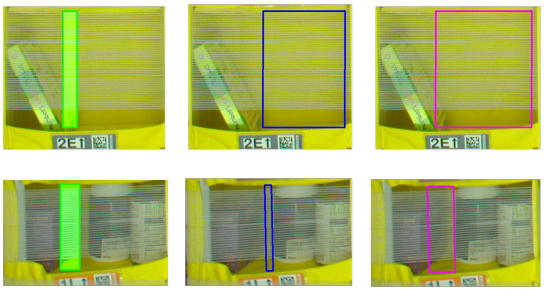


Fig. 6: Illustration of additional free space discovered with perspective correction. Green shade represents the difference between the raw (2D) and the perspective-corrected free space, blue and magenta box represents the raw (2D) and perspective-corrected free space, respectively.

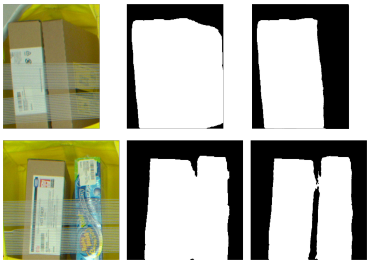


Fig. 7: Occupancy correction through depth. From left to right: RGB image, raw 2D occupancy map, depth-corrected occupancy map. In occupancy maps, white pixels represents occupied space and black pixels represent unoccupied space.

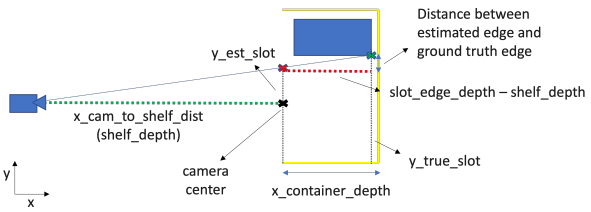


Fig. 8: Linear space (slot) correction through depth.

Model	Images w/o bands	Images w/ bands	Images w/ bands inpainted
Baseline Monocular	0.00914	0.05838	0.03246
Baseline Stereo	0.00876	0.0426	0.0327
Real Barrier Monocular	0.00917	0.01001	n/a
Real Barrier Stereo	0.0112	0.00893	n/a
Rendered Barrier Monocular	0.01006	0.03113	n/a
Rendered Barrier Stereo	0.00864	0.0201	n/a

TABLE II: MAE (m) for real shelf dataset

models don't lose performance even when the barriers are not present, meaning that we do not need separate models for cases with and without translucent barriers.

Models trained on images with rendered translucent barriers perform better than baseline models for predicting depth through barriers, but perform worse than the models trained on data with real barriers. The Rendered Barrier Monocular Model has a 3 cm MAE and the Rendered Barrier Stereo Model has a 2 cm MAE. It's possible that this performance could be further improved by additional augmentations to the rendered barriers during data generation.

We conclude that the best way to see through the translucent barriers for this scene is using a stereo model which is trained with images that contain the barrier in the training data. The model is able to learn how to see through the bands and achieves performance that is approximately equivalent (sub-centimeter), to the performance on images without bands. All models in this section were pre-trained on corresponding simulation shelf data.

B. Simulated Tabletop Experiment

In the tabletop domain, we also trained and evaluated monocular and stereo models on simulated images with and without translucent barriers. In this dataset, there is a distinction between soft plastic and hard plastic: we want to ignore the soft plastic but not the hard plastic because hard plastic affects grasp actions and dimension estimation. The MAE and error curves for the different models are shown in Table III and Figure 9b respectively, and example images are shown in Fig. 11. In this domain, the baseline models do

Model	Images w/o bands	Images w/ bands
Baseline Monocular Model	0.00589	0.00799
Baseline Stereo Model	0.0133	0.0279
Monocular Model with Barrier	0.00461	0.00464
Stereo Model with Barrier	0.03321	0.03652

TABLE III: MAE (m) for simulated tabletop dataset

	Free space detected (px)	**Space of All Containers (px)**	**Detected free Space (%)**	**Items Filled into the Shelf**
Raw 2D space	25018	86828	28.81%	37
Perspective-corrected	26928	86828	31.01%	54
Improvement	1910	N/A	2.20%	17

TABLE IV: Comparison of Raw 2D Slots and Perspective-Corrected Slots

not have as significant a drop in performance between images with and without plastic. This is likely because some images in the dataset have hard plastic occluding the item, and the models are only trained to ignore soft plastic packaging.

The monocular model trained on data with barriers outperforms the baseline model by 42%. However, training on the barriers actually makes performance worse for the stereo models. In these experiments, monocular models (both baseline and our models) outperform stereo models, which was not true for the shelf environment. We hypothesize that the monocular model is able to differentiate between soft and hard plastic, but feature matching-based models, such as the learned stereo, struggle with whether to ignore a matched pair because 1/3 of the data contains hard plastic, which the model must learn, and the other 2/3 of the data contains soft plastic, which the model must ignore.

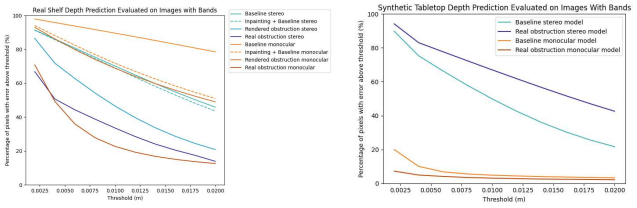
We do not evaluate the performance on inpainted images in this domain as generating corresponding data without barriers is trivial in simulation. We also don't train rendered barrier models for this dataset as the rendered and real barriers are equivalent in simulation.

C. Real-world Robotic Manipulation Experiment

We tested the perspective correction performance of the **Real Barrier Monocular Model** on 5 shelves total, each with a unique distribution of containers and are 39% full in terms of volume. We conducted A/B testing to compare the linear space detected by 2D object segmentation and by 3D depth-based perspective correction. Between the A/B tests on each shelf, items inside the containers are maintained in the same pose, and items to be filled into the pod are queued in exactly the same order. As shown in Table IV, the robot was only able to store 37 items into the shelves with the 2D segmentation-based space detection method, but was able to store 54 items (a 46% increase) with the 3D depth-corrected space detection method.

VII. CONCLUSION

We present a novel problem of estimating depth through translucent surfaces, which helps robots manipulate items behind or wrapped in materials like plastic packaging. In addition to a public dataset that will allow others to develop new methods to solve this problem, we also propose



(a) Real shelf depth prediction performance (b) Synthetic tabletop depth performance

Fig. 9: Percentage of pixels with error above the threshold (%) vs threshold

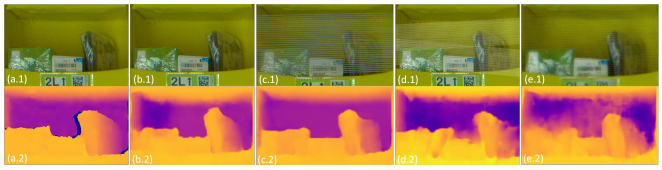


Fig. 10: Monocular depth prediction for each of the methods (a) 1. Container image without bands, 2. Ground truth depth (b) Container image without bands and the corresponding predicted depth (c) Container image with bands and the predicted depth (d) Container image with fake bands and the predicted depth (e) Container image with bands inpainted to remove bands and the predicted depth

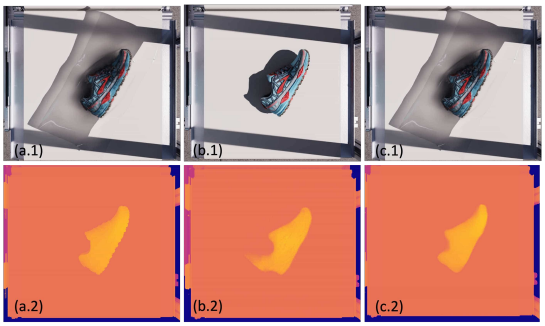


Fig. 11: (a) Item with plastic and the corresponding ground truth depth (b) Item without plastic, and its predicted depth (c) Item image with plastic and its predicted depth

several methods to solve the problem, including removing translucent barriers through inpainting and training learned monocular and stereo models on images with these depth occlusions. We show that both methods outperform baseline models trained on images without translucent barriers in a simulated tabletop domain and a real shelf domain. Directly training depth models on real images with barriers is the most effective way and is able to achieve sub-centimeter performance, but removing barriers through inpainting and rendering synthetic bands provides more efficient alternatives for data collection. We conclude that it is possible to predict depth through translucent barriers, and, through real robot experiments, we demonstrated that this capability allows us to better estimate item shape and dimension for manipulation in downstream tasks.

REFERENCES

[1] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation with the 3d lidar and stereo fusion," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2156–2163.

[2] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.

[3] Q. Li, M. Biswas, M. R. Pickering, and M. R. Frater, "Accurate depth estimation using structured light and passive stereo disparity estimation," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 969–972.

[4] "Depth from focus," *Pattern recognition letters*, vol. 5, no. 1, pp. 63–69, 1987.

[5] L. Zhang *et al.*, "Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multiview stereo," in *Proceedings ninth IEEE international conference on computer vision*. IEEE, 2003, pp. 618–625.

[6] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, 2021.

[7] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig, "Monocular depth estimation using deep learning: A review," *Sensors*, vol. 22, no. 14, p. 5353, 2022.

[8] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 4, pp. 1738–1764, 2020.

[9] M. Poggi, F. Tosi, K. Batsos, P. Mordohai, and S. Mattoccia, "On the synergies between machine learning and binocular stereo for depth estimation from images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5314–5334, 2021.

[10] J. Yoon, W. R. Zame, and M. Van Der Schaar, "Deep sensing: Active sensing using multi-directional recurrent neural networks," in *International Conference on Learning Representations*, 2018.

[11] M. W. Tao, J.-C. Su, T.-C. Wang, J. Malik, and R. Ramamoorthi, "Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 6, pp. 1155–1169, 2015.

[12] J. Jiang, G. Cao, J. Deng, T.-T. Do, and S. Luo, "Robotic perception of transparent objects: A review," *IEEE Transactions on Artificial Intelligence*, 2023.

[13] S. Eppel, H. Xu, Y. R. Wang, and A. Aspuru-Guzik, "Predicting 3d shapes, masks, and properties of materials inside transparent containers, using the transproteus cgi dataset," *Digital Discovery*, vol. 1, pp. 45–60, 2022.

[14] Y. R. Wang, Y. Zhao, H. Xu, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, "Mvtrans: Multi-view perception of transparent objects," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 3771–3778.

[15] Y. Furukawa, C. Hernández *et al.*, "Multi-view stereo: A tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 1-2, pp. 1–148, 2015.

[16] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion*." *Acta Numerica*, vol. 26, pp. 305–364, 2017.

[17] K. Zhou, X. Meng, and B. Cheng, "Review of stereo matching algorithms based on deep learning," *Computational intelligence and neuroscience*, vol. 2020, 2020.

[18] H. Zeng, B. Wang, X. Zhou, X. Sun, L. Huang, Q. Zhang, and Y. Wang, "Tsfe-net: Two-stream feature extraction networks for active stereo matching," *IEEE Access*, vol. 9, pp. 33 954–33 962, 2021.

[19] J. Wang, S. Zhang, Y. Wang, and Z. Zhu, "Learning efficient multi-task stereo matching network with richer feature information," *Neurocomputing*, vol. 421, pp. 151–160, 2021.

[20] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5695–5703.

[21] C. Wang, X. Bai, X. Wang, X. Liu, J. Zhou, X. Wu, H. Li, and D. Tao, "Self-supervised multiscale adversarial regression network for stereo disparity estimation," *IEEE Transactions on Cybernetics*, vol. 51, no. 10, pp. 4770–4783, 2020.

[22] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 919–21 928.

[23] C. Stucker and K. Schindler, "Resdepth: Learned residual stereo reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 184–185.

[24] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Neural Information Processing Systems*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:2255738

[25] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[26] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[27] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4009–4018.

[28] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," *arXiv preprint arXiv:2401.10891*, 2024.

[29] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *arXiv preprint arXiv:2404.15506*, 2024.

[30] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "Nerf: Neural radiance field in 3d vision, a comprehensive review," *arXiv preprint arXiv:2210.00379*, 2022.

[31] Y.-C. Guo, D. Kang, L. Bao, Y. He, and S.-H. Zhang, "Nerfren: Neural radiance fields with reflections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 409–18 418.

[32] I. Petrovska and B. Jutzi, "Vision through obstacles—3d geometric reconstruction and evaluation of neural radiance fields (nerfs)," *Remote Sensing*, vol. 16, no. 7, p. 1188, 2024.

[33] Z. Cai, X. Liu, G. Pedrini, W. Osten, and X. Peng, "Accurate depth estimation in structured light fields," *Optics Express*, vol. 27, no. 9, pp. 13 532–13 546, 2019.

[34] J. J. Clark, "Active photometric stereo." in *CVPR*, vol. 92, 1992, pp. 29–34.

[35] Jahne and Geissler, "Depth from focus with one image," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1994, pp. 713–717.

[36] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 633–640.

[37] Z. Shi, Y. Bahat, S.-H. Baek, Q. Fu, H. Amata, X. Li, P. Chakravarthula, W. Heidrich, and F. Heide, "Seeing through obstructions with diffractive cloaking," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[39] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang, "Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1757–1763.

[40] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 3634–3642.

[41] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-nerf: Using a neural radiance field to grasp transparent objects," *arXiv preprint arXiv:2110.14217*, 2021.

[42] Nvidia. (2024) Isaac sim introduction. what is isaac sim? - omniverse isaacsim latest documentation. [Online]. Available: http://web.archive.org/web/20080207010024/

[43] ——. (2024) Nvidia omniverse. [Online]. Available: https://www.nvidia.com/en-us/omniverse/

[44] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua, "Gated context aggregation network for image dehazing and deraining," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1375–1383.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 263–16 272.