Final Report: Detection of Mirrors in 3D Scenes from 2D Images

AJ Arnolie
Stanford University
ajarno@stanford.edu

Youngbae Son
Stanford University
ybson@stanford.edu

Chunwei Chen
Stanford University
hillchen@stanford.edu

https://github.com/ybson-git/CS231aProject

Abstract

In this project, we investigate the effect of mirrors on the task of 3D scene understanding both in regard to depth sensors as well as recent Monocular Depth Estimation methods. Using commercial RGB-D sensors, we first show that the presence of mirrors poses challenges for 3D reconstruction, as it is difficult to differentiate a mirrored scene from the actual scene using solely RGB data, and the depth measurements of mirrors produced by such sensors are often inaccurate. We then propose a method to mitigate existing limitations of depth estimation in scenes with mirrors that marginally improves upon the quality of previous approaches. We utilize Mirror3DNet, a mirror detection architecture that detects mirrors in 2D images, jointly predicts the corresponding mirror planes and mirror masks for each, and uses this information to refine depth estimation. Pairing this model with some of newest and top performing Monocular Depth Estimation Models that have been released in the past couple months, we show that this architecture can yield significant improvements in mirror depth estimation over commercial RGB-D sensors using solely RGB data, reducing mirror-induced depth estimation error. Additionally, we provide a small dataset of RGB-D data for scenes containing mirrors (captured with an Azure Kinect sensor), and we offer a widely accessible method for collecting RGB-D data through the use of the iPhone TrueDepth camera.

1. Introduction

Mirrors and specular surfaces appear incredibly often in our everyday lives, and thus, understanding these surfaces is critical to gaining a full 3D understanding of scenes in a wide variety of scenarios. Though recent progress certainly has been made in the fields depth sensing and 3D reconstruction using computer vision-based techniques, specular surfaces such as mirrors are a significant source of error for many computer vision architectures. This is largely because it is extremely challenging to distinguish between the scene a mirror reflects and the actual scene. As a result, many systems, including both commercial depth sensors and computer vision architectures, have trouble accurately estimating the depth of these areas in a scene. As mirrors are highly reflective surfaces, the appearance of such surfaces in depth measurements can cause either no signal or highly unreliable depth estimates to return. Such inaccurate depth estimation can pose challenges for extensive 3D reconstruction and for a number of different applications that require a sound 3D understanding of a scene.

Figure 1. Mirrors appear everywhere in our everyday lives: Plane mirrors in rooms and convex mirrors on roads, for example. Being able to reconstruct 3D scene accurately in these scenarios will be valuable for future applications including virtual house visits and autonomous driving tasks.

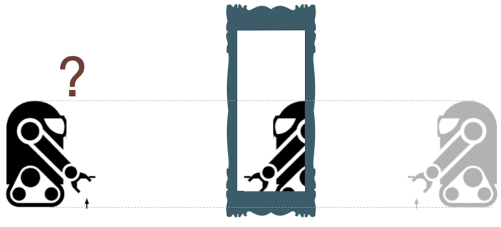

Figure 2. The presence of mirrors makes 3D understanding and 3D reconstruction difficult: RGB images do not tell us whether we are looking at mirror or an opening to another scene.

In terms of mitigating these specular surface-related issues, it follows that accurate detection of where mirrors are in a scene would be incredibly helpful in reducing the depth estimation error of these surfaces, allowing us to de-

velop more reliable methods for 3D reconstruction. Having this data would allow us to either either selectively omit the reflected areas from being used towards an inaccurate 3D reconstruction, or perhaps use this data to better reconstruct the 3D scene with the understanding that it is indeed a reflection. This is the method we propose in this paper, and we will be exploring how this is currently being done and how it might be improved.

2. Background and Related Work

2.1. Monocular Depth Estimation

Monocular Depth Estimation, the task of estimating the depth of a 3D scene from a single 2D image, is an especially challenging task in the field of computer vision and one of the more crucial building blocks for many of the more complex vision-based tasks such as navigation and planning for autonomous driving systems and 3D scene reconstruction. Depth provides critical information about the 3D structure of a scene, and Monocular Depth Estimation allows us to estimate this new dimension of data from the small amount of information a single RGB image provides.

Over the past few decades, significant progress has been made in the spaces of Stereo Depth Estimation ([20], [17], [2]) and Multi-Camera/Video Depth Estimation ([24], [8]) using Structure-from-Motion techniques. This is because Stereo and Multi-Camera situations provide significantly more useful information in terms of determining depth as they give us multiple points with which to find a 3D correspondence, removing much of the ambiguity that makes the Monocular Depth Estimation problem so difficult.

Still, with more recent rapid development in novel Deep Learning techniques, we are beginning to see more and more promising performance from many new models on the task of Monocular Depth estimation ([1], [5], [13], [18], [21], [25]). We will discuss a number of these recently developed methods in greater detail later in this paper. Still, though a variety of very effective methods have been developed for the task of Monocular Depth Estimation over the span of the past couple years, it it certainly still an open field with a number of unsolved issues. For example, these methods have significant difficulties handling object occlusions, ambiguous textures and 3D structures, and especially, depth for reflective surfaces such as mirrors.

2.2. Mirror Detection and Mirror Depth Estimation

As mentioned in the previous section, one of the most widely-recognized challenges associated with Monocular Depth Estimation and with most vision-based tasks in general is the handling of cases in which mirrors are present within a scene. In these cases, it often becomes incredibly difficult for depth sensors and depth estimation models to detect the difference between some reflected virtual scene and the true scene from 2D images, a task that is even challenging for humans on occasion.

It is challenging to train and test Monocular Depth Estimation networks that can effectively handle mirrors because the ground truth depth data for many of the most popular depth datasets, including ScanNet[4], Matterport3D[3], and NYU Depth V2[16], does not sufficiently account for reflective surfaces such as mirrors and glass for which these datasets either have missing or incredibly noisy depth data. Thus, one often proposed solution to these challenges presented by mirrors is to add the step of Mirror Detection. The goal of the Mirror Detection task is to identify mirrors in a scene and predict some representation of those mirrors that can then be used to further inform the depth prediction task. These detections are often represented as 2D masks of pixels corresponding to mirrors in the 3D scene or 3D planes estimating the true location of the mirror within the scene.

In recent years, there has been increased interest in the task of Mirror Detection, particularly for the application of addressing the mirror-related issues with Depth Estimation as discussed in the previous section ([23], [22]). Additionally, [21], a model we will be considering and analyzing later in this paper, performs Mirror Detection for the sake of improving Monocular Depth Estimation as well. The method proposed in this paper requires the assumption that mirrors are planar, though this doesn't hold in all cases. We address the flaws in this assumption in Section 5 of this paper. The current demonstrations of these models, however, are generally limited to only using the most popular RGB-Depth datasets, namely Matterport3D, ScanNet, and NYU Depth V2, with no demonstration of real-life usage and no straightforward method through which to test custom images on the models. Therefore, one goal of our project is to bring up full end-to-end mirror detection from 2D images that can be implemented from the physical data using either Kinect RGB-D camera and iPhone camera.

3. Technical Approaches

In this section, we will address the three technical approaches towards solving the problem of depth estimation in scenes containing mirrors. First, we will discuss the performance of commercially-available depth sensors such as the Azure Kinect sensor and iPhone TrueDepth camera on this task.

3.1. Limitations of Commercial Depth Sensors

3.1.1 Azure Kinect

The Azure Kinect DK is a commercial Computer Vision Development Kit containing a 12-MP RGB camera and 1-MP depth sensor made for computer vision applications. With cross platform (Linux, Windows) Kinect SDK (K4A)

and SDK document support, developers have access to an RGB camera, depth sensor, and synchronized Depth-RGB data over the SDK APIs.

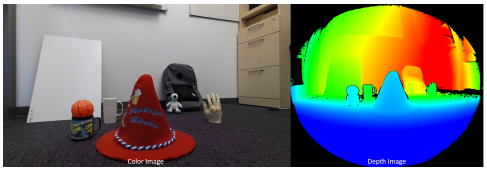

Figure 3. Azure Kinect SDK document examples.

In order to perform tests to understand the limitations of the Kinect sensor in situations where there are mirrors present, we utilized a Python wrapper pyk4a and Kinect SDK binary distribution to build out a data pipeline for device initialization, configuration (720p, 15fps), data collection, and visualization. Images passed through this data pipeline are returned as numpy arrays and behave like python objects.

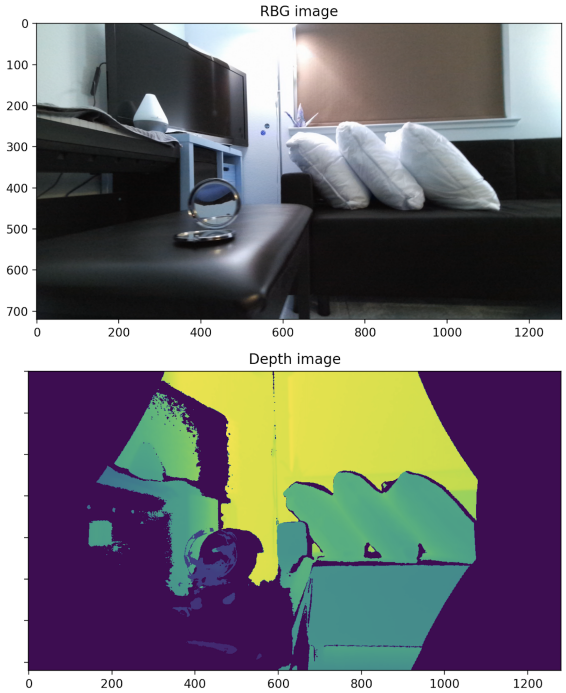

Figure 4. On the top is the 720p color image from Kinect RGB camera. On the bottom is the corresponding Depth output.

3.1.2 iPhone TrueDepth

Another potential, widely-accessible method for collecting depth information is through the use of the iPhone's TrueDepth Camera. The iPhone TrueDepth camera is a camera feature available in many of the iPhone models past the iPhone X that outputs a 2D image in which depth values are measured. The camera does this with a method quite similar to that of Azure Kinect in which, using an array of additional sensors including a proximity sensor, an infrared camera, and a dot projector, the 7 megapixel camera system uses LEDs to project a grid of thousands of infrared dots to record depth at real time speeds. Typically, this TrueDepth camera (which tends to be an iPhone's front-facing camera) is only used for face recognition during sign in, but it is possible to access this data for other purposes, as we have done for the sake of data collection.

Figure 5. On the left are the original RGB images, captured using an iPhone XR. On the right are the corresponding iPhone TrueDepth depth maps.

In order to collect data from the iPhone TrueDepth camera, we set up two useful test iPhone apps in Swift. The first app streams iPhone TrueDepth data live to the phone screen and allowed us to quickly and easily test the capabilities of the TrueDepth camera. The second app allows us to capture the depth and RGB data for a single frame of a scene from the iPhone to be passed to our model for testing. We have these apps running on an iPhone XR using the live app testing feature provided by XCode for iPhone.

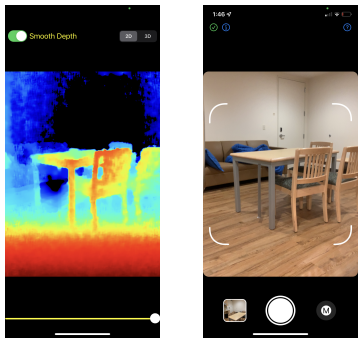

Figure 6. On the left, a test app streams and visualizes the iPhone XR's TrueDepth estimation of a scene live to the phone screen. On the right, a test app captures and saves the depth map and RGB image for a single frame to be passed the model for testing.

The experiments we performed with both Azure Kinect and iPhone TrueDepth indicate one thing although the vi-

sualization of the depth appears fairly accurate, there is extreme inaccuracy occurring around the mirror area of the image. The depth sensor predicts that either there is a high amount of depth in the area beyond the mirror plane (big value) or no depth (0 value) like a hole when it is in reality a flat plane.

3.2. State-of-the-Art Monocular Depth Estimation Models

In addition to raw depth measurement as described above, significant research has gone into the task of extracting detailed depth information from a single RGB image, a task also known as Monocular Depth Estimation (MDE). Through the development of our project, we researched some of the newest developments and best-performing models in the hope of applying these models to our overall architecture. But, as stated in the introduction, such depth estimation techniques tend to have challenges when it comes to the estimation of the depth maps of scenes containing mirrors or reflective surfaces of some kind. In order to collect a baseline for depth estimation in these cases, we first tested four of the newest and most promising depth estimation modules on mirror-based scenarios. Below, we have included overviews of the technical approaches for each of the Monocular Depth Estimation models that we considered for these tests.

All of the following models were trained on the NYU Depth V2 dataset (https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html), the same dataset the Mirror3DNet model uses to train and test on. The NYU Depth V2 dataset is a large collection of images of indoor scenes with both RGB and depth data provided for each. The data for this dataset is collected using a Azure Kinect sensor. Below, we provide an overview of the implementation and performance of each of these models within the context of the mirror detection architecture.

3.2.1 From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation (BTS) [13]

Published in 2021, the Monocular Depth Estimation model introduced in this paper argues that one of the greater challenges associated with many deep Convolutional Neural Network approaches to the Monocular Depth Estimation task is that such models with encoder-decoder architectures often have difficulty recovering dense depth prediction information from the latent space representation produced by the encoder. Though there are certainly effective ways to mitigate this issue (Multi-Layer Deconvolutional Networks, skip connections, etc.), this paper suggest a more effective way to do this would be to use local planar guidance layers,

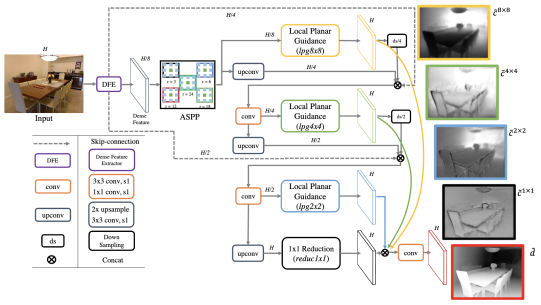

Figure 7. Network architecture for Big to Small model.

layers which use the local planar assumption to guide feature maps at each resolution level throughout the decoder of the network. This is one of the models that was used and evaluated for performance in the original Mirror3DNet paper, and we consider the results of this model in comparison to some of the newer Monocular Depth Estimation modules introduced below.

3.2.2 Enforcing Geometric Constraints of Virtual Normal for Depth Prediction (VNL) [25]

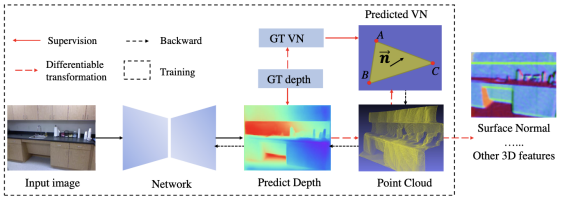

Figure 8. Network architecture for Virtual Normal Depth Prediction model.

Rather than solely focusing on pixel-wise accuracy, this Virtual Normal Depth Prediction model emphasizes the importance of 3D geometric constraints in the task of predicting depth from a single image and thus, rather than proposing major changes to network architecture, the paper proposes novel modifications to the loss calculation process used to train the model. More specifically, this architecture adds a loss term that focuses on enforcing the 3D geometric constraints provided by "virtual normal" directions sampled from points in the reconstructed 3D space. Practically, the model generates a 3D point cloud representation of the scene from the depth prediction made by the encoder-decoder network. Then, the model randomly samples three such points from this point cloud forming a "virtual plane" and enforces the constraint that local surface normals should be close to the predicted "virtual normal" corresponding to the "virtual plane". This model is among the newest and top-performing models in terms of Monocular depth estimation on the NYU V2 Dataset [16].

3.2.3 Vision Transformers for Dense Prediction (DPT) [18]

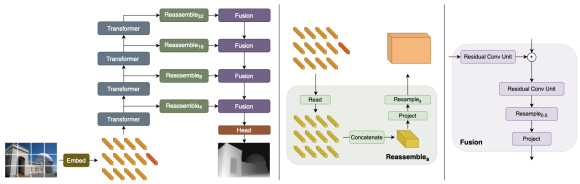

Figure 10. Network architecture for Vision Transformers model.

This work, published in 2021, introduces an approach to the Monocular Depth Estimation task which makes use of vision transformers rather than Convolutional Neural Networks to form the backbone of the model. More specifically, this model works by dividing an inputted image into a set of tokens (produced by performing feature extraction on a set of non-overlapping patches from the image). These tokens are then passed through a series of vision transformer networks from which a reassembled representation of the image is generated at each step. From here, this series of reassembled image representation is fused using a series of fusion blocks in which features are combined using Residual Convolutional Units and then upsampled. The resulting features are passed to an output head which then produces a final predicted depth map.

3.2.4 Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth (GLP) [12]

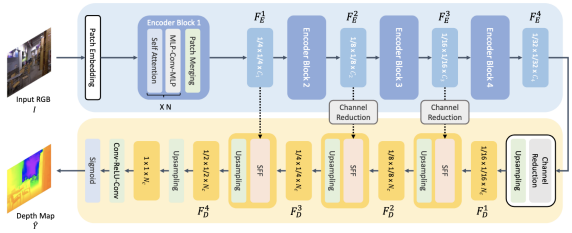

Figure 11. Network architecture for Global-Local Path Networks model.

Finally, the GLP model is one of the newest and most high-performing additions to the space of Monocular Depth Estimation, accepted to CVPR just a month ago. This model uses an encoder-decoder framework with a hierarchical transformer encoder (encodes multi-scale features to capture the global context of an image) and a lightweight decoder that considers the local connectivity of features in a scene when generating the final estimated depth map output. More specifically, the decoder features a series of Selective Feature Fusion modules that take local-context features from layers of the encoder as well as global-context features from the previous layer of the decoder and produce some hybrid feature that considers both contexts. This model produces state-of-the-art results for the Monocular Depth Estimation task on the NYU V2 dataset, and in the GLP paper, the model shows greater performance than the other models introduced above on the dataset across a number of different metrics.

3.2.5 Monocular Depth Estimation Model Comparisons

Based on with these four models and considering test outputs such as the example displayed below, we concluded that while Mirror3D used the Big to Small Monocular Depth Estimation model as their baseline depth estimation module, the more recently-published GLP model implementation in fact leads to better depth estimation performances as it produced the least noisy, most detailed depth estimations out of the models we tested. Additionally, to verify this, we performed evaluation of these Monocular Depth Estimation models on a common test set of the NYU V2 Dataset, using the raw depth information for each scene to calculate an overall Root Mean Squared Error (RMSE) for each approach. We've included the results of this testing in Table 1 below.

These values largely match the performances of the models reported in each of the papers discussed above with some minor discrepancies in terms of RMSE values. As these values also seemed to match visual comparisons of the results, an example of which can be found below in Figure 9, in the following section, we will focus on the results for refined scene depth estimation generated

Figure 9. (1) Input RGB Image, and the corresponding (2) BTS Estimated Depth Map, (3) VNL Estimated Depth Map, (4) DPT Estimated Depth Map, and (5) GLP Estimated Depth Map.

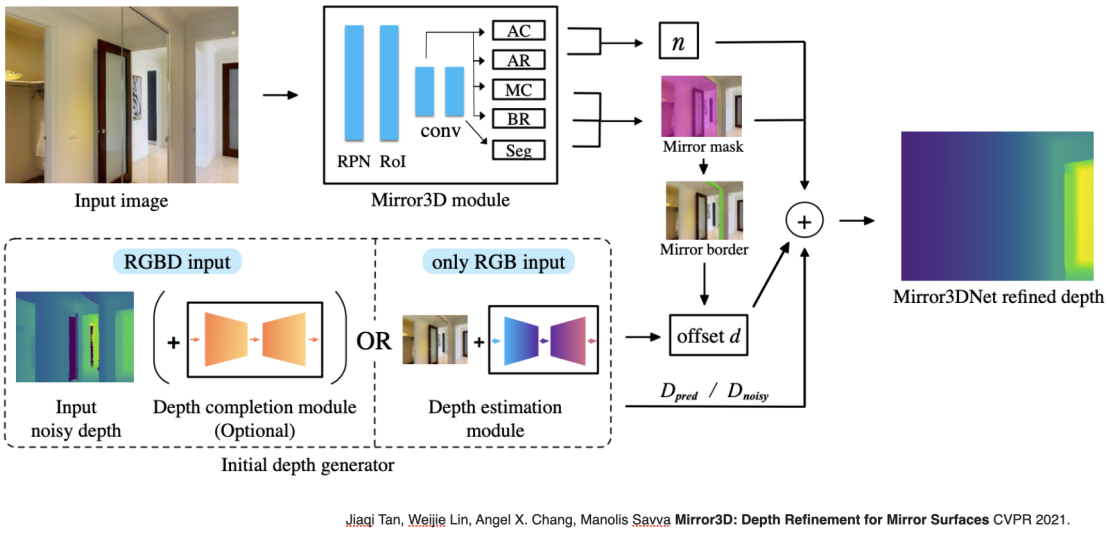

Figure 12. The structure of Mirror3DNet, available in https://github.com/AJArnolie/mirror3d which is also published in [21]

MDE Method	RMSE
Virtual Normal Depth Prediction (VNL) [25]	3.082
Big to Small (BTS) [13]	0.359
Dense Prediction Transformers (DPT) [18]	0.357
Global-Local Path Networks (GLP) [12]	**0.344**

Table 1. Results of MDE models on a test set of the NYU V2 Depth dataset.

specifically using an initial depth estimation from this Global-Local Path Networks model [12]. This model also happens to be the top performing Monocular Depth Estimation model (with available source code) on the NYU Depth V2 dataset according to metrics recorded on the PapersWithCode site for the NYU Depth V2 Dataset (https://paperswithcode.com/sota/monocular-depth-estimation-on-nyu-depth-v2).

3.3. CNN for Mirror Detection from 2D Images

Still, even considering the top performing depth estimation models tested in Section 3.2.4, it is clear to see that mirrors appearing in the scene leads to significant error in terms of depth-estimation. In order to address this issue, we now propose the addition of the Mirror3DNet module [23] to our overall model's architecture. Mirror3DNet is a model with two modules with the ultimate goal of producing a refined depth map estimation of a 3D scene containing mirrors. The first module, which we have been discussing up to this point, performs depth estimation on the input and outputs some initial depth map estimate for the scene. The second module detects mirrors within the scene, predicts their locations, and then outputs a predicted normal vector n for the mirror plane and a mirror mask for each mirror. Within this step, the predicted masks for each mirror are generated using a mirror mask segmentation MaskR-CNN module [9], and the normal vectors for the corresponding mirror planes are predicted using a 3D mirror plane estimation module inspired by PlaneRCNN [15]. The model then takes the resulting normal vectors and mirror masks and uses these values to modify the initial depth estimate with the ultimate goal of improving the accuracy of the depth estimation of the mirror surfaces within the scene (Figure 12). Note that this model architecture is build on the assumption that the mirror in the scene will planar. Thus, the overall architecture consists of three modules: mirror mask segmentation, mirror plane estimation, and depth estimation or completion.

The open source code we have applied as baseline is available at https://github.com/AJArnolie/mirror3d, a GitHub repository that corresponds to the paper published here [21]. Building upon this basic Mirror3DNet model, we restructured the code infrastructure of each of the relevant code repositories such that it can easily process any inputted RGB image, including those captured by smartphone cameras and Kinect sensors, and then we integrated some of the most recent and top performing Monocular Depth Estimation models into our our model as discussed in the previous section.

4. Experiment

After implementing and testing this model, we collected hundreds of example RGB-D images via the techniques de-

scribed in the previous section and ran the model on just the RGB data of the resulting images to retrieve a refined scene depth estimate for each. A portion of this data along with the resulting predictions of the model can be found in our project's GitHub repository. We showcase the two examples of scenes containing mirrors introduced in Figure 13, where in the first case, a mirror is hanging on the wall in the center of the image, and in the other, we are looking at a set of smaller mirrored objects sitting on top of a desk.

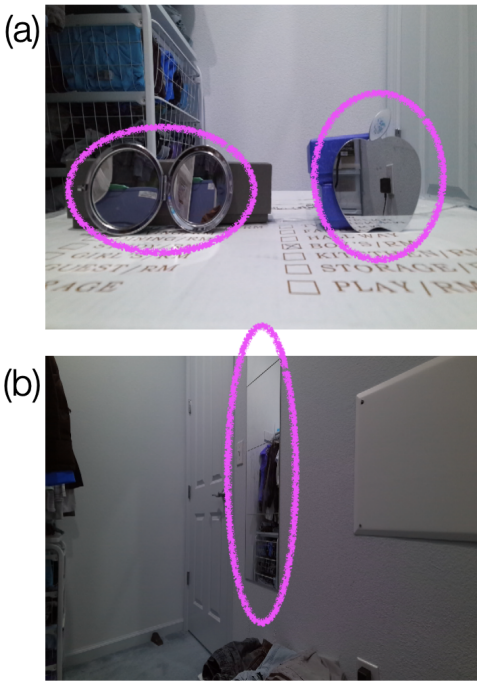

Figure 13. Two cases demonstrated here in the report. (a) shows the two mirrored objects each with different shape sitting on top of a box. We placed normal objects right behind the mirrors to compare with the depth measurement from the mirror. (b) shows the narrow mirror hanging on the wall. The depth from the neighboring wall can be compared with the depth of the mirror. Both (a) and (b) after going through various depth extraction and estimation procedures will be shown in Figures 14 and 15 respectively.

To best demonstrate the current problem and our solution, the experimental data is configured as follows: First, we show the inaccuracies of depth estimation using our Kinect sensor measurement. This is expected to be due to transmitted IR wave from the sensor gets fully bounced or scattered on the mirror surface and therefore does not bounce back to IR receiver in the sensor. Second, we show that from the same image, only based on RGB data, using the approaches mentioned in Section 3.3 can segment the mirror in the scene and lead to correct depth estimation. Consider Figures 14 and 15 on the following page to view an example of the results we received from this experimentation.

5. Conclusion

In this report, we experimentally confirmed the limitations of RGB-D sensors for depth prediction in the scenes with mirrors, and we successfully demonstrated that our model is able to accurately predict the depth of mirrors in a scene from just an RGB image. Depth measurements from the Kinect IR sensor were rather unpredictable, giving us both false highs and false lows (each case demonstrated in Figure 14 and 15), making it almost impossible to discern the true depth of the mirror just from the raw depth data. But, as shown above, the application of our model successfully handles this error in both the false high and false low cases, giving accurate estimates for the true depth of the mirror planes in both cases. We believe that commercial applications of this result could become valuable where the scene to be reconstructed has mirrors in it, such as the rooms in a house for a virtual house visit, or for mirrors on curved roads in certain autonomous driving situations.

One limitation we have found with the current work in this space is that, when detected, mirrors are always assumed to be a perfectly flat plane in the scene, which is usually true but may not always be the case. This assumption helped to simplify the implementation of the Mirror3DNet model, but it would be interesting to investigate how we might go about generalizing these concepts into something that is applicable to curved and oddly-shaped specular surfaces. Some works such as [14] (which addresses the convex mirror case) have begun exploring this idea, though this seems to be a relatively untouched subject within the field of Monocular Depth Estimation at the moment, and it was a challenge to find much more literature regarding this topic.

One additional limitation of the current model is that it does not have the capability to use information from the reflection of the mirror to improve its understanding of the 3D scene. As long as the surface of the mirror can be solved ([11], [10], [22]), it is possible to use the presence of mirrors to our advantage to gain even more useful information about the scene that might not be visible from the viewpoint of the camera. As of now, the existing opportunity has not been utilized, though this could potentially be done utilizing the predicted mirror normal vectors and mirror masks we are already generating. To generalize this even further, integrating the capability of mirror surface reconstruction [6], [19], [7] after mirror has been segmented in the scene would add more value to this work and could be potentially impactful future work.

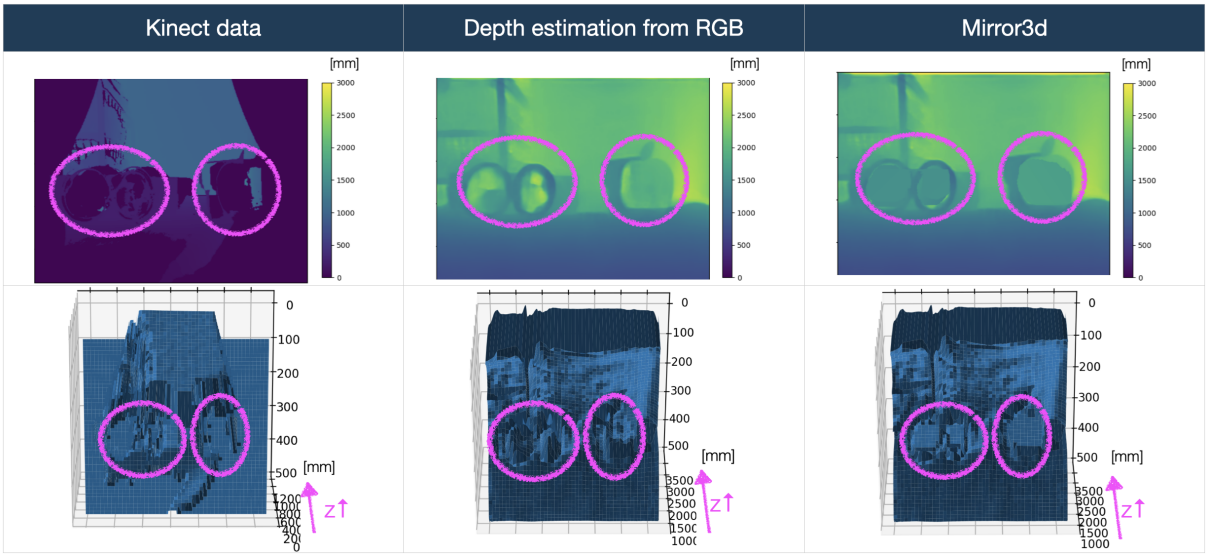

Figure 14. The various depth extraction methods from Figure 13(a), looking at two portable mirrored objects on the desk. The left column demonstrates raw depth measurement data from Kinect sensor and its 3D plot. We note that the depth of mirror is perceived to be zero, right in front of the camera. In the middle column, we show the depth estimation results using just the model in Section 3.2.4, without the model detection method. As shown from the results, this time, the mirrors are perceived as an open space, and the depth measured is even longer than the object behind it. In the right column we show the result with Mirror3DNet. Now the depth of the mirror is accurately captured, and the mirror surface is perceived to be closer than the object behind it. Therefore, Mirror3DNet successfully captures the real depth profile of the scene.

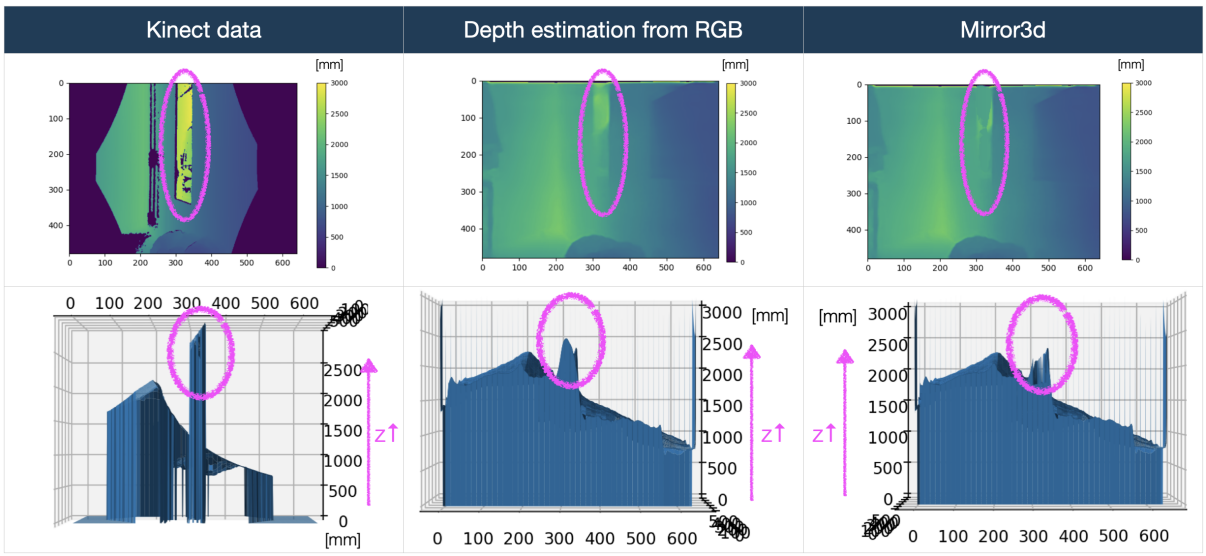

Figure 15. The various depth extraction methods from Figure 13(b), looking at a mirror hanging on the wall. The left column demonstrates raw depth measurement data from Kinect sensor and its 3D plot. We note that the depth of mirror is perceived to be large thus it is depicted as some opening or window on the wall. We note that this measurement is in contrast with the same depth result shown in Figure 14. This indicates that false depth detection of IR sensors on mirrors can happen both ways, and therefore from the depth data alone, it's hard to predict mirror surface. In the middle column, we show the depth estimation results using just the model in Section 3.2.4, without the model detection method. The mirrors are still perceived as an open space, and the depth measured is even longer than the object behind it. In the right column we show the result with Mirror3DNet. While the edge of the mirror still has deeper depth than the neighboring walls, the center of the mirror is accurately captured, and the mirror surface is now more aligned to the wall. Therefore, Mirror3DNet successfully captures the real depth profile of the scene.

References

[1] 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 (CVPR)*, pages 4009–4018, June 2021. 2

[2] R. Chabra, J. Straub, C. Sweeney, R. A. Newcombe, and H. Fuchs. Stereodrnet: Dilated residual stereo net. *CoRR*, abs/1904.02251, 2019. 2

[3] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[4] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2

[5] 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 (CVPR)*, June 2018. 2

[6] H. Han, S. Wu, Z. Song, and J. Zhao. An accurate phase measuring deflectometry method for 3d reconstruction of mirror-like specular surface. In *2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)*, pages 20–24, Los Alamitos, CA, USA, mar 2019. IEEE Computer Society. 7

[7] K. Han, M. Liu, D. Schnieders, and K.-Y. K. Wong. Fixed viewpoint mirror surface reconstruction under an uncalibrated camera. *IEEE Transactions on Image Processing*, 30:2141–2154, 2021. 7

[8] W. Hartmann, S. Galliani, M. Havlena, K. Schindler, and L. V. Gool. Learned multi-patch similarity. *CoRR*, abs/1703.08836, 2017. 2

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[10] H. H. Huynh, T. N. Nguyen, and J. Meunier. Matching-based depth camera and mirrors for 3d reconstruction. *Three-Dimensional Imaging, Visualization, and Display 2018*, May 2018. 7

[11] M. Kanbara, N. Ukita, M. Kidode, and N. Yokoya. 3d scene reconstruction from reflection images in a spherical mirror. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 874–879, 2006. 7

[12] D. Kim, W. Ga, P. Ahn, D. Joo, S. Chun, and J. Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. 2022. 5, 6

[13] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2, 4, 6

[14] J. Lin, G. Wang, and R. W. Lau. Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[15] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019. 6

[16] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 4

[17] R. Nevatia. Depth measurement by motion stereo. *Computer Graphics and Image Processing*, 5:203–214, 1976. 2

[18] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2, 5, 6

[19] S. Savarese and P. Perona. Local analysis for 3d reconstruction of specular surfaces - part ii. In *Proceedings of the 7th European Conference on Computer Vision-Part II*, ECCV '02, page 759–774, Berlin, Heidelberg, 2002. Springer-Verlag. 7

[20] A. Saxena, J. Schulte, A. Y. Ng, et al. Depth estimation using monocular and stereo cues. In *IJCAI*, volume 7, pages 2197–2203, 2007. 2

[21] J. Tan, W. Lin, A. X. Chang, and M. Savva. Mirror3D: Depth refinement for mirror surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 6

[22] T. Whelan, M. Goesele, S. J. Lovegrove, J. Straub, S. Green, R. Szeliski, S. Butterfield, S. Verma, and R. Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM Trans. Graph.*, 37(4), jul 2018. 2, 7

[23] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin, and R. W. Lau. Where is my mirror? In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 6

[24] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[25] W. Yin, Y. Liu, C. Shen, and Y. Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019. 2, 4, 6