OASIS: A Large-Scale Dataset for Single Image 3D in the Wild

Weifeng Chen[1,2] Shengyi Qian[1] David Fan[2] Noriyuki Kojima[1] Max Hamilton[1] Jia Deng[2]

[1]University of Michigan, Ann Arbor [2]Princeton University

{wfchen,syqian,kojimano,johnmaxh}@umich.edu dfan@alumni.princeton.edu,jiadeng@princeton.edu

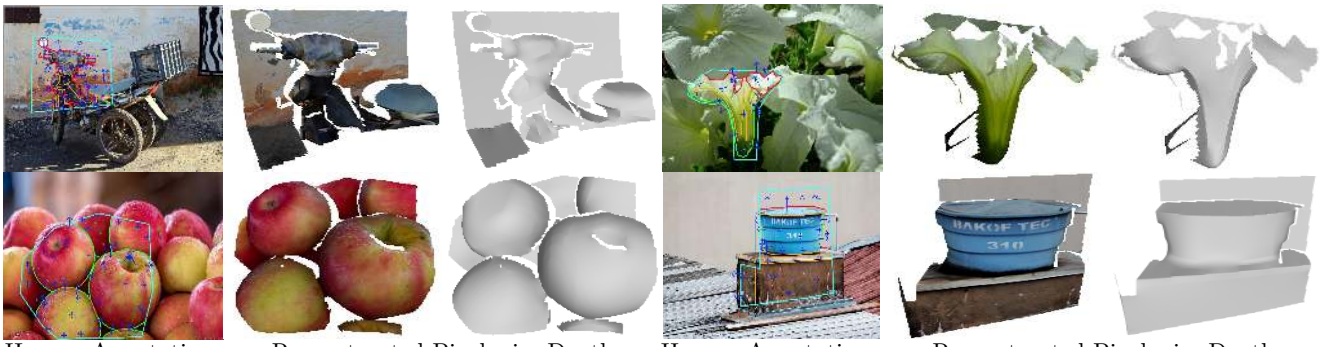


Figure 1. We introduce Open Annotations of Single-Image Surfaces (OASIS), a large-scale dataset of human annotations of 3D surfaces for 140,000 images in the wild. More examples in the supplementary material.

Abstract

Single-view 3D is the task of recovering 3D properties such as depth and surface normals from a single image. We hypothesize that a major obstacle to single-image 3D is data. We address this issue by presenting Open Annotations of Single Image Surfaces (OASIS), a dataset for single-image 3D in the wild consisting of annotations of detailed 3D geometry for 140,000 images. We train and evaluate leading models on a variety of single-image 3D tasks. We expect OASIS to be a useful resource for 3D vision research. Project site: `https://pvl.cs.princeton.edu/OASIS`.

1. Introduction

Single-view 3D is the task of recovering 3D properties such as depth and surface normals from a single RGB image. It is a core computer vision problem of critical importance. 3D scene interpretation is a foundation for understanding events and planning actions. 3D shape representation is crucial for making object recognition robust against changes in viewpoint, pose, and illumination. 3D from a single image is especially important due to the ubiquity of monocular images and videos. Even with a stereo camera with which 3D can be reconstructed by triangulating matching pixels from different views, monocular 3D cues are still necessary in textureless or specular regions where it is difficult to reliably match pixel values.

Single-image 3D is challenging. Unlike multiview 3D, it is ill-posed and resists tractable analytical formulation except in the most simplistic settings. As a result, data-driven approaches have shown greater promise, as evidenced by a plethora of works that train deep networks to map an RGB image to depth, surface normals, or 3D models [11, 17, 36, 14, 43, 24]. However, despite substantial progress, the best systems today still struggle with handling scenes "in the wild"— arbitrary scenes that a camera may encounter in the real world. As prior work has shown [5], state-of-art systems often give erroneous results when presented with unfamiliar scenes with novel shapes or layouts.

We hypothesize that a major obstacle of single-image 3D is data. Unlike object recognition, whose progress has been propelled by datasets like ImageNet [10] covering diverse object categories with high-quality labels, single-image 3D has lacked an ImageNet equivalent that covers diverse scenes with high-quality 3D ground truth. Existing datasets are restricted to either a narrow range of scenes [31, 9] or simplistic annotations such as sparse relative depth pairs or surface normals [5, 7].

In this paper we introduce *Open Annotations of Single-Image Surfaces* (OASIS), a large-scale dataset for single-image 3D in the wild. It consists of human annotations that

enable pixel-wise reconstruction of 3D surfaces for 140,000 randomly sampled Internet images. Fig. 1 shows the human annotations of example images along with the reconstructed surfaces.

A key feature of OASIS is its rich annotations of human 3D perception. Six types of 3D properties are annotated for each image: occlusion boundary (depth discontinuity), fold boundary (normal discontinuity), surface normal, relative depth, relative normal (orthogonal, parallel, or neither), and planarity (planar or not). These annotations together enable a reconstruction of pixelwise depth.

To construct OASIS, we created a UI for interactive 3D annotation. The UI allows a crowd worker to annotate the aforementioned 3D properties. It also provides a live, rotatable rendering of the resulting 3D surface reconstruction to help the crowd worker fine-tune their annotations.

It is worth noting that 140,000 images may not seem very large compared to millions of images in datasets like ImageNet. But the number of images can be a misleading metric. For OASIS, annotating one image takes 305 seconds on average. In contrast, verifying a single image-level label takes no more than a few seconds. Thus in terms of the total amount of human time, OASIS is already comparable to millions of image-level labels.

OASIS opens up new research opportunities on a wide range of single-image 3D tasks—depth estimation, surface normal estimation, boundary detection, and instance segmentation of planes—by providing in-the-wild ground truths either for the first time, or at a much larger scale than prior work. For depth estimation and surface normals, *pixelwise* ground truth is available for images in the wild for the first time—prior data in the wild provide only sparse annotations [5, 6]. For the detection of occlusion boundaries and folds, OASIS provides annotations at a scale 700 times larger than prior work—existing datasets [33, 15] have annotations for only about 200 images. For instance segmentation of planes, ground truth annotation is available for images in the wild for the first time.

To facilitate future research, we provide extensive statistics of the annotations in OASIS, and train and evaluate leading deep learning models on a variety of single-image tasks. Experiments show that there is a large room for performance improvement, pointing to ample research opportunities for designing new learning algorithms for single-image 3D. We expect OASIS to serve as a useful resource for 3D vision research.

2. Related Work

3D Ground Truth from Depth-Sensors and Computer Graphics Major 3D datatsets are either collected by sensors [31, 12, 29, 30, 9] or synthesized with Computer Graphics [4, 23, 32, 22, 26]. But due to the limitations of depth sensors and the lack of varied 3D assets for rendering, the diversity of scenes is quite limited. For example, sensor-based ground truth is mostly for indoor or driving scenes [31, 9, 23, 32, 12].

3D Ground Truth from Multiview Reconstruction Single-image 3D training data can also be obtained by applying classical Structure-from-Motion (SfM) algorithms on Internet images or videos [18, 38, 6]. However, classical SfM algorithms have many well known failure modes including scenes with moving objects and scenes with specular or textureless surfaces. In contrast, humans can annotate all types of scenes.

3D Ground Truth from Human Annotations Our work is connected to many previous works that crowdsource 3D annotations of Internet images. For example, prior work has crowdsourced annotations of relative depth [5] and surface normals [7] at sparse locations of an image (a single pair of relative depth and a single normal per image). Prior work has also aligned pre-existing 3D models to images [39, 34]. However, this approach has a drawback that not every shape can be perfectly aligned with available 3D models, whereas our approach can handle arbitrary geometry.

Our work is related to that of Karsch et al. [15], who reconstruct pixelwise depth from human annotations of boundaries, with the aid of a shape-from-shading algorithm [2]. Our approach is different in that we annotate not only boundaries but also surface normals, planarity, and relative normals, and our reconstruction method does not rely on automatic shape from shading, which is still unsolved and has many failure modes.

One of our inspirations is LabelMe3D [28], which annotated 3D planes attached to a common ground plane. Another is OpenSurfaces [3], which also annotated 3D planes. We differ from LabelMe3D and OpenSurfaces in that our annotations recover not only planes but also curved surfaces. Our dataset is also much larger, being $600\times$ the size of LabelMe3D and $5\times$ of OpenSurfaces in terms of the number of images annotated. It is also more diverse, because LabelMe3D and OpenSurface include only city or indoor scenes.

3. Crowdsourcing Human Annotations

We use random keywords to query and download Creative Commons Flickr images with a known focal length (extracted from the EXIF data). Each image is presented to a crowd worker for annotation through a custom UI as shown in Fig. 2 (a). The worker is asked to mask out a region that she wishes to work on with a polygon of her choice, with the requirement that the polygon covers a pair of randomly pre-selected locations. She then works on the annotations and iteratively monitors the generated mesh (detailed in Sec 4) from an interactive preview window (Fig. 2 (a)).

Figure 2. **(a)** Our UI allows a user to annotate rich 3D properties and includes a preview window for interactive 3D visualization. **(b)** An illustration of the depth scaling procedure in our backend.

Occlusion Boundary and Fold An occlusion boundary denotes locations of depth discontinuity, where the surface on one side is physically disconnected from the surface on the other side. When it is drawn, the worker also specifies which side of the occlusion is closer to the viewer, i.e. depth order of the surfaces on both sides of the occlusion. Workers need to distinguish between two kinds of occlusion boundaries. *Smooth occlusion* (green in Fig 2 (a)) is where the the closer surface smoothly curves away from the viewer, and the surface normals should be orthogonal to the occlusion line and parallel to the image plane, and pointing toward the further side. *Sharp occlusion* (red in Fig 2 (a)) has none of these constraints. On the other hand, *fold* denotes locations of surface normal discontinuity, where the surface geometry changes abruptly, but the surfaces on the two sides of the fold are still physically attached to each other (orange in Fig 2 (a)).

Occlusion boundaries segment a region into subregions, each of which is a *continuous surface* whose geometry can change abruptly but remains physically connected in 3D. Folds further segment a continuous surface into *smooth surfaces* where the geometry vary smoothly without discontinuity of surface normals.

Surface Normal The worker first specifies if a smooth surface is planar or curved. She annotates one normal at each planar surface which indicates the orientation of the plane. For each curved surface, she annotates normals at as many locations as she sees fit. A normal is visualized as a blue arrow originating from a green grid (see supplementary material), rendered in perspective projection according to the known focal length. Such visualization helps workers perceive the normal in 3D [7]. To rotate and adjust the normal, the worker only needs to drag the mouse.

Relative Normal Finally, to annotate normals with higher accuracy, the worker specifies the *relative normal* between each pair of planar surfaces. She chooses between *Neither*, *Parallel* and *Orthogonal*. Surfaces pairs that are parallel or orthogonal to each other then have their normals adjusted automatically to reflect the relation.

Interactive Previewing While annotating, the worker can click a button to see a visualization of the 3D shape constructed from the current annotations (detailed later in Sec. 4). Workers can rotate or zoom to inspect the shape from different angles in a preview window (Fig 2 (a)). She keeps working on it until she is satisfied with the shape.

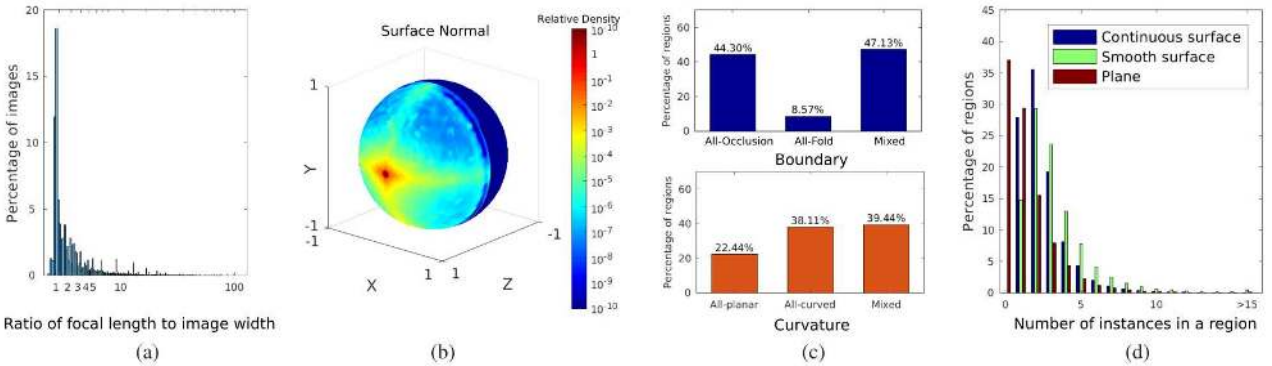
Quality Control Completing our 3D annotation task requires knowledge of relevant concepts. To ensure good quality of the dataset, we require each worker to complete a training course to learn concepts such as occlusions, folds and normals, and usage of the UI. She then needs to pass a qualification quiz before being allowed to work on our annotation task. Besides explicitly selecting qualified workers, we also set up a separate quality verification task on each collected mesh. In this task, a worker inspects the mesh to judge if it reflects the image well. Only meshes deemed high quality are accepted.

To improve our annotation throughput, we collected annotations from two sources: Amazon Mechanical Turk, which accounts for 31% of all annotations, and a data annotation company that employs full-time annotators, who supplied the rest of the annotations.

4. From Human Annotations to Dense Depth

Because humans do not directly annotate the depth value of each pixel, we need to convert the human annotations to pixelwise depth in order to visualize the 3D surface.

Generating Dense Surface Normals We first describe how we generate dense surface normals from annotations. We assume the normals to be smoothly varying in the spatial domain, except across folds or occlusion boundaries where the normals change abruptly. Therefore, our system propagates the known normals to the unknown ones by requiring the final normals to be smooth overall, but stops the propa-



(a) (b) (c) (d)

Figure 3. Statistics of OASIS. (a) The distribution of focal length (unit: relative length to the image width). (b) The distribution of surface normals. (c) Boundary: the ratio of regions containing only occlusion, only fold, and both. Curvature: the distribution of regions containing only planes, only curved surfaces, and both. (d) The frequency distribution of each surface type in a region.

gation at fold and occlusion lines.

More concretely, let N_p denote the normal at pixel p on a normal map N, and F, O denotes the pixels belong to the folds and occlusion boundaries. We have a set of known normals $\tilde{N}$ at locations P_{known} from (1) surface normal annotations by workers, and (2) the pre-computed normals along the smooth occlusion boundaries as mentioned in Sec 3. Each pixel p has four neighbors $\Phi(p)$. If p is on an occlusion boundary, its neighbors on the closer side of this boundary are $\Gamma_O(p)$. If p is on a fold line, only its neighbors $\Gamma_F(p)$ on one fixed random side of this line are considered. We solve for the optimal normal N^* using LU factorization and then normalize it into unit norm:

$$N^* = \underset{N}{\operatorname{argmin}} \sum_{p \notin F \cup O} \sum_{\substack{q \in \Phi(p) \\ q \notin F \cup O}} |N_p - N_q|^2 + \sum_{p \in O} \sum_{q \in \Gamma_O(p)} |N_p - N_q|^2 + \sum_{p \in F} \sum_{q \in \Gamma_F(p)} |N_p - N_q|^2 \quad (1)$$

$$\text{s.t.} \quad N_p = \tilde{N}_p, \forall p \in P_{known} \quad (2)$$

Generating Dense Depth Our depth generation pipeline consists of two stages: First, from surface normals and focal length, we recover the depth of each *continuous surface* through integration [25]. Next, we adjust the depth order among these surfaces by performing surface-wise depth scaling (Fig. 2 (b)), i.e. each surface has its own scale factor.

Our design is motivated by this fact: in single-view depth recovery, depth within continuous surface can be recovered only up to an ambiguous scale; thus different surfaces may end up with different scales, leading to incorrect depth ordering between surfaces. But workers already decide which side of an occlusion boundary is closer to the viewer. Based on such knowledge, we correct depth order by scaling the depth of each surface.

We now describe the details. Let $\mathbf{S}$ denotes the set of all continuous surface. From integration, we obtain the depth Z_S of each $S \in \mathbf{S}$. We then solve for a scaling factor X_S for each S, which is used in scaling depth Z_S. Let $\mathbf{O}$ denote the set of occlusion boundaries. Along $\mathbf{O}$, we densely sample a set of point pairs $\mathbf{B}$. Each pair $(p, q) \in \mathbf{B}$ has p lying on the closer side of one of the occlusion boundaries $O_i \in \mathbf{O}$ and q the further side. The continuous surface a pixel p lies on is $S(p)$, and its depth is Z_p. The set of optimal scaling factors $\mathbf{X}^*$ is solved for as follows:

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} \sum_{S \in \mathbf{S}} X_S \quad (3)$$

$$\text{s.t.} \quad X_{S(p)} Z_p + \epsilon \leq X_{S(q)} Z_q, \forall (p, q) \in \mathbf{B} \quad (4)$$

$$X_S \geq \eta, \forall S \in \mathbf{S} \quad (5)$$

where $\epsilon > 0$ is a minimum separation between surfaces, and $\eta > 0$ is a minimum scale factor. Eq.(4) requires the surfaces to meet the depth order constraints specified by point pairs $(p, q) \in \mathbf{B}$ after scaling. Meanwhile, Eq.(3) constrains the value of $\mathbf{X}$ so that they do not increase indefinitely. After correcting the depth order, the final depth for surface S is $X_S^* Z_S$. We normalize and reproject the final depth to 3D as point clouds, and generate 3D meshes for visualization.

5. Dataset Statistics

Statistics of Surfaces Fig. 3 plots various statistics of the 3D surfaces. Fig. 3 (a) plots the distribution of focal length. We see that focal lengths in OASIS vary greatly: they range from wide angle to telezoom, and are mostly $1\times$ to $10\times$ of the width of the image. Fig. 3 (b) visualizes the distribution of surface normals. We see that a substantial proportion of normals point directly towards the camera, suggesting that parallel-frontal surfaces frequently occur in natural scenes. Fig. 3 (c) presents region-wise statistics. We see that most regions (90%+) contain occlusion boundaries and close to half have both occlusion boundaries and folds (top). We also see that most regions (70%+) contain at least one curve surface (bottom). Fig. 3 (d) shows the histogram of

	NYU Depth [31] (depth mean: 2.471 m, depth std: 0.754 m)			Tanks & Temples [16] (depth mean: 4.309m, depth std: 3.059m)		
	Human-Human	Human-Sensor	CNN-Sensor	Human-Human	Human-Sensor	CNN-Sensor
Depth (EDist)	0.078m	0.095m	0.097m [17]	0.194m	0.213m	0.402m [17]
Normals (MAE)	13.13°	17.82°	14.19° [44]	14.33°	20.29°	29.11° [44]
Post-Rotation Depth (EDist)	0.037m	0.048m	-	0.082m	0.080m	-
Depth Order (WKDR)	5.68%	8.67%	11.90%	9.28%	10.80%	32.13%

Table 1. Depth and normal difference between different humans (Human-Human), between human and depth sensor (Human-Sensor), and between ConvNet and depth sensor (CNN-Sensor). The results are averaged over all human pairs.

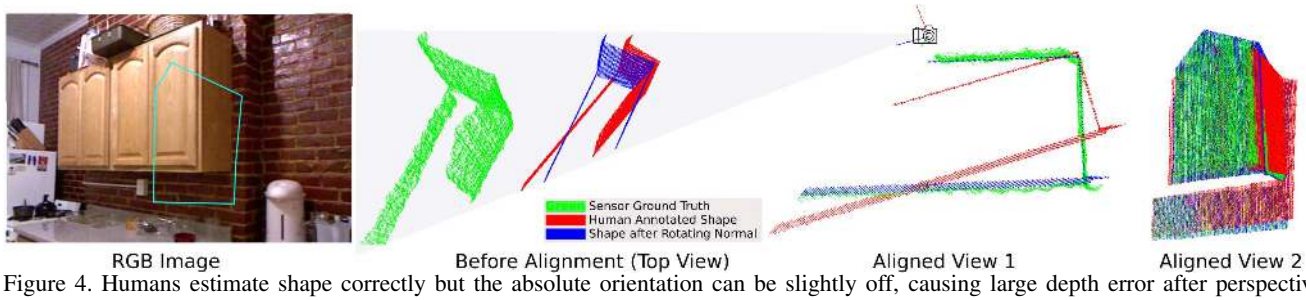


Figure 4. Humans estimate shape correctly but the absolute orientation can be slightly off, causing large depth error after perspective back-projection into 3D. Depth error drops significantly (from 0.07m to 0.01m) after a global rotation of normals.

the number of different kinds of surfaces in an annotated region. We see that most regions consist of multiple disconnected pieces and have non-trivial geometry in terms of continuity and smoothness.

Annotation Quality We study how accurate and consistent the annotations are. To this end, we randomly sample 50 images from NYU Depth [31] and 70 images from Tanks and Temples [16], and have 20 workers annotate each image. Tab. 1 reports the depth and normal difference between human annotations, between human annotations and sensor ground truth, and between predictions from state-of-the-art ConvNets and sensor ground truth. Depth difference is measured by the mean Euclidean distance (EDist) between corresponding points in two point clouds, after aligning one to the other through a global translation and scaling (surface-wise scaling for human annotations and CNN predictions). Normal difference is measured in Mean Angular Error (MAE). We see in Tab. 1 that human annotations are highly consistent with each other and with sensor ground truth, and are better than ConvNet predictions, especially when the ConvNet is not trained and tested on the same dataset.

We observe that humans often estimate the shape correctly, but the overall orientation can be slightly off, causing a large depth error against sensor ground truth (Fig. 4). This error can be particularly pronounced for planes close to orthogonal to the image plane. Thus we also compute the error after a rotational alignment with the sensor ground truth—we globally rotate the human annotated normals (up to 30 degrees) before generating the shape. After accounting for this global rotation of normals, human-sensor depth difference is further reduced by 47.96% (relative) for NYU and 62.44% (relative) for Tanks and Temples; a significant drop of normal error is also observed in human-human difference.

We also measure the qualitative aspect of human annotations by evaluating the WKDR metric [5], i.e. the percentage of point pairs with inconsistent depth ordering between query and reference depth. Depth pairs are sampled in the same way as [5]. Tab. 1 again shows that human annotations are qualitatively accurate and highly consistent with each other.

It is worth noting that metric 3D accuracy is not required for many tasks such as navigation, object manipulation, and semantic scene understanding—humans do well without perfect metric accuracy. Therefore human perception of depth alone can be the gold standard for training and evaluating vision systems, regardless of its metric accuracy. As a result, our dataset would still be valuable even if it were less metrically accurate than it is currently.

6. Experiments

To facilitate future research, we use OASIS to train and evaluate leading deep learning models on a suite of single-image 3D tasks including depth estimation, normal estimation, boundary detection, plane segmentation. Qualitative results are shown in Fig. 5. A train-val-test split of 110K, 10K, 20K is used for all tasks.

For each task we estimate human performance to provide an upperbound accounting for the variance of human annotations. We randomly sample 100 images from the test set, and have each image re-annotated by 8 crowd workers. That is, each image now has "predictions" from 8 different humans. We evaluate each prediction and report the mean as the performance expected of an average human.

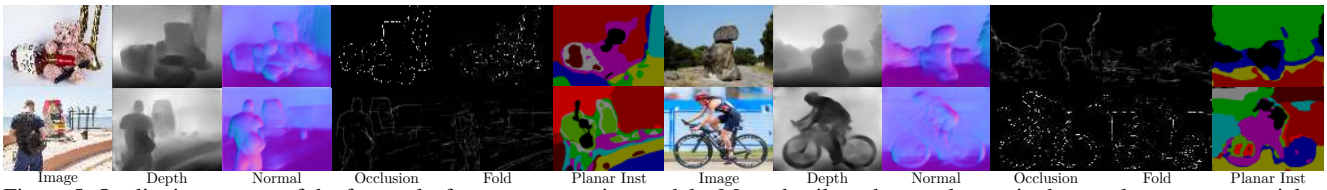


Figure 5. Qualitative outputs of the four tasks from representative models. More details and examples are in the supplementary material.

6.1. Depth Estimation

We first study single-view depth estimation. OASIS provides pixelwise *metric* depth in the wild. But as discussed in Sec 4, due to inherent single-image ambiguity, depth in OASIS is independently recovered within each continuous surface, after which the depth undergoes a surface-wise scaling to correct the depth order. The recovered depth is only accurate up to scaling within each continuous surface and ordering between continuous surfaces.

Given this, in OASIS we provide metric depth ground truths that is surface-wise accurate up to a scaling factor. This new form of depth necessitates new evaluation metrics and training losses.

Depth Metric The images in OASIS have varied focal lengths. This means that to evaluate depth estimation, we cannot simply use pixelwise difference between a predicted depth map and the ground truth map. This is because the predicted 3D shape depends greatly on the focal length—given the same depth values, decreasing the focal length will flatten the shape along the depth dimension. In practice, the focal length is often unknown for a test image. Thus, we require a depth estimator to predict a focal length along with depth. Because the predicted focal length may differ from the ground truth focal length, pixelwise depth difference is a poor indicator of how close the predicted 3D shape is to the ground truth.

A more reasonable metric is the Euclidean distance between the predicted and ground-truth 3D point cloud. Concretely, we backproject the predicted depth Z to a 3D point cloud $\mathbf{P} = \{(X_p, Y_p, Z_p)\}$ using f (the predicted focal length), and ground truth depth Z^* to $\mathbf{P}^* = \{(X_p^*, Y_p^*, Z_p^*)\}$ using f^* (the ground truth focal length). We then calculate the distance between $\mathbf{P}$ and $\mathbf{P}^*$.

The metric also needs to be invariant to surface-wise depth scaling and translation. Therefore we introduce a surface-wise scaling factor $\lambda_{S_i} \in \mathbf{\Lambda}$, and a surface-wise translation $\delta_{S_i} \in \mathbf{\Delta}$, to align each predicted surface $S_i \in \mathbf{S}$ in $\mathbf{P}$ to the ground truth point cloud $\mathbf{P}^*$ in a least square manner. The final metric, which we call Locally Scale-Invariant RMSE (LSIV_RMSE), is defined as:

$$LSIV\_RMSE(Z, Z^*) = \min_{\mathbf{\Lambda}, \mathbf{\Delta}} \sum_p \Big(\frac{(X_p^*, Y_p^*, Z_p^*)}{\sigma(X^*)} - \lambda_{S(p)}(X_p, Y_p, Z_p) - (0, 0, \delta_{S(p)}) \Big)^2, \quad (6)$$

where $S(p)$ denotes the surface a pixel p is on. The ground truth point cloud $\mathbf{P}^*$ is normalized to a canonical scale by the standard deviation of its X coordinates $\sigma(X^*)$. Under this metric, as long as $\mathbf{P}$ is accurate up to scaling and translation, it will align perfectly with $\mathbf{P}^*$, and get 0 error.

Note that LSIV_RMSE ignore the ordering between two separate surfaces; it allows objects floating in the air to be arbitrarily scaled. This is typically not an issue because in most scenes there are not many objects floating in the air. But we nonetheless also measure the correctness of depth ordering. We report WKDR [5], which is the percentage of point pairs that have incorrect depth order in the predicted depth. We evaluate on depth pairs sampled in the same way as [5], i.e. half are random pairs, half are from the same random horizontal lines.

Models We train and evaluate two leading depth estimation networks on OASIS: the Hourglass network [5], and ResNetD [38], a dense prediction network based on ResNet50. Each network predicts a metric depth map and a focal length, which are together used to backproject pixels to 3D points, which are compared against the ground truth to compute the LSIV_RMSE metric, which we optimize as the loss function during training. Note that we do not supervise on the predicted focal length.

We also evaluate leading pre-trained models that estimate single-image depth on OASIS, including FCRN [17] trained on ILSVRC [27] and NYU Depth [31], Hourglass [18] trained on MegaDepth [18], ResNetD [38] trained on a combination of datasets including ILSVRC [27], Depth in the Wild [5], ReDWeb [38] and YouTube3D [6]. For networks that do not produce a focal length, we use the validation set to find the best focal length that leads to the smallest LSIV_RMSE, and use this focal length for each test image. In addition, we also evaluate *plane*, a naive baseline that predicts a uniform depth map.

Tab. 2 reports the results. In terms of metric depth, we see that networks trained on OASIS perform the best. This is expected because they are trained to predict a focal length and to directly optimize the LSIV_RMSE metric. It is noteworthy that ImageNet pretraining provides a significant benefit even for this purely geometrical task. Off-the-shelf models do not perform better than the naive baseline, probably because they were not trained on diverse enough scenes or were not trained to optimize metric depth error. In terms of relative depth, it is interesting to see that ResNetD trained

Method	Training Data	LSIV_RMSE	WKDR
FCRN [17]	ImageNet [27] + NYU [31]	0.67	39.94%
Hourglass [5, 18]	MegaDepth [18]	0.67	38.37%
ResNetD [38, 6]	ImageNet [27] + YouTube3D [6]+ ReDWeb [38] + DIW [5]	0.66	34.03%
ResNetD [38]	ImageNet [27] + OASIS	0.37	32.04%
ResNetD [38]	OASIS	0.47	38.79%
Hourglass [5]	OASIS	0.47	39.64%
Plane	-	0.67	100.00%
Human (Approx)	-	0.24	19.33%

Table 2. Depth estimation performance of different networks on OASIS (lower is better). For networks that do not produce a focal length, we use the best focal length leading to the smallest error.

on ImageNet and OASIS performs the best, even though the training loss does not enforce depth ordering. We also see that there is still a significant gap between human performance and machine performance. At the same time, the gap is not hopelessly large, indicating the effectiveness of a large training set.

Method	Training Data	OASIS						
		Angle Distance		% Within t°			Relative Normal	
		Mean	Median	11.25°	22.5°	30°	AUC_o	AUC_p
Hourglass [7]	OASIS	23.34	18.08	31.44	59.79	72.25	0.5508	0.5439
Hourglass [7]	SNOW [7]	30.74	26.65	14.33	40.84	56.73	0.5329	0.4714
Hourglass [7]	NYU [31]	34.69	28.76	14.65	38.49	52.06	0.5415	0.5061
PBRS [44]	NYU [31]	38.09	33.00	11.94	32.58	45.29	0.5729	0.5227
Front_Facing	-	31.20	24.76	27.36	46.62	56.94	0.5000	0.5000
Human (Approx)	-	17.43	13.08	43.89	75.94	84.72	0.8870	0.6439

Table 3. Surface normal estimation on OASIS.

Method	Training Data	DIODE [35]				ETH3D [30]			
		Angle Distance	% Within t°			Angle Distance	% Within t°		
		Mean	11.25°	22.5°	30°	Mean	11.25°	22.5°	30°
Hourglass [7]	OASIS	34.57	13.71	35.69	49.65	34.51	23.52	52.04	62.73
Hourglass [7]	SNOW [7]	40.10	8.29	27.20	40.67	45.71	10.69	31.16	43.16
Hourglass [7]	NYU [31]	42.23	10.97	29.76	41.35	41.84	21.94	44.05	53.81
PBRS [44]	NYU [31]	42.59	9.96	29.08	40.72	39.91	18.68	44.76	56.08
Front_Facing	-	47.76	5.62	18.70	28.05	58.97	11.84	23.75	30.19

Table 4. Cross-dataset generalization.

6.2. Surface Normal Estimation

We now turn to single-view surface normal estimation. We evaluate on absolute normal, i.e. the pixel-wise predicted normal values, and *relative normal*, i.e. the parallel and orthogonal relation predicted between planar surfaces.

Absolute Normal Evaluation We use standard metrics proposed in prior work [37]: the mean and median of angular error measured in degrees, and the percentage of pixels whose angular error is within γ degrees.

We evaluate on OASIS four state-of-the-art networks that are trained to directly predict normals: (1) Hourglass [7] trained on OASIS, (2) Hourglass trained on the Surface Normal in the Wild (SNOW) dataset [7], (3) Hourglass trained on NYU Depth [31], and (4) PBRS, a normal estimation network by Zhang et al. [44] trained on NYU Depth [31]. We also include Front_Facing, a naive baseline predicting all normals to be orthogonal to the image plane.

Tab. 3 reports the results. As expected, the Hourglass network trained on OASIS performs the best. Although SNOW is also an in-the-wild dataset, the same network trained on it does not perform as well, but is still better than training on NYU. Notably, the human-machine gap appears fairly small numerically (17.43 versus 23.34 in mean angle error). However, we observe that the naive baseline can achieve 31.20; thus the dynamic range of this metric is small to start with, due to the natural distribution of normals in the wild. In addition, a close examination of the results suggests that these standard metrics of surface normals do not align well with perceptual quality. In natural images there can be large areas that dominate the metric but have uninteresting geometry, such as a blank wall in the background. For example, in Fig. 6, a neural network gets the background correct, but largely misses the important details in the foreground. This opens up an interesting research question about developing new evaluation metrics.

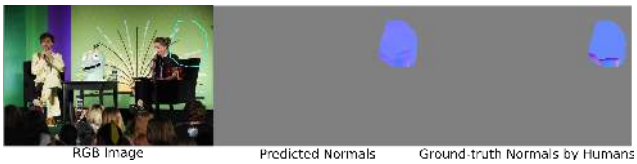


Figure 6. Limitations of standard metrics: a deep network gets low mean angle error but important details are wrong.

Relative Normal Evaluation We also evaluate the predicted normals in terms of relative relations, specifically orthogonality and parallelism. Getting these relations correct is important because it can help find vanishing lines and perform self-calibration.

We first define a metric to evaluate relative normal. From the human annotations, we first sample an equal number of point pairs from surface pairs that are parallel, orthogonal, and neither. Given a predicted normal map, we look at the two normals at each point pair and measure the angle θ between them. We consider them orthogonal if $|cos(\theta - 90^\circ)| < cos(\Theta_o)$, and parallel if $|cos(\theta)| > cos(\Theta_p)$, where Θ_o, Θ_p are thresholds. We then plot the Precision-and-Recall curve for orthogonal by varying Θ_o, and measure its Area Under Curve $\mathbf{AUC_o}$, using *neither* and *parallel* pairs as negative examples. Varying Θ_p and using *neither* and *orthogonal* as negative examples, we obtain $\mathbf{AUC_p}$ for parallel.

Tab. 3 reports results of relative normal evaluation. Notably, all methods perform similarly, and all perform very poorly compared to humans. This suggests that existing approaches to normal estimation have limitations in capturing orthogonality and parallelism, indicating the need for further research.

Cross-Dataset Generalization Next we study how networks trained on OASIS generalize to other datasets. Surface normal estimation is ideal for such evaluation because unlike depth, which is tricky to evaluate on a new dataset due to scale ambiguity and varying focal length, a normal estimation network can be directly evaluated on a new dataset without modification.

We train the same Hourglass network on OASIS, and NYU, and report their performance on two benchmarks not seen in training: DIODE [35] and ETH3D [30]. From Tab. 4 we see that training on NYU underperforms on all benchmarks, showing that networks trained on scene-specific datasets have difficulties generalizing to diverse scenes. Training on OASIS outperforms on all benchmarks, demonstrating the effectiveness of diverse annotations.

6.3. Fold and Occlusion Boundary Detection

Occlusion and fold are both important 3D cues, as they tell us about physical connectivity and curvature: *Occlusion* delineates the boundary at which surfaces are physically disconnected to each other, while *Fold* is where geometry changes abruptly but the surfaces remain connected.

Task We investigate joint boundary detection and occlusion-versus-fold classification: deciding whether a pixel is a boundary (fold or occlusion) and if so, which kind it is. Prior work has explored similar topics: Hoiem et al. [13] and Stein et al. [33] handcraft edge or motion features to perform occlusion detection, but our task involves folds, not just occlusion lines.

Metric \ Model	Edge: All Fold	Edge: All Occ	HED [40]	Hourglass [5]	Human (Approx)
ODS	0.123	0.539	0.533	0.585	0.810
OIS	0.129	0.576	0.584	0.639	0.815
AP	0.020	0.440	0.466	0.547	0.642

Table 5. Boundary detection performance on OASIS.

Evaluation Metric We adopt metrics similar to standard ones used in edge detection [1, 40]: F-score by optimal threshold per image (OIS), by fixed threshold (ODS) and average precision (AP). For a boundary to be considered correct, it has to be labeled correctly as either occlusion or fold. More details on the metrics can be found in the supplementary material.

To perform joint detection of fold and occlusion, we adapt and train two networks on OASIS: Hourglass [5], and a state-of-the-art edge detection network HED [40]. The networks take in an image, and output two probabilities per pixel: p_e is the probability of being an boundary pixel (occlusion or fold), and p_f is the probability of being a fold pixel. Given a threshold τ, pixels whose $p_e < \tau$ are neither fold nor occlusion. Pixels whose $p_e > \tau$ are fold if $p_F > 0.5$ and otherwise occlusion.

As baselines, we also investigate how a generic edge detector would perform on this task. We use HED network trained on BSDS dataset [1] to detect image edges, and classify the resulting edges to be either all occlusion (*Edge: All Occ*) or all fold (*Edge: All Fold*).

All results are reported on Tab 5. Hourglass outperforms HED when trained on OASIS, and significantly outperforms both the All-Fold and All-Occlusion baselines, but still underperforms humans by a large margin, suggesting that fold and occlusion boundary detection remains challenging in the wild.

6.4. Instance Segmentation of Planes

Our last task focuses on instance segmentation of planes in the wild. This task is important because planes often have special functional roles in a scene (e.g. supporting surfaces, walls). Prior work has explored instance segmentation of planes, but is limited to indoor or driving environments [21, 42, 20, 41]. Thanks to OASIS, we are able to present the first-ever evaluation of this task in the wild.

We follow the way prior work [21, 20, 42] performs this task: a network takes in an image, and produces instance masks of planes, along with an estimate of planar parameters that define each 3D plane. To measure performance, we report metrics used in instance segmentation literature [19]: the average precision (AP) computed and averaged across a range of overlap thresholds (ranges from 50% to 95% as in [19, 8]). A ground truth plane is considered correctly detected if it overlaps with one of the detected planes by more than the overlap threshold, and we penalize multiple detection as in [8]. We also report the AP at 50% overlap ($AP^{50\%}$) and 75% overlap ($AP^{75\%}$).

PlanarReconstruction by Yu et al. [42] is a state-of-the-art method for planar instance segmentation. We train PlanarReconstruction on three combinations of data: (1) ScanNet [9] only as done in [42], (2) OASIS only, and (3) ScanNet + OASIS. Tab. 6 compares their performance.

As expected, training on ScanNet alone performs the worse, because ScanNet only has indoor images. Training on OASIS leads to better performance. Leveraging both ScanNet and OASIS is the best overall. But even the best network significantly underperforms humans, suggesting ample space for improvement.

Method	Training Data	AP	$AP^{50\%}$	$AP^{75\%}$
PlanarReconstruction [42]	ScanNet [9]	0.076	0.161	0.065
	OASIS	0.127	0.250	0.112
	ScanNet [9] + OASIS	0.139	0.264	0.130
Human (Approx)	-	0.461	0.542	0.476

Table 6. Planar instance segmentation performance on OASIS.

7. Conclusion

We have presented OASIS, a dataset of rich human 3D annotations. We trained and evaluated leading models on a variety of single-image tasks. We expect OASIS to be a useful resource for 3D vision research.

Acknowledgement This work was partially supported by a National Science Foundation grant (No. 1617767), a Google gift, and a Princeton SEAS innovation grant.

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image seg-

mentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2011.

[2] Jonathan T Barron and Jitendra Malik. Color constancy, intrinsic images, and shape estimation. In *European Conference on Computer Vision*, pages 57–70. Springer, 2012.

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)*, 32(4), 2013.

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012.

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016.

[6] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5604–5613, 2019.

[7] Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy*, pages 22–29, 2017.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas A Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, volume 2, page 10, 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] 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.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[13] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, 2011.

[14] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–630, 2018.

[15] Kevin Karsch, Zicheng Liao, Jason Rock, Jonathan T. Barron, and Derek Hoiem. Boundary cues for 3d object shape recovery. In *CVPR*, 2013.

[16] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[17] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[18] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019.

[21] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2588, 2018.

[22] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[23] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. *arXiv preprint arXiv:1612.05079*, 2016.

[24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[25] Yvain Quéau, Jean-Denis Durou, and Jean-François Aujol. Normal integration: a survey. *Journal of Mathematical Imaging and Vision*, 60(4):576–593, 2018.

[26] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[28] Bryan C Russell and Antonio Torralba. Building a database of 3d scenes from user annotations. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2711–2718. IEEE, 2009.

[29] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. 3-d depth reconstruction from a single still image. *International journal of computer vision*, 76(1):53–69, 2008.

[30] Thomas Schöps, Johannes L Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017, 2017.

[31] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[32] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[33] Andrew N Stein and Martial Hebert. Occlusion boundaries from motion: Low-level detection and mid-level reasoning. *International journal of computer vision*, 82(3):325, 2009.

[34] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018.

[35] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

[36] Peng Wang, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, and Alan L Yuille. Surge: Surface regularized geometry estimation from a single image. In *Advances in Neural Information Processing Systems*, pages 172–180, 2016.

[37] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–547, 2015.

[38] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018.

[39] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. In *European Conference on Computer Vision*, pages 160–176. Springer, 2016.

[40] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[41] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[42] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1037, 2019.

[43] Bernhard Zeisl, Marc Pollefeys, et al. Discriminatively trained dense surface normal estimation. In *European conference on computer vision*, pages 468–484. Springer, 2014.

[44] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.