Mirror3D: Depth Refinement for Mirror Surfaces

Jiaqi Tan Weijie Lin Angel X. Chang Manolis Savva
Simon Fraser University
https://3dlg-hcvc.github.io/mirror3d/

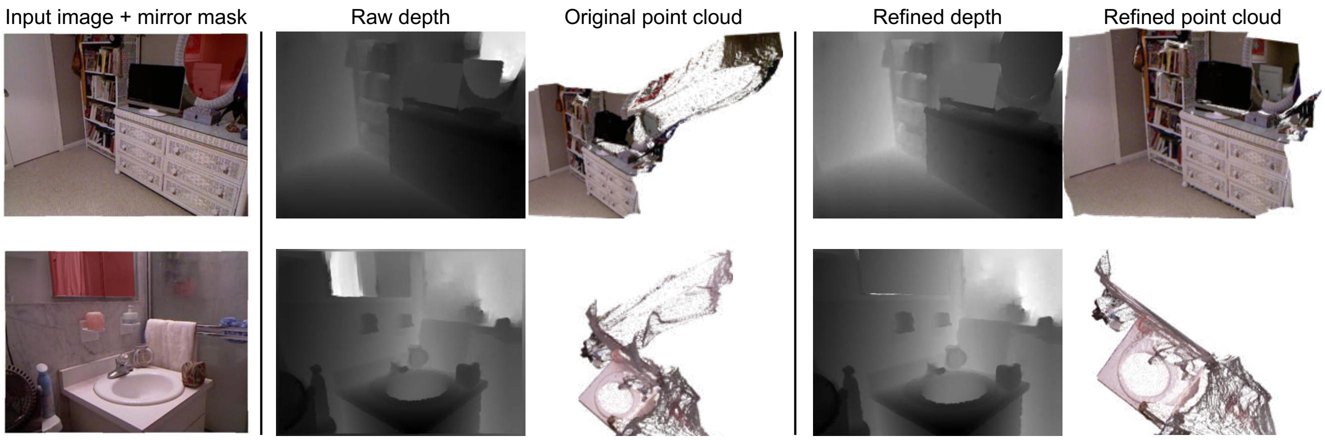

Figure 1: We present the task of 3D mirror plane prediction and depth refinement. First, we annotate several popular RGBD datasets (Matterport3D [6], ScanNet [7], NYUv2 [32]) with 3D mirror planes. Our benchmarks show that both existing RGBD dataset 'ground truth' raw depth data, and state-of-the-art depth estimation and depth completion methods exhibit dramatic errors on mirror surfaces. We propose an architecture for 3D mirror plane estimation that refines depth estimates and produces more reliable reconstructions (compare left and right depth and point cloud pairs from NYUv2 [32] dataset).

Abstract

Despite recent progress in depth sensing and 3D reconstruction, mirror surfaces are a significant source of errors. To address this problem, we create the Mirror3D dataset: a 3D mirror plane dataset based on three RGBD datasets (Matterpot3D, NYUv2 and ScanNet) containing 7,011 *mirror instance masks and 3D planes. We then develop Mirror3DNet: a module that refines raw sensor depth or estimated depth to correct errors on mirror surfaces. Our key idea is to estimate the 3D mirror plane based on RGB input and surrounding depth context, and use this estimate to directly regress mirror surface depth. Our experiments show that Mirror3DNet significantly mitigates errors from a variety of input depth data, including raw sensor depth and depth estimation or completion methods.*

1. Introduction

Recent years have seen much progress in 3D reconstruction methods. It is now possible to acquire 3D reconstructions of interiors with high quality geometry and texture. However, these reconstructions fail spectacularly when used in environments with mirrors and glass, both of which are prevalent in indoor spaces.

The fragility of many reconstruction methods is due to a reliance on accurate active depth sensing. Commodity sensors such as the Microsoft Kinect and Intel RealSense employ active infrared or time of flight depth sensing which requires strong signal return from sensed surfaces. Unfortunately, highly glossy and reflective surfaces such as mirrors cause either no signal return or highly unreliable depth estimates. Though it may seem that reflectors are a 'corner case' with few affected pixels, the resulting errors are catastrophic to reconstruction algorithms, leading to loss of camera pose tracking and geometry artifacts (see Figure 1).

Recently, Whelan et al. [34] demonstrated that by detecting mirror planes it is possible to mitigate the above issues and achieve high-quality reconstructions in scenes with many reflectors. However, their method relies on the use of an AprilTag [27] attached to a custom camera rig. The use of such custom hardware is not always feasible. In this paper, we propose a method for identifying mirrors and estimating mirror surface depth on RGBD data collected with commodity hardware.

Our key idea is to identify mirror regions based on color information (RGB), model the mirror as a plane and use an estimated mirror normal and information from the mirror's surroundings to predict the mirror's position in 3D. Our work is related to the recently proposed task of mirror segmentation in 2D [21, 38]. However, we operate in 3D and focus on using 3D mirror plane estimates to improve the reliability of depth data.

To estimate the prevalence of mirrors in 3D environments and understand the severity of the above reconstruction issues, we first annotate 3D mirror planes in three popular RGBD datasets: Matterport3D [6], ScanNet [7], and NYUv2 [32]. These annotations create 'true' ground truth for observed mirror surfaces that was previously unavailable. We find the prominence of mirrors varies between datasets depending on their acquisition procedure, leading to a corresponding range of depth data issues and reconstruction failures. Using this data, we introduce the task of 3D mirror detection from RGB and RGBD data, and establish initial benchmark results by: i) applying state-of-the-art depth estimation and depth completion approaches to directly estimate mirror depth values; and ii) propose a simple architecture combining a MaskRCNN [14] module and a PlaneRCNN [23] module to segment mirror surfaces, estimate mirror 3D planes and refine mirror surface depth estimates. Overall, we make the following contributions:

- Introduce the task of 3D mirror plane prediction from single-view RGB and RGBD data
- Provide 3D mirror plane annotations for three RGBD datasets (Matterport3D, NYUv2 and ScanNet)
- Establish benchmarks for RGB and RGBD-based 3D mirror plane prediction, and evaluate depth completion and depth estimation approaches on the task
- Present Mirror3DNet, an architecture that predicts a 3D mirror normal and mirror segmentation to refine raw sensor depth or the output of state-of-the-art depth completion and estimation methods

2. Related work

We summarize related work on mirror detection, 3D plane estimation, and depth estimation and completion.

Mirror detection and correction. The challenges of dealing with reflective and transparent surfaces in 3D reconstruction and robotics have long been recognized. Yang and Wang [37] proposed a sensor fusion technique for dealing with LiDAR sensor failures on mirror and glass surfaces. Käshammer and Nüchter [16] detect mirrors and correct laser-scanned point clouds based on heuristics using known mirror dimensions. The general problem of mirror surface reconstruction is addressed by work on reflectometry. Early work focuses on the stereo image setting [2], active illumination hardware setups [3], or single image input but relying on detecting reflection correspondences of reference target objects [24]. In recent years, there has been renewed interest in identifying mirrors, glass, and transparent objects for improved reconstruction [34] and for manipulation in robotics [33]. Whelan et al. [34] rely on observing the reflection of an AprilTag [27] attached to a custom scanner. Work on mirror detection for uncontrolled RGB images using deep learning is less explored, with a few recent works. Yang et al. [38] introduce a network and dataset for identifying mirrors in 2D images, and Lin et al. [21] extend the earlier work by extracting mirror context features to improve mirror detection. In contrast, we study 3D mirror plane estimation in RGBD datasets without relying on custom hardware or other assumptions on the capture setup.

3D plane detection and plane reconstruction. There is recent work on detecting 3D planes from single-view images using neural networks [22, 23, 36, 40]. Unlike this work, our focus is not on creating a planar segmentation of the entire observed scene. We focus specifically on identifying mirror regions and corresponding 3D planes. Identifying mirrors in RGB images is challenging, as mirrors contain a reflection of the environment which can be difficult to distinguish from non-reflected regions. Mirror detection remains challenging even with RGBD data as mirror depth values tend to be noisy and unreliable.

Depth estimation and depth completion. Depth estimation and completion are recently popular tasks. Typically, the term depth estimation is used when the input is only color (RGB) and has no depth information. The term depth completion is used when the input is RGBD, where the D (depth) channel is noisy and may have missing values. Existing methods for single-view depth estimation [1, 4, 9, 10, 18–20, 25, 30, 39] and depth completion [15, 26, 28, 31, 41] improve depth prediction for the entire image, relying on reconstructed 3D mesh data that is assumed to provide accurate depth. Chabra et al. [5] show that an exclusion mask for noisy areas such as reflective surfaces can result in better reconstruction. We leverage 3D mirror plane estimates to improve the accuracy of existing depth estimation and completion methods. Moreover, we note that both depth estimation and completion methods are typically evaluated on ground truth data that does not account for noise due to mirrors and glass. Widely employed datasets such as NYUv2[32], Matterport3D[6, 41], ScanNet [7], and SUN3D [12, 35] have noisy or missing depth for reflectors (see Figure 1). Thus, these regions are typically ignored or evaluated with incorrect values as the ground truth. We contribute ground truth 3D mirror annotations for three RGBD datasets, allowing for correct benchmarking of mirror surface depth estimation and completion.

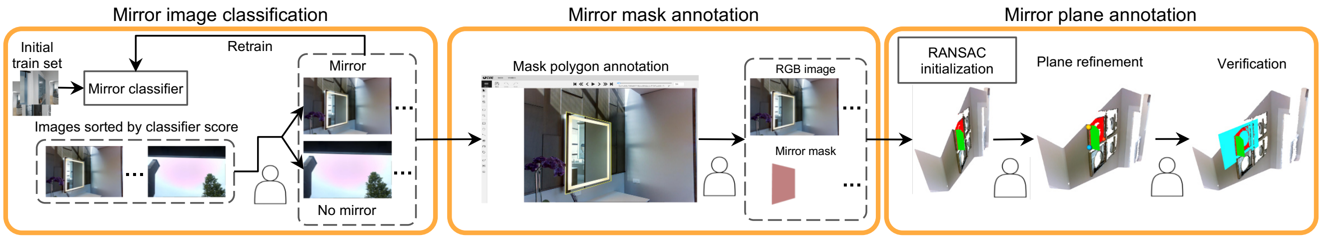

Figure 2: Our human-in-the-loop mirror mask and 3D plane annotation workflow. We leverage an iteratively trained mirror image classifier to assist a user in rapidly selecting all mirror images in an input image dataset. The images are then annotated with mirror surface polygon masks. The masks and their corresponding depth images are used to initialize mirror plane estimates with a RANSAC approach. The user then refines and verifies all 3D mirror plane estimates.

3. Mirror3D dataset

To enable benchmarking of 3D mirror plane prediction and surface depth estimation, we create Mirror3D: the first large-scale dataset of mirror annotations for RGBD images. We design a human-in-the-loop workflow that enables efficient iterative annotation of mirror masks and mirror 3D planes. Using this workflow, we annotate RGBD images containing mirrors in three common RGBD datasets (NYUv2 [32], Matterport3D [6], and ScanNet [7]) to create an aggregated dataset that contains 7,011 annotated 3D mirror planes in 5,894 RGBD frames.

3.1. Dataset construction

Our annotation workflow consists of three stages: i) mirror image classification, ii) mirror mask annotation, and iii) mirror plane annotation. Figure 2 shows the overall pipeline. We first pretrain a ResNet-50 classifier on a seed dataset of 'mirror present' and 'no mirror present' RGB images from Structured3D [42] (roughly half of the images contained mirrors, approximately 7,000 images total). We then sort all input RGBD dataset candidate frames by using the mirror classification score. An annotator confirms whether images contain a mirror, splitting them into 'mirror' and 'no mirror' sets in batches of 100 images. As this verification proceeds through each batch, the classifier is finetuned on the dataset of newly annotated 'mirror' and 'no mirror' images to improve the efficacy of the classification score sorting whenever less than 20 additional mirror images are added from a single 100-image batch. The annotator was instructed to stop looking for additional mirror images when less than 5% of a batch contains mirrors. This first stage was performed by one annotator and took approximately 28 hours in total over 8 batch iterations.

In the second stage, we used the CVAT[1] annotation interface to define mirror mask polygons for all mirror images. The annotators specified two types of mirror mask polygons: a 'coarse' mask and a 'detailed' mask. The coarse mask ignores occluding objects and small mirror protrusions, while the detailed mask is a pixel-accurate boundary of the visible mirror surface in each image. This stage was performed by four annotators over approximately 67 hours.

[1] github.com/openvinotoolkit/cvat

	Matterport3D	ScanNet	NYUv2	**Total**
mirror planes	4,662	2,218	131	7,011
RGBD images	3,782	1,987	125	5,894
RGBD panoramas	2,468	–	–	2,468
3D scenes	79	282	96	457

Table 1: Summary statistics for our Mirror3D dataset.

The last stage used a 3D interface developed with Open3D [43] that allows inspection of an initial 3D mirror plane normal estimate obtained by filtering mirror mask border depth values using RANSAC [11]. The annotators could check that the plane is correct, refine the 3D plane estimate by specifying three plane points on the point cloud and manually adjusting, or indicate that it is impossible to define a plane (in cases where the point cloud is extremely noisy). After the 3D planes are annotated, we generated turntable videos of the point clouds with the annotated planes from a frontal and top-down view to allow for quick verification. Planes identified as incorrect were further adjusted and corrected. This last stage took a total of approximately 50 hours of annotation effort by the same expert annotator who carried out the first stage.

In total, the annotators inspected approximately 30,000 RGBD images to obtain the final 7,011 3D mirror plane annotations. The mirror image classification stage took 3s per image on average. The mirror mask annotation stage took 25s and 57s per mirror instance for coarse mask and detailed mask respectively. Finally, the mirror plane annotation stage takes 20s per mirror instance on average.

3.2. Dataset statistics

At the end of our annotation, we end up with a total of 7,011 mirror 3D plane and instance mask annotations across

RGBD images from all three source datasets. Table 1 shows a more detailed breakdown of dataset statistics. We split this data following the standard training, validation and test splits of each source dataset. We note that mirrors are found in roughy 22.9% of Matterport3D panoramas (from a total of 79 out of 90 Matterport3D residences) and 96 of 464 (20%) NYUv2 scenes, which are relatively high fractions compared to only 282 out of approximately 1,500 ScanNet scenes (18.8%). We hypothesize that this is due to the capture setup and methodology employed by the ScanNet authors who may have avoided scanning scene regions that contain mirrors. The supplement provides some additional analysis of the annotated data in terms of mirror region location, prominence and distribution across the images.

4. Mirror3DNet: refining mirror depth

Having constructed our dataset of 3D mirror plane annotations, we would now like to address the task of predicting 3D mirrors from a single-view RGB or RGB-D image, and using these predictions to improve estimated depth values on mirror surfaces. Because mirrors are often planar, we make the simplifying assumption that we can model a mirror using a mirror mask and a mirror plane. We will show that we can improve depth estimation from RGB-only image input or depth completion results from an RGBD image using the predicted 3D region and plane. Thus, the input to our approach is either an RGB image or an RGBD image with missing or incorrect depth for mirrors. The output is an estimated 3D mirror plane and an RGBD image with refined depth for the mirror.

Our goal is not to propose a new approach for depth estimation or depth completion, but rather to benchmark state-of-the-art approaches for both and demonstrate that we can leverage a simple mirror-aware architecture to improve depth output for both families of approaches. To do this, we propose the Mirror3DNet architecture (see Figure 3). This architecture uses a Mask R-CNN [14] module and a 3D mirror plane estimation module inspired by PlaneRCNN [23] to predict mirror masks and planes respectively. A depth estimation or completion module may be used first to predict depth values given an input RGB or RGBD image, and then Mirror3DNet is applied to predict the mirror mask and mirror plane. We evaluate this architecture's performance on depth prediction as well as mirror mask prediction and mirror normal estimation.

Thus, the overall architecture consists of three modules: mirror mask segmentation, mirror plane estimation, and depth estimation or completion. PlaneRCNN uses a warping loss to enforce consistency of reconstructed 3D planes from nearby viewpoints. Our problem setting only assumes a single image input, so we replace the warping loss module with a mirror plane estimation module which refines the depth on the mirror region.

4.1. Mirror mask and plane prediction

In this part of the architecture, we address detection and segmentation of the mirror surface. To do this, we employ a mirror classification (MC) branch and a mirror bounding box regression (MR) branch, as well as an instance segmentation (Seg) branch, all based on the Mask R-CNN module. We also add an anchor normal classification (AC) branch and an anchor normal regression (AR) branch after the ROI pooling layer to predict the mirror plane normal.

Since it is challenging to directly regress the mirror normal values, these latter two branches follow an approach similar to PlaneRCNN and decompose the regression into a classification phase and a residual prediction phase. In the classification phase we obtain a coarse orientation of the mirror normal by classifying into one of a few anchor normal orientations using the AC branch. We use 10 mirror anchor normals from a k-means clustering of all mirror normals in the annotated Matterport3D training set. During training, the module predicts one anchor normal for each positive proposal. For supervision, we assign each mirror instance to its closest mirror anchor normal.

Given the anchor normal classification we then regress the residual using the AR branch to form the final normal vector. The final normal n is the sum of the anchor normal and the normal residual. The ground truth mirror residual of an instance is the distance vector between ground truth mirror and the ground truth mirror anchor normal.

4.2. Depth estimation and completion

As Figure 3 shows, our network architecture accommodates a depth estimation or completion module. The details of this part depend on the input type. That is, whether we take an RGB image or RGBD data including noisy depth. If the input is an RGB image, we first use a depth estimation module to produce an initial depth map estimate. Then the rest of the Mirror3D architecture refines the depth map into D_{pred}. If the input is an RGBD image with a noisy depth map D_{noisy}, we directly carry it forward for refinement through the architecture.

To estimate the 3D mirror plane we need to combine the mirror mask and mirror normal estimates from the previous stage with a depth offset d for the plane. Since depth values on the mirror surface itself are missing or unreliable, we rely on mirror border regions which are often non-reflective materials and significantly more reliable. Thus, we compute the average depth of points that fall on pixels within a small offset of the mirror mask border. Given the predicted mirror segmentation mask, we create a mirror border mask region by expanding outwards by 50 pixels. This threshold corresponds to roughly 9% of average image dimensions and we have empirically found it to be reasonable. We then take the average depth of all points in this mirror mask as the offset of the mirror plane d.

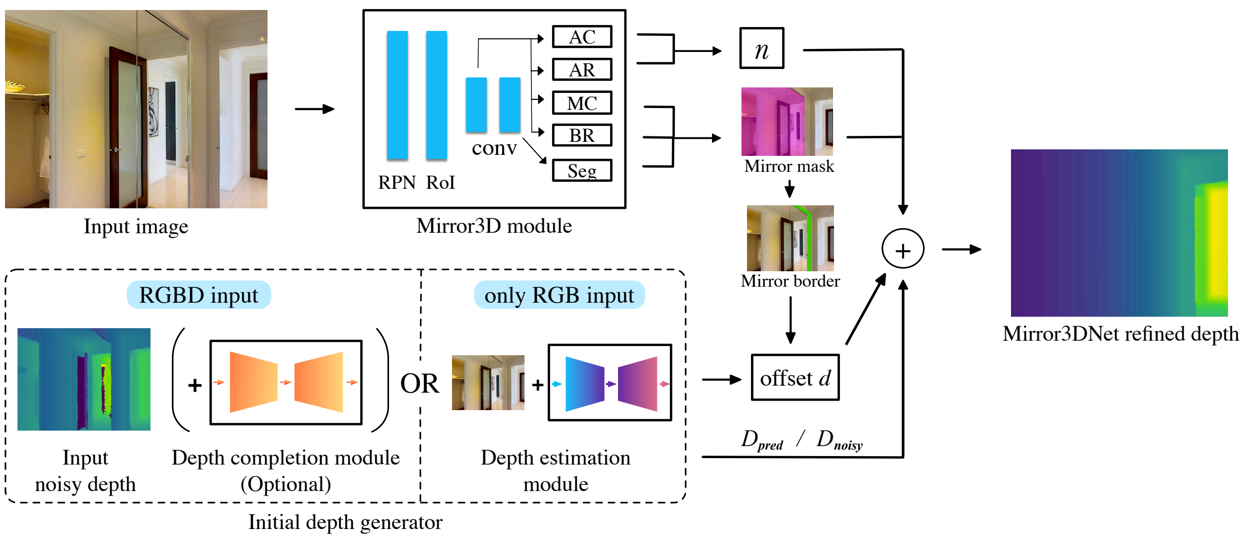

Figure 3: Overall diagram showing how the Mirror3DNet architecture can be used for either an RGB image or an RGBD image input. For an RGB input, we refine the depth of the predicted depth map D_{pred} output by a depth estimation module. For RGBD input, we refine a noisy input depth D_{noisy}. The Mirror3DNet module predicts a mirror normal n and mirror mask. Then, by computing an offset depth d based on depth values at the mirror border, we can determine the position and orientation of the 3D mirror plane, and produce refined output depth values that improve mirror surface depth accuracy.

The Mirror3DNet module is trained under three types of loss terms: i) mirror segmentation loss, ii) mirror anchor normal classification loss, and iii) mirror anchor normal regression loss. In mirror segmentation, we use a cross-entropy loss for mirror classification and mirror mask prediction. We use a smooth L1 loss for mirror bounding box regression, and a cross-entropy loss for mirror anchor normal classification. For mirror anchor normal regression, we use a smooth L1 loss. The total loss $\mathcal{L}$ is then:

$$\begin{aligned}\mathcal{L} &= \text{CE}(a_i, a_i^*) + \text{Smooth}_{L1}(r_i, r_i^*) \\ &+ \text{CE}(c_i, c_i^*) + \text{Smooth}_{L1}(b_i, b_i^*) + \text{CE}(m_i, m_i^*)\end{aligned}$$

where a_i and a_i^* are the predicted and ground truth mirror anchor normal class, r_i and r_i^* are the 1×3 predicted mirror regression vector and ground truth mirror regression vector, c_i and c_i^* are the predicted mirror proposal class and ground truth proposal class, b_i and b_i^* are the predicted and ground truth mirror bounding box parameter, m_i and m_i^* are the predicted and ground truth mirror mask.

4.3. Implementation details

We implement our network architecture in PyTorch [29]. We used the Adam [17] optimizer with initial learning rate set to 10^{-4}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and without weight decay. The RestNet-50 [13] backbone was initialized with weights pretrained on ImageNet [8]. We train our model for 50,000 iteration with batch size 32. We only supervise on pixels that have a ground truth depth of less than 10 meters.

5. Experiments

5.1. Datasets

We carry out our quantitative and qualitative evaluation on the NYUv2 and Matterport3D datasets. To show the importance of accurate mirror depth we evaluate against both original 'raw depth' ground truth depth and our annotated mirror depth. Therefore, we employ the following sets of depth data in our experiments:

NYUv2-raw: the 1449 annotated RGBD frames from NYUv2 (795 frames in train and 654 frames in test), captured with a Kinect depth sensor and exhibiting noise and missing data on mirror surfaces.

NYUv2-ref: the corrected version of NYUv2-raw, with mirror surface depth computed using our 3D mirror plane annotations.

MP3D-mesh: depth rendered from the Matterport3D reconstructed mesh, as used by Zhang and Funkhouser [41].

MP3D-mesh-ref: the corrected version of the above using our 3D mirror plane annotations.

5.2. Evaluation metrics

Here, we define the metrics that we use to measure mirror mask and plane prediction, and depth estimation and completion accuracy.

Depth estimation and completion. We follow the evaluation protocol of Zhang and Funkhouser [41], and employ root mean squared error (RMSE), relative error (Rel), and δ_i. The δ_i metric denotes the percentage of predicted pixels where the relative error is less than a threshold i. Specifically, i is chosen to be equal to $1.05, 1.10, 1.25, 1.25^2$ and 1.25^3. Here, the larger is i, the more sensitive the δ_i metric. Larger values of δ_i reflect a more accurate prediction. Since depth estimation methods are often limited by the size-distance ambiguity in predicting depth values, we also calculate a scale-invariant root mean squared error (s-RMSE) as introduced by Eigen et al. [9]. The main paper reports a subset of these metrics, with more complete results found in the supplement.

RMSE: $\sqrt{1/|P| \sum_{p \in P} \|D^*(p) - D(p)\|^2}$ where $D(p)$ and $D^*(p)$ is depth and ground truth depth at point p.

s-RMSE: $\sqrt{1/|P| \sum_{p \in P} \|D^*(p) - sD(p)\|^2}$ where s is the least square root of $1/n \sum (D^* - sD)^2$.

Rel: $1/|P| \sum_{p \in P} |D^*(p) - D(p)| / D^*(p)$.

SSIM: the structural similarity index measure, $\frac{(2\mu_{D^*(p)}\mu_{D(p)} + c_1)(2\sigma_{D^*(p)D(p)} + c_2)}{(\mu^2_{D^*(p)} + \mu^2_{D(p)} + c_1)(\sigma^2_{D^*(p)} + \sigma^2_{D(p)} + c_2)}$ where we have set $c_1 = 0.0001, c_2 = 0.0009$.

$\boldsymbol{\delta_i}$: percentage of pixels within error range i, where the error range is defined by $\max\left(D^*(p)/D(p), D(p)/D^*(p)\right) < i$.

All the above metrics are reported separately for depth points within the ground truth mirror region, points outside the mirror, and together for all depth points in each frame.

Mirror mask prediction and plane estimation. We adopt the commonly used mean average precision (mAP) to evaluate mirror mask segmentation predictions, and a number of mirror normal prediction error metrics.

Seg-AP: segmentation AP, with IoU threshold starting from 0.50 to 0.95 at 0.05 steps. The final AP score is the average over the 10 threshold steps.

AC-AP: the anchor normal classification AP, with same IoU thresholds as Seg-AP. We use the anchor normal classification score to filter positive samples at inference time.

AngErr: the angle between predicted mirror normal and ground truth mirror normal: $\arccos(m^* \cdot m / |m^*||m|)$.

AR-L2: the mirror normal L2 error, measuring anchor normal (AR) regression performance. Calculated from the L2 distance between predicted mirror normal and ground truth mirror normal: $\sqrt{(m^*/|m^*| - m/|m|)^2}$

5.3. Qualitative evaluation

Figure 4 shows two qualitative comparison examples. The top set of visualizations shows the improvements on NYUv2-raw and NYUv2-ref data using our Mirror3DNet module to enhance results from several depth estimation and depth completion approaches. We note that outlier depth points behind the mirror surface are significantly reduced after using Mirror3DNet (see depth RMSE images and top-down point cloud visualizations). The improvements are consistent across input depth approach, whether RGB-based depth estimation is employed or RGBD-based depth completion is employed. The bottom set of visualizations shows comparisons on the MP3D-mesh and MP3D-mesh-ref datasets. This data exhibits fewer depth outliers due to the input being rendered from a 3D reconstructed mesh that incorporates manual mirror region correction. Despite this, the Mirror3DNet module still improves depth accuracy on the mirror surface, in particular for RGB-based depth estimation approaches.

	Seg-AP ↑	AC-AP ↑	AR-L2 ↓	AngErr ↓
Mirror3DNet (ours)	**0.271**	0.072	**0.159**	**9.10**
PlaneRCNN [23]	0.208	**0.076**	0.194	11.2

Table 2: Evaluation of mirror mask segmentation and mirror normal prediction. The PlaneRCNN [23] baseline is trained on the NYUv2-ref dataset. We observe that our Mirror3DNet module improves on mask segmentation, and normal estimation metrics.

			RMSE ↓			SSIM ↑		
Input	Train	Method	Mirror	Other	All	Mirror	Other	All
RGBD	*	Mirror3DNet	0.891	0.077	0.309	0.721	0.984	0.946
RGBD	ref	saic [31]	1.081	**0.074**	0.391	0.669	**0.928**	0.884
RGBD	raw	saic [31]	1.170	0.077	0.417	0.658	0.926	0.882
RGBD	raw	saic [31] + Mirror3DNet	**0.874**	0.095	**0.314**	**0.718**	0.922	**0.888**
RGB	ref	BTS [19]	**0.472**	0.351	**0.391**	0.825	0.832	0.821
RGB	ref	VNL [39]	6.169	5.804	5.882	0.228	0.203	0.204
RGB	raw	BTS [19]	0.971	**0.315**	0.547	0.691	**0.856**	0.819
RGB	raw	VNL [39]	3.939	2.265	2.725	0.384	0.629	0.583
RGB	raw	BTS [19] + Mirror3DNet	0.801	0.317	0.481	**0.753**	**0.856**	**0.827**
RGB	raw	VNL [39] + Mirror3DNet	3.462	2.262	2.554	0.444	0.628	0.593

Table 3: Depth prediction evaluation on NYUv2-ref dataset, for images containing mirrors.

5.4. Quantitative evaluation

Mirror 3D plane prediction. We first quantify the improvements introduced by our Mirror3DNet architecture over a baseline PlaneRCNN module for predicting the mirror 3D plane. Table 2 reports overall quantitative metrics for the NYUv2-ref dataset. We see that the Mirror3DNet module leads to improved mirror mask segmentation and plane normal estimates. Please refer to the supplement for additional results reporting a set of ablation experiments that we use to choose the mirror anchor normal count and the mirror border width.

Mirror depth refinement. We conducted a series of experiments to quantify the improvements in depth value prediction offered by our Mirror3DNet architecture, and summarize overall trends here. Please refer to the supplement for additional results and metrics. We used three different datasets as the ground truth for the purposes of the evalu-

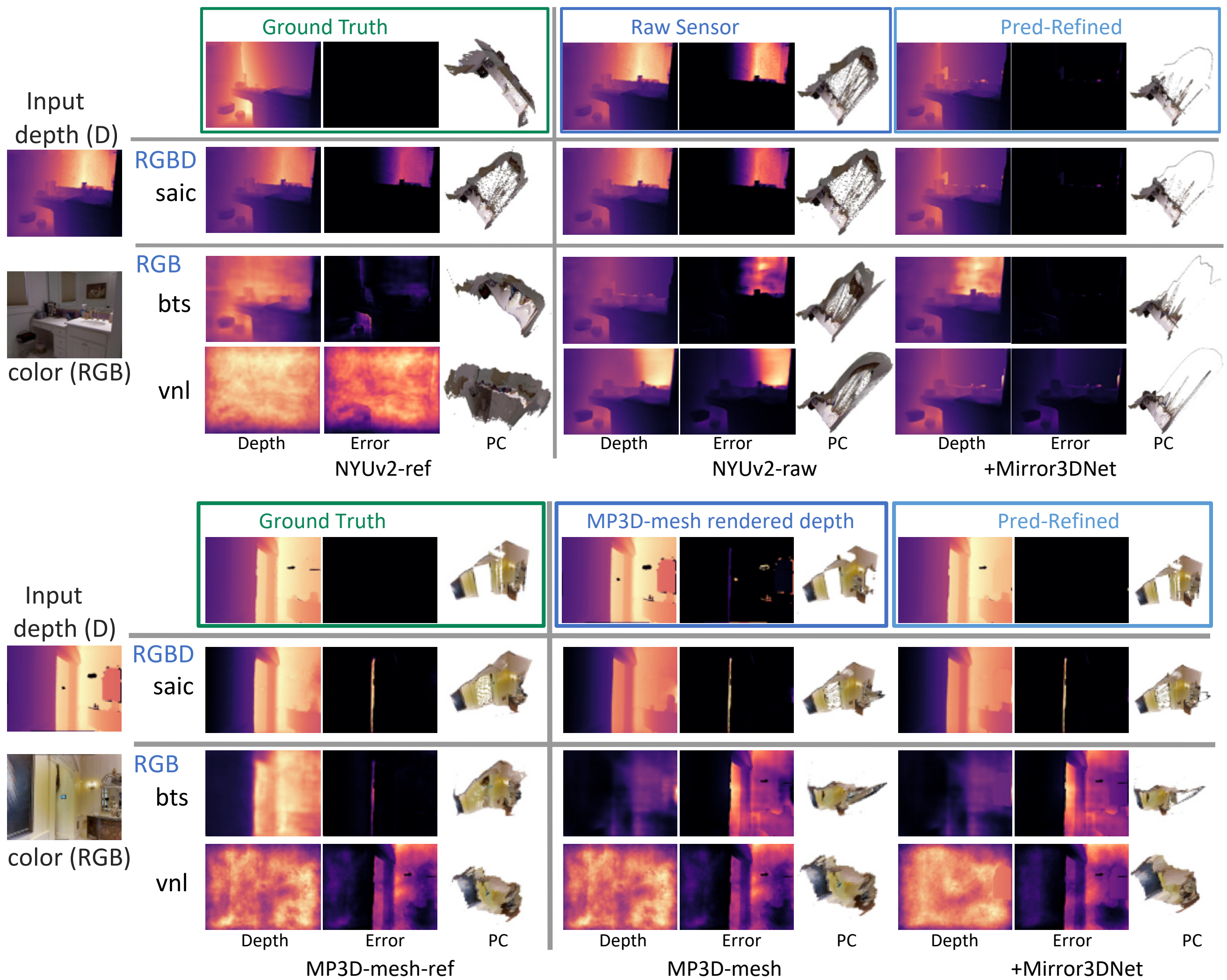

Figure 4: Visualizations of depth frames, depth errors against ground truth (RMSE mapped to colormap), and resulting 3D point clouds (PC) for NYUv2 (top) and Matterport3D (bottom). We compare the output depth from state-of-the-art RGB-based depth estimation approaches (bts [19], vnl [39]) and an RGB-based depth completion approach (saic [31]). We contrast the outputs from these approaches when trained directly on the corrected datasets which leverage our 3D mirror plane annotations (NYUv2-ref and MP3D-mesh-ref), against output of the approaches when trained on the uncorrected original datasets (NYUv2-raw and MP3D-mesh-ref), and against outputs after refinement using our Mirror3DNet module. Overall, we observe that mirror depth errors are significantly reduced, as seen by the reduced RMSE error in mirror regions, and the reduced prominence of depth outlier points.

ation reported here: NYUv2-ref, NYUv2-raw, and MP3D-mesh-ref. Results with the first dataset (NYUv2-ref) treated as the ground truth are in Table 3. These results show how far from the 'correct ground truth' the various depth estimation and prediction methods are, and how much of an improvement Mirror3DNet can provide. We note that methods trained on the corrected NYUv2-ref depth have significantly better performance on mirror area depth prediction. The Mirror3DNet module also helps improve depth accuracy for mirrors significantly when applied to methods trained on NYUv2-raw, which would be the practical input at capture time.

We can contrast the above set of results against the results reported in Table 4, where the ground truth is assumed to be the original, uncorrected NYUv2-raw dataset. Here, we can make the observation that evaluating on raw depth as ground truth gives a different (and inaccurate) ranking of methods compared to the corrected ground truth in NYUv2-

Input	Train	Method	RMSE ↓			SSIM ↑		
			Mirror	Other	All	Mirror	Other	All
RGBD	*	Mirror3DNet	0.414	0.036	0.179	0.885	0.995	0.974
RGBD	ref	saic [31]	0.201	**0.042**	0.085	0.846	**0.935**	0.919
RGBD	raw	saic [31]	**0.102**	**0.042**	**0.054**	**0.862**	0.933	**0.920**
RGBD	raw	saic [31] + Mirror3DNet	0.484	0.070	0.216	0.788	0.929	0.901
RGB	ref	BTS [19]	1.024	0.355	0.538	**0.644**	0.833	0.796
RGB	ref	VNL [39]	5.238	5.798	5.741	0.264	0.204	0.209
RGB	raw	BTS [19]	**0.962**	**0.316**	**0.452**	0.621	**0.858**	**0.821**
RGB	raw	VNL [39]	3.113	2.254	2.491	0.433	0.633	0.599
RGB	raw	BTS [19] + Mirror3DNet	1.098	0.318	0.510	0.624	0.857	0.817
RGB	raw	VNL [39] + Mirror3DNet	2.902	2.252	2.716	0.438	0.631	0.598

Table 4: Depth prediction evaluation on NYUv2-raw dataset, for images containing mirrors.

Input	Train	Method	RMSE ↓			SSIM ↑		
			Mirror	Other	All	Mirror	Other	All
sensor-D	*	*	2.605	0.901	1.268	0.215	0.762	0.669
mesh-D	*	*	0.631	0.000	0.177	0.794	1.000	0.970
RGBD (sensor-D)	*	Mirror3DNet	1.542	0.897	1.136	0.586	0.798	0.744
RGBD (mesh-D)	*	Mirror3DNet	0.428	0.016	0.150	0.881	0.998	0.978
RGBD	mesh-ref	saic[31]	**0.308**	**0.316**	**0.358**	**0.861**	0.899	**0.899**
RGBD	mesh	saic[31]	0.984	0.320	0.595	0.692	**0.909**	0.864
RGBD	mesh	Mirror3DNet + saic [31]	0.786	0.321	0.553	0.786	0.908	0.870
RGB	mesh-ref	BTS[19]	**0.572**	**0.634**	**0.658**	**0.788**	**0.776**	**0.769**
RGB	mesh-ref	VNL[39]	1.364	1.410	1.408	0.620	0.630	0.623
RGB	mesh	BTS[19]	1.142	1.033	1.097	0.669	0.757	0.733
RGB	mesh	VNL[39]	1.400	1.432	1.429	0.456	0.440	0.421
RGB	mesh	BTS[19] + Mirror3DNet	1.156	1.034	1.092	0.746	0.757	0.739
RGB	mesh	VNL[39] + Mirror3DNet	1.390	1.424	1.423	0.612	0.475	0.470

Table 5: Depth prediction evaluation on MP3D-mesh-ref dataset, for images containing mirrors.

ref. Not surprisingly, the ranking on NYUv2-raw gives an edge to the methods that were trained on the raw depth. This is however, a misleading result, as the 'assumed to be correct ground truth' is highly inaccurate for mirror regions.

Lastly, in Table 5 we can see results using the MP3D-mesh-ref dataset as the ground truth. The overall trends remain the same, but we note that on this data there are less pronounced differences in performance (i.e. it is still better to train on refined depth). We hypothesize that this is due to mesh cleanup and post-processing relying on human-provided annotation of some reflective and transparent surfaces. In other words, the Matterport reconstruction pipeline that was used to produce the mesh from which MP3D-mesh data is rendered already incorporates a degree of mirror surface correction. While this annotation is not explicitly specified by the Matterport3D dataset, the mesh reconstruction itself makes use of it, and the rendered mesh depth will thus have cleaner depth than without human intervention. Our goal with Mirror3DNet is to automate this depth refinement so that human intervention is not required at capture time or mesh reconstruction post-processing.

5.5. Limitations and future work

Our mirror depth refinement approach is a simple first step but it is subject to several limitations that suggest future work directions. Firstly, we assume planar mirrors and use a fixed border width to estimate the mirror depth offset. Both are assumptions that should can be lifted using more advanced mirror detection approaches. Moreover, in this paper we only focused on using 3D mirror plane estimates to refine depth values at the mirror surface. Observations of reflected objects in mirrors can be leveraged to further improve reconstruction of other surfaces beyond the mirror. The estimated 3D mirror planes can also be used to create more realistic rendered visuals from 3D reconstructions, by simulating reflections and light propagation from the mirror surfaces.

6. Conclusion

In this paper, we tackled the problem of 3D mirror plane prediction and mirror depth refinement. We created Mirror3D: a large-scale dataset of 3D mirror plane annotations based on three popular RGBD datasets which we use to obtain corrected ground truth depth on mirror surfaces. Using this data, we develop Mirror3DNet: a mirror depth refinement architecture that can be used to refine depth estimation or depth completion output. Our experiments show that mirror depth errors in popular RGBD datasets are prevalent, and that treating existing depth data as ground truth can misrepresent depth prediction method performance. Moreover, we show that our Mirror3DNet architecture helps to improve mirror depth estimates from depth estimation and depth completion approaches, significantly mitigating 3D reconstruction artifacts due to mirror surfaces.

Acknowledgements. We thank Hanxiao Jiang, Yongsen Mao, and Yiming Zhang for their help with the annotation. We also thank the anonymous reviewers for their suggestions. This research was enabled in part by support provided by WestGrid and Compute Canada. AXC is supported by a Canada CIFAR AI Chair, and MS by a Canada Research Chair and NSERC Discovery Grant.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.

[2] Jonathan Balzer, Sebastian Hofer, and Jürgen Beyerer. Multiview specular stereo reconstruction of large mirror surfaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2537–2544. IEEE, 2011.

[3] Jonathan Balzer, Daniel Acevedo-Feliz, Stefano Soatto, Sebastian Höfer, Markus Hadwiger, and Jürgen Beyerer. Cavlectometry: Towards holistic reconstruction of large mirror objects. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 448–455. IEEE, 2014.

[4] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 3174–3182, 2017.

[5] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs. StereoDRNet: Dilated residual stereonet. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11786–11795, 2019.

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *Proc. of International Conference on 3D Vision (3DV)*, 2017.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255. Ieee, 2009.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[10] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3):611–625, 2017.

[11] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[12] Maciej Halber and Thomas Funkhouser. Fine-to-coarse global registration of RGB-D scans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, July 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017.

[15] Yu-Kai Huang, Tsung-Han Wu, Yueh-Cheng Liu, and Winston H Hsu. Indoor depth completion with boundary consistency and self-attention. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[16] PF Käshammer and A Nüchter. Mirror identification and correction of 3D point clouds. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(5):109, 2015.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019.

[19] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[20] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single RGB images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017.

[21] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

[22] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. PlaneNet: Piece-wise planar reconstruction from a single RGB image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2579–2588, 2018.

[23] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. PlaneRCNN: 3D plane detection and reconstruction from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4450–4459, 2019.

[24] Miaomiao Liu, Richard Hartley, and Mathieu Salzmann. Mirror surface reconstruction from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 129–136, 2013.

[25] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5667–5675, 2018.

[26] Raul de Queiroz Mendes, Eduardo Godinho Ribeiro, Nicolas dos Santos Rosa, and Valdir Grassi Jr. On deep learning techniques to boost monocular depth estimation for autonomous navigation. *arXiv preprint arXiv:2010.06626*, 2020.

[27] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407. IEEE, May 2011.

[28] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. *arXiv preprint arXiv:2007.10042*, 2020.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[30] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[31] Dmitry Senushkin, Ilia Belikov, and Anton Konushin. Decoder modulation for indoor depth completion. *arXiv preprint arXiv:2005.08607*, 2020.

[32] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Eur. Conf. Comput. Vis.*, pages 746–760. Springer, 2012.

[33] Thomas Weng, Amith Pallankize, Yimin Tang, Oliver Kroemer, and David Held. Multi-modal transfer learning for grasping transparent and specular objects. *IEEE Robotics and Automation Letters*, 5(3):3791–3798, 2020.

[34] Thomas Whelan, Michael Goesele, Steven J Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, and Richard A Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM Trans.*

Graph., 37(4):102–1, 2018.

[35] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3D: A database of big spaces reconstructed using SFM and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013.

[36] Fengting Yang and Zihan Zhou. Recovering 3D planes from a single image via convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 85–100, 2018.

[37] Shao-Wen Yang and Chieh-Chih Wang. Dealing with laser scanner failure: Mirrors and windows. In *2008 IEEE International Conference on Robotics and Automation*, pages 3009–3015. IEEE, 2008.

[38] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *Int. Conf. Comput. Vis.*, pages 8809–8818, 2019.

[39] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Int. Conf. Comput. Vis.*, pages 5684–5693, 2019.

[40] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3D reconstruction via associative embedding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1029–1037, 2019.

[41] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single RGB-D image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 175–185, 2018.

[42] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *Eur. Conf. Comput. Vis.*, 2019.

[43] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.